



Technical university of Crete

Department of

“Production and management engineering”



Thesis “Predicting the stock market: A neuro-fuzzy approach”

Project Supervisor : Dr. George Atsalakis PhD

Author : Konstantinos Matzouranis

Student Number : 200010102

ACKNOWLEDGEMENTS

The accomplishment of the current dissertation required the valuable help and support of several people.

First of all, the author would like to thank the supervisor of this thesis Professor, Dr.George Atsalakis for his support guidance and advice throughout the project.

Furthermore, it is essential for the author to express his appreciation to Mr. Rapanakis and Mr. Droulias for their invaluable advising and help within this year and to Mrs Metaxa for her support and understanding throughout long hours of study and research!

Finally, the author is thankful to his parents Mihalis and Katerina and his sister Efrossini, for their constant encouragement and support throughout his studies.

ABSTRACT

The purpose of this project is to study and analyze the way through which scientists have tried throughout the years, to come up with a successful scientific way of predicting the stock market.

Stock market has always been a challenge, for investors, scientist, academics, and in general lots of people around the globe. The belief that a way of predicting the stocks exists and therefore should be discovered, has tortured people since the beginning of all stock markets. Many attempts have been made throughout the years. Sometimes with scientific methods and algorithms, other times by experience and observation, and many times by unofficial methods (prays, mediums, palm reading etc!).

In this study the literature review of all available scientific methods which have had some results the last few decades, is presented, as well as the way these methods work their successes and their failures.

In the introduction of the project, a brief historical presentation is made and the general purpose and scope are been stated.

In chapter 2, a large literature review is made for stock prediction methods, and we analyze some methods mostly used in the past to predict stocks.

The fundamental and technical analyses are described, as well as the efficient market hypothesis and Elliott's wave theory.

In chapter 3, we talk about neural networks, and the way they affected stock market's prediction as well as other scientific areas.

Chapter 4 is a presentation of fuzzy logic and its applications in business and finance. Finally, chapter 5 examines the combination of neural systems and fuzzy logic, into big neuro-fuzzy systems which have been the new approach in predicting stocks over the last two decades.

TABLE OF CONTENTS

Acknowledgements.....	i
Abstract.....	ii
Chapter 1. Introduction	
Introduction	1
Chapter 2. Literature review	
2.1. Stock market prediction	5
2.2.1 Fundamental analysis	6
2.2.2 Technical analysis.....	26
2.2.3 Efficient market hypothesis	33
2.2.4 Elliott wave principle.....	49
Chapter 3. Artificial Neural Networks	
3.1. The biological example	60
3.2. Historical Background	63
3.3 The basic artificial model	65
3.4 General framework	70
3.5 Training of ANN	73
3.6. Paradigms of learning.....	75
3.7 Applications of ANN in real case scenarios.....	83

Chapter 4. Fuzzy logic

4.1. Definition and terminology	93
4.2 Basic concepts.....	95
4.3 Operations on fuzzy sets	105
4.4. Membership functions.....	108
4.5. Fuzzy rules.....	113
4.6. Fuzzy inference systems	119
4.7 Applications of fuzzy logic in finance.....	128

Chapter 5. Neuro fuzzy systems

5.1. The basics.....	132
5.2 Types of neuro fuzzy systems.....	134
5.3 ANFIS	140
5.4 Trainable neuro fuzzy and training procedures	144
5.5 Prediction of stocks using neuro fuzzy networks.....	149
References	153

Chapter 1

Introduction

The prediction of stock market movement has been an issue of interest for centuries. Despite years of study and the latest technology, it seems that no method has been discovered that consistently works. Previous attempts at solving this problem have employed a wide of array tools and strategies. Among these, many utilized some type of neural network in an attempt to learn from historical data and apply those findings to predict current data (White, 1998; Roman and Jameel, 1996; Wu and Lu, 1993). Other approaches use technical analysis to locate patterns in data that suggest certain movement trends (Chenowith, 1996). These patterns in the data are constructed by conforming to a series of rules based on the stock price and volume of trading over a series of days. Technical analysis is based on the assumption that previous market data reflects all of the relevant information about a stock and therefore predictions can be made from this. Still another method, known as fundamental analysis, looks at a business' financial statements, competitors, and business strategies in addition to evaluating historical data in an effort to make a prediction. The goal of fundamental analysis is to determine the value of a stock based on the previously mentioned factors and to act on the assumption that the actual stock price will eventually reflect the determined value. As a result, fundamental analysis usually works best over longer periods of time, whereas technical analysis is more suitable for short term trading (Murphy, 1999). Neural networks offer the ability to process large amounts of data quickly and retain information learned from that data to be used again in the future. Technical analysis provides a slightly more concrete method of evaluation at the cost of locating and implementing the algorithms yourself. Fundamental analysis seems optimal as it utilizes the largest information set and has many well known practitioners such as Warren Buffett and Peter Lynch, but is more difficult to implement via an Artificial Intelligence system mainly because such information is not often

publicly available. Furthermore, assessment techniques for fundamental analysis are far more complicated and difficult to automate. Lastly, gains realized by fundamental analysis usually take years to mature fully (Murphy 1999), so comparative research studies would not be feasible.

The difficulty with technical analysis is that a complete pattern is required to make an accurate prediction on the stock movement (Murphy 1999). Ideally, such a prediction should be made before the pattern was completed to enable prediction.

Historically, the efforts made to predict the stock market, are as old as the stock market itself. Scientists have been trying to predict the stock market forever in order to maximize profits. But what is so special about predicting stocks?

Stock time series have a number of specific properties that although not unique when examined one by one, together make the prediction task rather unusual and call for specific considerations when developing and evaluating a prediction system.

1) Prediction of stocks is generally believed to be a very difficult task. The most common viewpoint especially among academics is that the risk of predicting stocks is comparable to that of inventing a perpetuum mobile or solving problems like the quadrature of the circle.

2) The prediction process resembles very much a random-walk. The autocorrelation for day to day changes is very low. Hawawini and Keim conclude on several studies that the serial correlation in stock price time series is economically and statistically insignificant.

3) The process is “regime shifting”, meaning that the underlying process is time varying. The noise level and volatility in the time series change as the stock markets move in and out of periods of “turbulence”, “hausse” and “baisse” (rise and fall). This causes great problems for conventional algorithms for time series production.

On the positive side, the application of a prediction system for stock prices is somewhat special:

1) A successful prediction algorithm does not have to provide predictions for all points in the time series. The important measure of success is the hit rate and the generated profit at the points where the algorithms produces predictions. The behavior in between these points is not equally interesting. A missed opportunity to profit does not mean lost money!!

Another interesting property of this problem area is that the prediction algorithms themselves get integrated in the data generating process. Assume as an example that a new, sophisticated algorithm show a superior and unquestionable predictive power when applied to both historical data and real trading. As this technique becomes public knowledge and commonly used, the market participants will include the new algorithm as a new tool together with all previously suggested and used stock evaluation methods. The predictive power of the new algorithm will then decrease until it becomes useless. Viewed this way, what remains to make profit on in trading is no more than the residual from the superimposed prediction algorithms that comprise the process of price generation. If the superimposed algorithms do a good job, the residual should be indeed the near random walk that we are facing.

The scope of this paper is to make a deep approach to the problem of stock prediction as it has been dealt with from scientists. The many approaches of stock prediction are presented thoroughly and analyzed through examples. In the last three chapters, the discussion turns to neural networks and fuzzy logic, the innovative theories that changed the way predictions are made through the last few decades and their applications to stock market prediction. As we will see, throughout this paper, there have been minor or major successes in the field of stock market prediction during all these years, as well as minor or major failures. The attempt to predict efficiently the stock market, is a field that always will intrigue scientists since, many believe that a pattern can be successfully created in order to maximize profits. Up to now, despite some descent attempts, there hasn't been a theory that will be able to predict efficiently. Will someone be able to "crack" the code and simply transform investments into pure functions and mathematics? The question will remain unanswered as I predict for the years to come

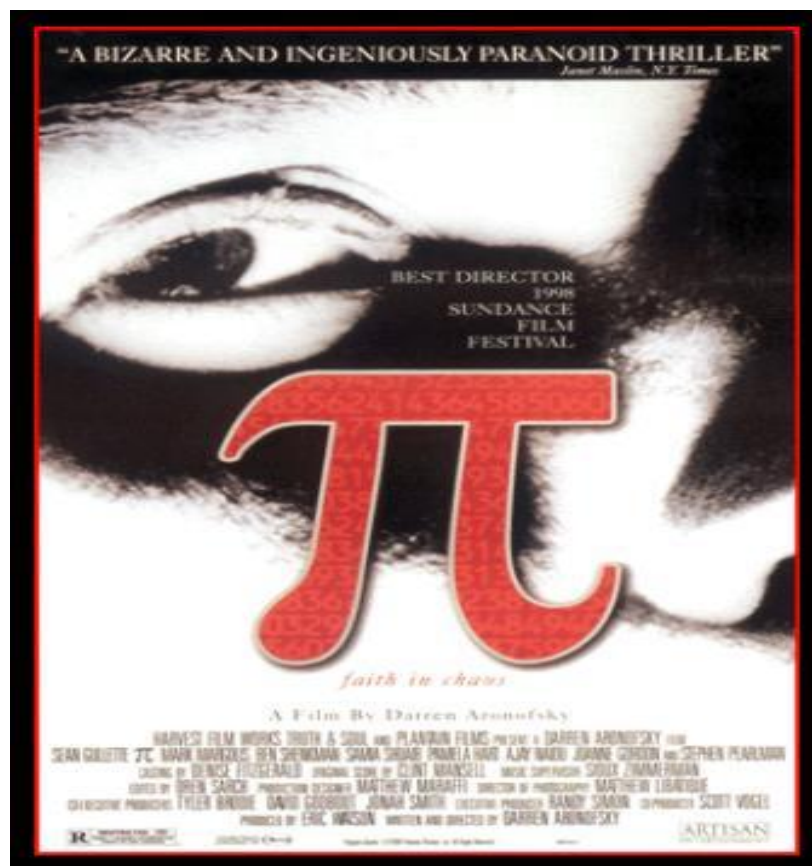


Figure 1: The movie pi in 1998 was the first cinematographic attempt to present the problematic of stock prediction. A deranged man tries to invent a pattern in order to successfully reveal the secrets of stock market and predict its movements.

CHAPTER 2

Literature Review

2.1 STOCK MARKET PREDICTION

Stock Market prediction has always had a certain appeal for researchers. While numerous scientific attempts have been made, no method has been discovered to accurately predict stock price movement. The difficulty of prediction lies in the complexities of modeling market dynamics. Even with a lack of consistent prediction methods, there have been some mild successes. Stock Prices are considered to be very dynamic and susceptible to quick changes because of the underlying nature of the financial domain and in part because of the mix of known parameters (Previous Days Closing Price, P/E Ratio etc.) and unknown factors (like Election Results, Rumors etc.)

An intelligent trader would predict the stock price and buy a stock before the price rises, or sell it before its value declines. Though it is very hard to replace the expertise that an experienced trader has gained, an accurate prediction algorithm can directly result into high profits for investment firms, indicating a direct relationship between the accuracy of the prediction algorithm and the profit made from using the algorithm. In this chapter, we will analyze the existing forecasting methods, before the use of neural networks and fuzzy logic was completely developed. We will examine their theoretical background and present their successes as well as their critics.

The most well-known and therefore used stock market prediction techniques (as presented below), are: Fundamental analysis, Technical analysis, Efficient market hypothesis (also known as the random walk theory) and Elliott wave theory.

2.2 DIFFERENT APPROACHES TO STOCK PRICE ANALYSIS

2.2.1 Fundamental Analysis

Fundamental analysis is the examination of the underlying forces that affect the well being of the economy, industry groups, and companies. As with most analysis, the goal is to derive a forecast and profit from future price movements. At the company level, fundamental analysis may involve examination of financial data, management, business concept and competition. At the industry level, there might be an examination of supply and demand forces for the products offered. For the national economy, fundamental analysis might focus on economic data to assess the present and future growth of the economy. To forecast future stock prices, fundamental analysis combines economic, industry, and company analysis to derive a stock's current fair value and forecast future value. If fair value is not equal to the current stock price, fundamental analysts believe that the stock is either over or under valued and the market price will ultimately gravitate towards fair value. Fundamentalists do not heed the advice of the random walkers and believe that markets are weak-form efficient. By believing that prices do not accurately reflect all available information, fundamental analysts look to capitalize on perceived price discrepancies.

One of the most famous and successful fundamental analysts is the Oracle of Omaha, Warren Buffett, who is well known for successfully employing fundamental analysis to pick securities. His abilities have turned him into a billionaire

The end goal of performing fundamental analysis is to produce a value that an investor can compare with the security's current price, with the aim of figuring out what sort of position to take with that security (underpriced = buy, overpriced = sell or short). This method of security analysis is considered to be the opposite of technical analysis. Fundamental analysis is about using real data to evaluate a security's value. Although most analysts use fundamental analysis to value stocks, this method of valuation can be used for just about any type of security.

For example, an investor can perform fundamental analysis on a bond's value by looking at economic factors, such as interest rates and the overall state of the economy, and information about the bond issuer, such as potential changes in credit ratings. For assessing stocks, this method uses revenues, earnings, future growth, return on equity, profit margins and other data to determine a company's underlying value and potential for future growth. In terms of stocks, fundamental analysis focuses on the financial statements of the company being evaluated. The biggest part of fundamental analysis involves delving into the financial statements. Also known as quantitative analysis, this involves looking at revenue, expenses, assets, liabilities and all the other financial aspects of a company. Fundamental analysts look at this information to gain insight on a stock's future performance.

The Very Basics

Economic Forecast

First and foremost in a top-down approach would be an overall evaluation of the general economy. The economy is like the tide and the various industry groups and individual companies are like boats. When the economy expands, most industry groups and companies benefit and grow. When the economy declines, most sectors and companies usually suffer. Many economists link economic expansion and contraction to the level of interest rates. Interest rates are seen as a leading indicator for the stock market as well. Below is a chart of the S&P 500 and the yield on the 10-K year note over the last 30 years. Although not exact, a correlation between stock prices and interest rates can be seen. Once a scenario for the overall economy has been developed, an investor can break down the economy into its various industry groups.



Figure 2.1:10-year K note

When talking about stocks, fundamental analysis is a technique that attempts to determine a security's value by focusing on underlying factors that affect a company's *actual* business and its future prospects. On a broader scope, you can perform fundamental analysis on industries or the economy as a whole. The term simply refers to the analysis of the economic well-being of a financial entity as opposed to only its price-movements.

Fundamental analysis serves to answer questions, such as:

- Is the company's revenue growing?
- Is it actually making a profit?
- Is it in a strong-enough position to beat out its competitors in the future?
- Is it able to repay its debts?
- Is management trying to "cook the books"?

Of course, these are very involved questions, and there are literally hundreds of others you might have about a company. It all really boils down to one question: Is the company's stock a good investment? Fundamental analysis is the toolbox to answer this question.

Fundamentals: Quantitative and Qualitative

One could define fundamental analysis as “researching the fundamentals”, but that doesn’t tell you a whole lot unless you know what fundamentals are. As we mentioned in the introduction, the big problem with defining fundamentals is that it can include anything related to the economic well-being of a company. Obvious items include things like revenue and profit, but fundamentals also include everything from a company’s market share to the quality of its management. The various fundamental factors can be grouped into two categories: quantitative and qualitative. The financial meaning of these terms isn’t all that different from their regular definitions. Here is how the MSN Encarta dictionary defines the terms:

- ***Quantitative*** – capable of being measured or expressed in numerical terms.
- ***Qualitative*** – related to or based on the quality or character of something, often as opposed to its size or quantity.

In our context, quantitative fundamentals are numeric, measurable characteristics about a business. It’s easy to see how the biggest source of quantitative data is the financial statements. You can measure revenue, profit, assets and more with great precision. Turning to qualitative fundamentals, these are the less tangible factors surrounding a business - things such as the quality of a company’s board members and key executives, its brand-name recognition, patents or proprietary technology.

Quantitative Meets Qualitative

Neither qualitative nor quantitative analysis is inherently better than the other. Instead, many analysts consider qualitative factors in conjunction with the hard, quantitative factors. Take the Coca-Cola Company, for example. When

examining its stock, an analyst might look at the stock's annual dividend payout, earnings per share, P/E ratio and many other quantitative factors. However, no analysis of Coca-Cola would be complete without taking into account its brand recognition. Anybody can start a company that sells sugar and water, but few companies on earth are recognized by billions of people. It's tough to put your finger on exactly what the Coke brand is worth, but you can be sure that it's an essential ingredient contributing to the company's ongoing success.

The Concept of Intrinsic Value

Before we get any further, we have to address the subject of intrinsic value. One of the primary assumptions of fundamental analysis is that the price on the stock market does not fully reflect a stock's "real" value. After all, why would we be doing price analysis if the stock market were always correct? In financial jargon, this true value is known as the intrinsic value. For example, let's say that a company's stock was trading at \$20. After doing extensive homework on the company, you determine that it really is worth \$25. In other words, you determine the intrinsic value of the firm to be \$25. This is clearly relevant because an investor wants to buy stocks that are trading at prices significantly below their estimated intrinsic value.

This leads us to one of the second major assumptions of fundamental analysis: *in the long run, the stock market will reflect the fundamentals*. There is no point in buying a stock based on intrinsic value if the price never reflected that value. Nobody knows how long "the long run" really is. It could be days or years.

This is what fundamental analysis is all about. By focusing on a particular business, an investor can estimate the intrinsic value of a firm and thus find opportunities where he or she can buy at a discount. If all goes well, the investment will pay off over time as the market catches up to the

fundamentals.

The big unknowns are:

- 1) You don't know if your estimate of intrinsic value is correct; and
- 2) You don't know how long it will take for the intrinsic value to be reflected in the marketplace.

Fundamental analysis seeks to determine the intrinsic value of a company's stock. But since qualitative factors, by definition, represent aspects of a company's business that are difficult or impossible to quantify, incorporating that kind of information into a pricing evaluation can be quite difficult.

Business Model

Even before an investor looks at a company's financial statements or does any research, one of the most important questions that should be asked is: What exactly does the company do? This is referred to as a company's business model – it's how a company makes money. One can get a good overview of a company's business model by checking out its website or reading the first part of its 10-K filing.

Sometimes business models are easy to understand. Take McDonalds, for instance, which sells hamburgers, fries, soft drinks, salads and whatever other new special they are promoting at the time. It's a simple model, easy enough for anybody to understand.

Other times, we'd be surprised how complicated it can get. Boston Chicken Inc. is a prime example of this. Back in the early '90s its stock was the darling of Wall Street. At one point the company's CEO bragged that they were the "first new fast-food restaurant to reach \$1 billion in sales since 1969". The problem is, they didn't make money by selling chicken. Rather, they made their money from royalty fees and high-interest loans to franchisees. Boston Chicken was really nothing more than a big franchisor. On top of this, management was aggressive with how it recognized its revenue. As soon as it

was revealed that all the franchisees were losing money, the house of cards collapsed and the company went bankrupt.

At the very least, we should understand the business model of any company we invest in. The "*Oracle of Omaha*", Warren Buffett, rarely invests in tech stocks because most of the time he doesn't understand them. This is not to say the technology sector is bad, but it's not Buffett's area of expertise; he doesn't feel comfortable investing in this area. Similarly, unless you understand a company's business model, you don't know what the drivers are for future growth, and you leave yourself vulnerable to being blindsided like shareholders of Boston Chicken were.

Competitive Advantage

Another business consideration for investors is competitive advantage. A company's long-term success is driven largely by its ability to maintain a competitive advantage - and keep it. Powerful competitive advantages, such as Coca Cola's brand name and Microsoft's domination of the personal computer operating system, create a moat around a business allowing it to keep competitors at bay and enjoy growth and profits. When a company can achieve competitive advantage, its shareholders can be well rewarded for decades.

Management

Just as an army needs a general to lead it to victory, a company relies upon management to steer it towards financial success. Some believe that management is *the* most important aspect for investing in a company. It makes sense - even the best business model is doomed if the leaders of the company fail to properly execute the plan.

Corporate Governance

Corporate governance describes the policies in place within an organization denoting the relationships and responsibilities between management,

directors and stakeholders. These policies are defined and determined in the company charter and its bylaws, along with corporate laws and regulations. The purpose of corporate governance policies is to ensure that proper checks and balances are in place, making it more difficult for anyone to conduct unethical and illegal activities.

Good corporate governance is a situation in which a company complies with all of its governance policies and applicable government regulations in order to look out for the interests of the company's investors and other stakeholders.

Financial statements

Financial statements are the medium by which a company discloses information concerning its financial performance. Followers of fundamental analysis use the quantitative information gleaned from financial statements to make investment decisions. Before we jump into the specifics of the three most important financial statements - income statements, balance sheets and cash flow statements - we will briefly introduce each financial statement's specific function, along with where they can be found.

The Major Statements

The Balance Sheet

The balance sheet represents a record of a company's assets, liabilities and equity at a particular point in time. The balance sheet is named by the fact that a business's financial structure balances in the following manner:

$$\text{Assets} = \text{Liabilities} + \text{Shareholders' Equity}$$

Assets represent the resources that the business owns or controls at a given point in time. This includes items such as cash, inventory, machinery and buildings. The other side of the equation represents the total value of the financing the company has used to acquire those assets. Financing comes as a result of liabilities or equity. Liabilities represent debt (which of course must be paid back), while equity represents the total value of money that the

owners have contributed to the business - including retained earnings, which is the profit made in previous years.

The Income Statement

While the balance sheet takes a snapshot approach in examining a business, the income statement measures a company's performance over a specific time frame. Technically, you could have a balance sheet for a month or even a day, but you'll only see public companies report quarterly and annually.

The income statement presents information about revenues, expenses and profit that was generated as a result of the business' operations for that period.

Statement of Cash Flows

The statement of cash flows represents a record of a business' cash inflows and outflows over a period of time. Typically, a statement of cash flows focuses on the following cash-related activities:

- *Operating Cash Flow (OCF)*: Cash generated from day-to-day business operations
- *Cash from investing (CFI)*: Cash used for investing in assets, as well as the proceeds from the sale of other businesses, equipment or long-term assets
- *Cash from financing (CFF)*: Cash paid or received from the issuing and borrowing of funds

The cash flow statement is important because it's very difficult for a business to manipulate its cash situation. There is plenty that aggressive accountants can do to manipulate earnings, but it's tough to fake cash in the bank. For this reason some investors use the cash flow statement as a more conservative measure of a company's performance.

10-K and 10-Q

Now that we have an understanding of what the three financial statements represent, let's discuss where an investor can go about finding them. In the United States, the Securities And Exchange Commission (SEC) requires all companies that are publicly traded on a major exchange to submit periodic filings detailing their financial activities, including the financial statements mentioned above.

Some other pieces of information that are also required are an auditor's report, management discussion and analysis (MD&A) and a relatively detailed description of the company's operations and prospects for the upcoming year. All of this information can be found in the business' annual 10-K and quarterly 10-Q filings, which are released by the company's management and can be found on the internet or in physical form.

The 10-K is an annual filing that discloses a business's performance over the course of the fiscal year. In addition to finding a business's financial statements for the most recent year, investors also have access to the business's historical financial measures, along with information detailing the operations of the business. This includes a lot of information, such as the number of employees, biographies of upper management, risks, future plans for growth, etc. Businesses also release an annual report, which some people also refer to as the 10-K. The annual report is essentially the 10-K released in a fancier marketing format. It will include much of the same information, but not all, that you can find in the 10-K. The 10-K really is boring - it's just pages and pages of numbers, text and legalese. But just because it's boring doesn't mean it isn't useful. There is a lot of good information in a 10-K, and it's required reading for any serious investor. You can think of the 10-Q filing as a smaller version of a 10-K. It reports the company's performance after each fiscal quarter. Each year three 10-Q filings are released - one for each of the first three quarters. (**Note:** There is no 10-Q for the fourth quarter, because the 10-K filing is released during that time). Unlike the 10-K filing, 10-Q filings

are not required to be audited. Here's a tip if you have trouble remembering which is which: think "Q" for quarter.

Management Discussion and Analysis (MD&A)

As a preface to the financial statements, a company's management will typically spend a few pages talking about the recent year (or quarter) and provide background on the company. This is referred to as the management discussion and analysis (MD&A). In addition to providing investors a clearer picture of what the company does, the MD&A also points out some key areas in which the company has performed well.

The Auditor's Report

The auditors' job is to express an opinion on whether the financial statements are reasonably accurate and provide adequate disclosure. This is the purpose behind the auditor's report, which is sometimes called the "report of independent accountants". By law, every public company that trades stocks or bonds on an exchange must have its annual reports audited by a certified public accountants firm. An auditor's report is meant to scrutinize the company and identify anything that might undermine the integrity of the financial statements. The typical auditor's report is almost always broken into three paragraphs and written in the following fashion:

Independent Auditor's Report

Paragraph 1

Recounts the responsibilities of the auditor and directors in general and lists the areas of the financial statements that were audited.

Paragraph 2

Lists how the generally accepted accounting principles (GAAP) were applied, and what areas of the company were assessed.

Paragraph 3

Provides the auditor's opinion on the financial statements of the company being audited. This is simply an opinion, not a guarantee of accuracy.

While the auditor's report won't uncover any financial bombshells, audits give credibility to the figures reported by management. You'll only see unaudited financials for unlisted firms (those that trade OTCBB or on the Pink Sheets). While quarterly statements aren't audited, you should be very wary of any annual financials that haven't been given the accountants' stamp of approval.

The Notes to the Financial Statements

Just as the MD&A serves an introduction to the financial statements, the notes to the financial statements (sometimes called footnotes) tie up any loose ends and complete the overall picture. If the income statement, balance sheet and statement of cash flows are the heart of the financial statements, then the footnotes are the arteries that keep everything connected. Therefore, if you aren't reading the footnotes, you're missing out on a lot of information. The footnotes list important information that could not be included in the actual ledgers. For example, they list relevant things like outstanding leases, the maturity dates of outstanding debt and details on compensation plans, such as stock options, etc.

Generally speaking there are two types of footnotes:

Accounting Methods - This type of footnote identifies and explains the major accounting policies of the business that the company feels that you should be aware of. This is especially important if a company has changed accounting policies. It may be that a firm is practicing "cookie jar accounting" and is changing policies only to take advantage of current conditions in order to hide poor performance.

Disclosure - The second type of footnote provides additional disclosure that simply could not be put in the financial statements. The financial statements in an annual report are supposed to be clean and easy to follow. To maintain this cleanliness, other calculations are left for the footnotes. For example, details of long-term debt - such as maturity dates and the interest rates at which debt was issued - can give you a better idea of how borrowing costs are laid out. Other areas of disclosure include everything from pension plan

liabilities for existing employees to details about ominous legal proceedings involving the company.

The majority of investors and analysts read the balance sheet, income statement and cash flow statement but, for whatever reason, the footnotes are often ignored. What sets informed investors apart is digging deeper and looking for information that others typically wouldn't. No matter how boring it might be, read the fine print - it will make you a better investor.

Revenue as an investor signal

Revenue, also commonly known as sales, is generally the most straightforward part of the income statement. Often, there is just a single number that represents all the money a company brought in during a specific time period, although big companies sometimes break down revenue by business segment or geography.

The best way for a company to improve profitability is by increasing sales revenue. For instance, Starbucks Coffee has aggressive long-term sales growth goals that include a distribution system of 20,000 stores worldwide. Consistent sales growth has been a strong driver of Starbucks' profitability.

The best revenues are those that continue year in and year out. Temporary increases, such as those that might result from a short-term promotion, are less valuable and should garner a lower price-to-earnings multiple for a company.

What are the Expenses?

There are many kinds of expenses, but the two most common are the cost of goods sold (COGS) and selling, general and administrative expenses (SG&A). Cost of goods sold is the expense most directly involved in creating revenue. It represents the costs of producing or purchasing the goods or services sold by the company. For example, if Wal-Mart pays a supplier \$4 for a box of soap, which it sells to customers for \$5. When it is sold, Wal-Mart's cost of good sold for the box of soap would be \$4.

Next, costs involved in operating the business are SG&A. This category includes marketing, salaries, utility bills, technology expenses and other general costs associated with running a business. SG&A also includes depreciation and amortization. Companies must include the cost of replacing worn out assets. Remember, some corporate expenses, such as research and development (R&D) at technology companies, are crucial to future growth and should not be cut, even though doing so may make for a better-looking earnings report. Finally, there are financial costs, notably taxes and interest payments, which need to be considered.

Profits = Revenue – Expenses

Profit, most simply put, is equal to total revenue minus total expenses. However, there are several commonly used profit subcategories that tell investors how the company is performing. Gross profit is calculated as revenue minus cost of sales. Returning to Wal-Mart again, the gross profit from the sale of the soap would have been \$1 (\$5 sales price less \$4 cost of goods sold = \$1 gross profit).

Companies with high gross margins will have a lot of money left over to spend on other business operations, such as R&D or marketing. So be on the lookout for downward trends in the gross margin rate over time. This is a telltale sign of future problems facing the bottom line. When cost of goods sold rises rapidly, they are likely to lower gross profit margins - unless, of course, the company can pass these costs onto customers in the form of higher prices.

Operating profit is equal to revenues minus the cost of sales and SG&A. This number represents the profit a company made from its actual operations, and excludes certain expenses and revenues that may not be related to its central operations. High operating margins can mean the company has effective control of costs, or that sales are increasing faster than operating costs. Operating profit also gives investors an opportunity to do profit-margin comparisons between companies that do not issue a separate disclosure of their cost of goods sold figures (which are needed to do gross margin

analysis). Operating profit measures how much cash the business throws off, and some consider it a more reliable measure of profitability since it is harder to manipulate with accounting tricks than net earnings.

Net income generally represents the company's profit after all expenses, including financial expenses, have been paid. This number is often called the "bottom line" and is generally the figure people refer to when they use the word "profit" or "earnings".

When a company has a high profit margin, it usually means that it also has one or more advantages over its competition. Companies with high net profit margins have a bigger cushion to protect themselves during the hard times. Companies with low profit margins can get wiped out in a downturn. And companies with profit margins reflecting a competitive advantage are able to improve their market share during the hard times - leaving them even better positioned when things improve again.

The Snapshot of Health

The balance sheet, also known as the statement of financial condition, offers a snapshot of a company's health. It tells you how much a company owns (its assets), and how much it owes (its liabilities). The difference between what it owns and what it owes is its equity, also commonly called "net assets" or "shareholders equity".

The balance sheet tells investors a lot about a company's fundamentals: how much debt the company has, how much it needs to collect from customers (and how fast it does so), how much cash and equivalents it possesses and what kinds of funds the company has generated over time.

The Balance Sheet's Main Three

Assets, liability and equity are the three main components of the balance sheet. Carefully analyzed, they can tell investors a lot about a company's fundamentals:

Assets

There are two main types of assets: **current assets** and **non-current assets**. Current assets are likely to be used up or converted into cash within one business cycle - usually treated as twelve months. Three very important current asset items found on the balance sheet are: cash, inventories and accounts receivables.

Investors normally are attracted to companies with plenty of cash on their balance sheets. After all, cash offers protection against tough times, and it also gives companies more options for future growth. Growing cash reserves often signal strong company performance. Indeed, it shows that cash is accumulating so quickly that management doesn't have time to figure out how to make use of it. A dwindling cash pile could be a sign of trouble. That said, if loads of cash are more or less a permanent feature of the company's balance sheet, investors need to ask why the money is not being put to use. Cash could be there because management has run out of investment opportunities or is too short-sighted to know what to do with the money.

Inventories are finished products that haven't yet sold. As an investor, you want to know if a company has too much money tied up in its inventory. Companies have limited funds available to invest in inventory. To generate the cash to pay bills and return a profit, they must sell the merchandise they have purchased from suppliers. Inventory turnover (cost of goods sold divided by average inventory) measures how quickly the company is moving merchandise through the warehouse to customers. If inventory grows faster than sales, it is almost always a sign of deteriorating fundamentals.

Receivables are outstanding (uncollected bills). Analyzing the speed at which a company collects what it's owed can tell you a lot about its financial efficiency. If a company's collection period is growing longer, it could mean problems ahead. The company may be letting customers stretch their credit in order to recognize greater top-line sales and that can spell trouble later on, especially if customers face a cash crunch. Getting money right away is

preferable to waiting for it - since some of what is owed may never get paid. The quicker a company gets its customers to make payments, the sooner it has cash to pay for salaries, merchandise, equipment, loans, and best of all, dividends and growth opportunities.

Non-current assets are defined as anything not classified as a current asset. This includes items that are fixed assets, such as property, plant and equipment (PP&E). Unless the company is in financial distress and is liquidating assets, investors need not pay too much attention to fixed assets. Since companies are often unable to sell their fixed assets within any reasonable amount of time they are carried on the balance sheet at cost regardless of their actual value. As a result, it's possible for companies to grossly inflate this number, leaving investors with questionable and hard-to-compare asset figures.

Liabilities

There are current liabilities and non-current liabilities. Current liabilities are obligations the firm must pay within a year, such as payments owing to suppliers. Non-current liabilities, meanwhile, represent what the company owes in a year or more time. Typically, non-current liabilities represent bank and bondholder debt.

You usually want to see a manageable amount of debt. When debt levels are falling, that's a good sign. Generally speaking, if a company has more assets than liabilities, then it is in decent condition. By contrast, a company with a large amount of liabilities relative to assets ought to be examined with more diligence. Having too much debt relative to cash flows required to pay for interest and debt repayments is one way a company can go bankrupt. Look at the quick ratio. Subtract inventory from current assets and then divide by current liabilities. If the ratio is 1 or higher, it says that the company has enough cash and liquid assets to cover its short-term debt obligations.

$$\text{Quick Ratio} = \frac{\text{Current Assets} - \text{Inventories}}{\text{Current Liabilities}}$$

Equity

Equity represents what shareholders own, so it is often called shareholder's equity. As described above, equity is equal to total assets minus total liabilities.

$$\text{Equity} = \text{Total Assets} - \text{Total Liabilities}$$

The two important equity items are paid-in capital and retained earnings. Paid-in capital is the amount of money shareholders paid for their shares when the stock was first offered to the public. It basically represents how much money the firm received when it sold its shares. In other words, retained earnings are a tally of the money the company has chosen to reinvest in the business rather than pay to shareholders. Investors should look closely at how a company puts retained capital to use and how a company generates a return on it.

Most of the information about debt can be found on the balance sheet - but some assets and debt obligations are not disclosed there. For starters, companies often possess hard-to-measure intangible assets. Corporate intellectual property (items such as patents, trademarks, copyrights and business methodologies), goodwill and brand recognition are all common assets in today's marketplace. But they are not listed on company's balance sheets.

Strengths of Fundamental Analysis

Long-term Trends

Fundamental analysis is good for long-term investments based on long-term trends, very long-term. The ability to identify and predict long-term economic, demographic, technological or consumer trends can benefit patient investors who pick the right industry groups or companies.

Value Spotting

Sound fundamental analysis will help identify companies that represent a good value. Some of the most legendary investors think long-term and value. Graham and Dodd, Warren Buffett and John Neff are seen as the champions of value investing. Fundamental analysis can help uncover companies with valuable assets, a strong balance sheet, stable earnings, and staying power.

Business Acumen

One of the most obvious, but less tangible, rewards of fundamental analysis is the development of a thorough understanding of the business. After such painstaking research and analysis, an investor will be familiar with the key revenue and profit drivers behind a company. Earnings and earnings expectations can be potent drivers of equity prices. Even some technicians will agree to that. A good understanding can help investors avoid companies that are prone to shortfalls and identify those that continue to deliver. In addition to understanding the business, fundamental analysis allows investors to develop an understanding of the key value drivers and companies within an industry. A stock's price is heavily influenced by its industry group. By studying these groups, investors can better position themselves to identify opportunities that are high-risk (tech), low-risk (utilities), growth oriented (computer), value driven (oil), non-cyclical (consumer staples), cyclical (transportation) or income-oriented (high yield).

Knowing Who's Who

Stocks move as a group. By understanding a company's business, investors can better position themselves to categorize stocks within their relevant industry group. Business can change rapidly and with it the revenue mix of a company. This happened to many of the pure Internet retailers, which were not really Internet companies, but plain retailers. Knowing a company's business and being able to place it in a group can make a huge difference in relative valuations.

Weaknesses of Fundamental Analysis

Time Constraints

Fundamental analysis may offer excellent insights, but it can be extraordinarily time-consuming. Time-consuming models often produce valuations that are contradictory to the current price prevailing on Wall Street. When this happens, the analyst basically claims that the whole street has got it wrong. This is not to say that there are not misunderstood companies out there, but it is quite brash to imply that the market price, and hence Wall Street, is wrong.

Industry/Company Specific

Valuation techniques vary depending on the industry group and specifics of each company. For this reason, a different technique and model is required for different industries and different companies. This can get quite time-consuming, which can limit the amount of research that can be performed. A subscription-based model may work great for an Internet Service Provider (ISP), but is not likely to be the best model to value an oil company.

Analyst Bias

The majority of the information that goes into the analysis comes from the company itself. Companies employ investor relations managers specifically to handle the analyst community and release information. As Mark Twain said, "there are lies, damn lies, and statistics." When it comes to massaging the data or spinning the announcement, CFOs and investor relations managers are professionals. Only buy-side analysts tend to venture past the company statistics. Buy-side analysts work for mutual funds and money managers. They read the reports written by the sell-side analysts who work for the big brokers (CIBC, Merrill Lynch, Robertson Stephens, CS First Boston, Paine Weber, and DLJ to name a few). These brokers are also involved in underwriting and investment banking for the companies. Even though there are restrictions in place to prevent a conflict of interest, brokers have an ongoing relationship with the company under analysis. When reading these

reports, it is important to take into consideration any biases a sell-side analyst may have. The buy-side analyst, on the other hand, is analyzing the company purely from an investment standpoint for a portfolio manager. If there is a relationship with the company, it is usually on different terms. In some cases this may be as a large shareholder.

Definition of Fair Value

When market valuations extend beyond historical norms, there is pressure to adjust growth and multiplier assumptions to compensate. If Wall Street values a stock at 50 times earnings and the current assumption is 30 times, the analyst would be pressured to revise this assumption higher. There is an old Wall Street adage: the value of any asset (stock) is only what someone is willing to pay for it (current price). Just as stock prices fluctuate, so too do growth and multiplier assumptions. Are we to believe Wall Street and the stock price or the analyst and market assumptions?

It used to be that free cash flow or earnings were used with a multiplier to arrive at a fair value. In 1999, the S&P 500 typically sold for 28 times free cash flow. However, because so many companies were and are losing money, it has become popular to value a business as a multiple of its revenues. This would seem to be OK, except that the multiple was higher than the PE of many stocks! Some companies were considered bargains at 30 times revenues.

Fundamental analysis in general, can be valuable, but it should be approached with caution. If you are reading research written by a sell-side analyst, it is important to be familiar with the analyst behind the report. We all have personal biases, and every analyst has some sort of bias. There is nothing wrong with this, and the research can still be of great value. Learn what the ratings mean and the track record of an analyst before jumping off the deep end. Corporate statements and press releases offer good information, but they should be read with a healthy degree of skepticism to separate the facts from the spin. Press releases don't happen by accident;

they are an important PR tool for companies. Investors should become skilled readers to weed out the important information and ignore the hype.

2.2.2 TECHNICAL ANALYSIS

In finance, **technical analysis** is a security analysis discipline for forecasting the direction of prices through the study of past market data, primarily price and volume.

The principles of technical analysis derive from the observation of financial markets over hundreds of years. The oldest known hints of technical analysis appear in Joseph de la Vega's accounts of the Dutch markets in the 17th century. In Asia, the oldest example of technical analysis is thought to be a method developed by Homma Munehisa during early 18th century which evolved into the use of candlestick techniques, and is today a main charting tool.

Dow Theory is based on the collected writings of Dow Jones co-founder and editor Charles Dow, and inspired the use and development of modern technical analysis from the end of the 19th century. Other pioneers of analysis techniques include Ralph Nelson Elliott, William Delbert Gann and Richard Wyckoff who developed their respective techniques in the early 20th century.

Many more technical tools and theories have been developed and enhanced in recent decades, with an increasing emphasis on computer-assisted techniques.

Technical analysis really just studies supply and demand in a market in an attempt to determine what direction, or trend, will continue in the future. In other words, technical analysis attempts to understand the emotions in the market by studying the market itself, as opposed to its components. Technical analysis is a method of evaluating securities by analyzing the statistics generated by market activity, such as past prices and volume. Technical analysts do not attempt to measure a security's intrinsic value (as fundamentalists do), but instead use charts and other tools to identify patterns that can suggest future activity.

Just as there are many investment styles on the fundamental side, there are also many different types of technical traders. Some rely on chart patterns; others use technical indicators and oscillators, and most use some combination of the two. In any case, technical analysts' exclusive use of historical price and volume data is what separates them from their fundamental counterparts. Unlike fundamental analysts, technical analysts don't care whether a stock is undervalued - the only thing that matters is a security's past trading data and what information this data can provide about where the security might move in the future.

Technical analysis is widely used among traders and financial professionals, and is very often used by active day traders, market makers, and pit traders. In the 1960s and 1970s it was widely dismissed by academics. In a recent review, Irwin and Park reported that 56 of 95 modern studies found it produces positive results, but noted that many of the positive results were rendered dubious by issues such as data snooping so that the evidence in support of technical analysis was inconclusive; it is still considered by many academics to be pseudoscience. Academics such as Eugene Fama say the evidence for technical analysis is sparse and is inconsistent with the *weak form* of the efficient market hypothesis. Users hold that even if technical analysis cannot predict the future, it helps to identify trading opportunities.

In the foreign exchange markets, its use may be more widespread than fundamental analysis. While some isolated studies have indicated that technical trading rules might lead to consistent returns in the period prior to 1987, most academic work has focused on the nature of the anomalous position of the foreign exchange market. It is speculated that this anomaly is due to central bank intervention. Recent research suggests that combining various trading signals into a Combined Signal Approach may be able to increase profitability and reduce dependence on any single rule.

One of the most important concepts in technical analysis is that of trend. The meaning in finance isn't all that different from the general definition of the term - a trend is really nothing more than the general direction in which a security or market is headed. Take a look at the chart below:

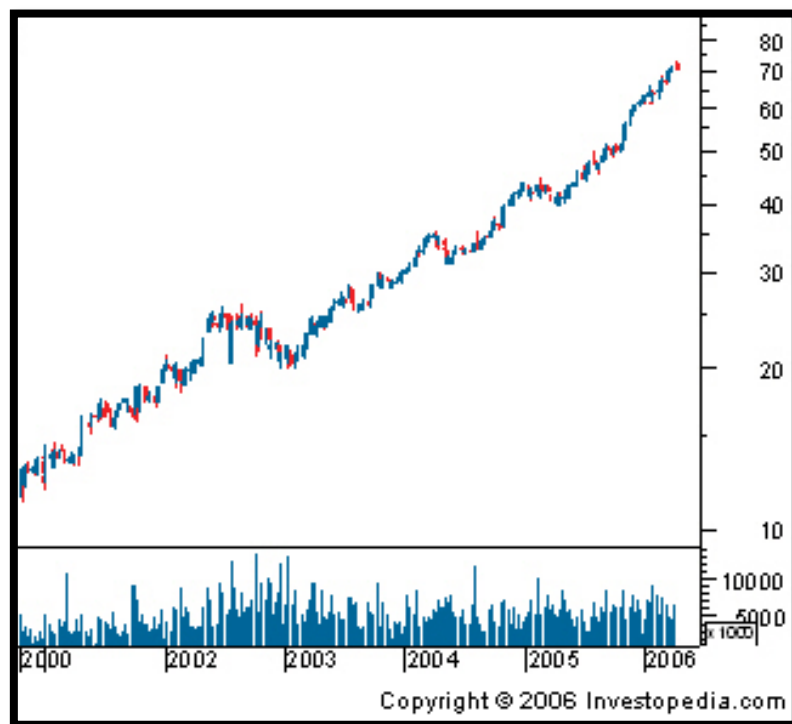


Figure 2.2

It isn't hard to see that the trend in Figure 2.2 is up. However, it's not always this easy to see a trend:

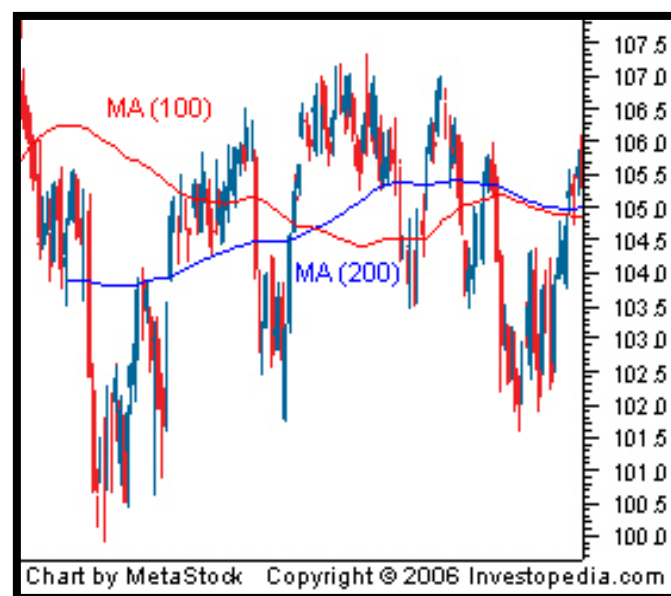


Figure 2.3

There are lots of ups and downs in this chart, but there isn't a clear indication of which direction this security is headed.

The field of technical analysis is based on three assumptions:

1. *The market discounts everything.*
2. *Price moves in trends.*
3. *History tends to repeat itself.*

1. The Market Discounts Everything

A major criticism of technical analysis is that it only considers price movement, ignoring the fundamental factors of the company. However, technical analysis assumes that, at any given time, a stock's price reflects everything that has or could affect the company - including fundamental factors. Technical analysts believe that the company's fundamentals, along with broader economic factors and market psychology, are all priced into the stock, removing the need to actually consider these factors separately. This only leaves the analysis of price movement, which technical theory views as a product of the supply and demand for a particular stock in the market.

2. Price Moves in Trends

In technical analysis, price movements are believed to follow trends. This means that after a trend has been established, the future price movement is more likely to be in the same direction as the trend than to be against it. Most technical trading strategies are based on this assumption. An example of a security that had an apparent trend is AOL from November 2001 through August 2002. A technical analyst or trend follower recognizing this trend would look for opportunities to sell this security. AOL consistently moves downward in price. Each time the stock rose, sellers would enter the market and sell the stock; hence the "zigzag" movement in the price. The series of "lower highs" and "lower lows" is a tell tale sign of a stock in a down trend. In

other words, each time the stock moved lower, it fell below its previous relative low price. Each time the stock moved higher, it could not reach the level of its previous relative high price.

Note that the sequence of lower lows and lower highs did not begin until August. Then AOL makes a low price that doesn't pierce the relative low set earlier in the month. Later in the same month, the stock makes a relative high equal to the most recent relative high. In this a technician sees strong indications that the down trend is at least pausing and possibly ending, and would likely stop actively selling the stock at that point.

3. History Tends To Repeat Itself

Another important idea in technical analysis is that history tends to repeat itself, mainly in terms of price movement. The repetitive nature of price movements is attributed to market psychology; in other words, market participants tend to provide a consistent reaction to similar market stimuli over time. Technical analysis uses chart patterns to analyze market movements and understand trends. Although many of these charts have been used for more than 100 years, they are still believed to be relevant because they illustrate patterns in price movements that often repeat themselves. Technical analysis can be used on any security with historical trading data. This includes stocks, futures and commodities, fixed-income securities, forex, etc. In this tutorial, we'll usually analyze stocks in our examples, but keep in mind that these concepts can be applied to any type of security. In fact, technical analysis is more frequently associated with commodities and forex, where the participants are predominantly traders.

The Differences

Charts vs. Financial Statements

At the most basic level, a technical analyst approaches a security from the charts, while a fundamental analyst starts with the financial statements.

By looking at the **balance sheet**, **cash flow statement** and **income**

statement, a fundamental analyst tries to determine a company's value. In financial terms, an analyst attempts to measure a company's intrinsic value. In this approach, investment decisions are fairly easy to make - if the price of a stock trades below its intrinsic value, it's a good investment. Although this is an oversimplification (fundamental analysis goes beyond just the financial statements) for the purposes of this tutorial, this simple tenet holds true.

Technical traders, on the other hand, believe there is no reason to analyze a company's fundamentals because these are all accounted for in the stock's price. Technicians believe that all the information they need about a stock can be found in its charts.

Time Horizon

Fundamental analysis takes a relatively long-term approach to analyzing the market compared to technical analysis. While technical analysis can be used on a timeframe of weeks, days or even minutes, fundamental analysis often looks at data over a number of years. The different timeframes that these two approaches use is a result of the nature of the investing style to which they each adhere. It can take a long time for a company's value to be reflected in the market, so when a fundamental analyst estimates intrinsic value, a gain is not realized until the stock's market price rises to its "correct" value. This type of investing is called value investing and assumes that the short-term market is wrong, but that the price of a particular stock will correct itself over the long run. This "long run" can represent a timeframe of as long as several years, in some cases.

Furthermore, the numbers that a fundamentalist analyzes are only released over long periods of time. Financial statements are filed quarterly and changes in earnings per share don't emerge on a daily basis like price and volume information. Also remember that fundamentals are the actual characteristics of a business. New management can't implement sweeping changes overnight and it takes time to create new products, marketing campaigns, supply chains, etc. Part of the reason that fundamental analysts use a long-term timeframe, therefore, is because the data they use to analyze

a stock is generated much more slowly than the price and volume data used by technical analysts.

Trading Versus Investing

Not only is technical analysis more short term in nature than fundamental analysis, but the goals of a purchase (or sale) of a stock are usually different for each approach. In general, technical analysis is used for a trade, whereas fundamental analysis is used to make an investment. Investors buy assets they believe can increase in value, while traders buy assets they believe they can sell to somebody else at a greater price. The line between a trade and an investment can be blurry, but it does characterize a difference between the two schools.

The Critics

Some critics see technical analysis as a form of *black magic*. Don't be surprised to see them question the validity of the discipline to the point where they mock its supporters. In fact, technical analysis has only recently begun to enjoy some mainstream credibility. While most analysts on Wall Street focus on the fundamental side, just about any major brokerage now employs technical analysts as well.

Can They Co-Exist?

Although technical analysis and fundamental analysis are seen by many as polar opposites - the oil and water of investing - many market participants have experienced great success by combining the two. For example, some fundamental analysts use technical analysis techniques to figure out the best time to enter into an undervalued security. Oftentimes, this situation occurs when the security is severely oversold. By timing entry into a security, the gains on the investment can be greatly improved. Alternatively, some technical traders might look at fundamentals to add strength to a technical signal. For example, if a sell signal is given through technical patterns and indicators, a technical trader might look to reaffirm his or her decision by looking at some key fundamental data. Oftentimes, having both the fundamentals and technicals on your side can provide the best-case scenario for a trade.

While mixing some of the components of technical and fundamental analysis is not well received by the most devoted groups in each school, there are certainly benefits to at least understanding both schools of thought.

Users of technical analysis are most often called technicians or market technicians. Some prefer the term technical market analyst or simply market analyst. An older term, chartist, is sometimes used, but as the discipline has expanded and modernized the use of the term chartist has become less popular.

Combination with other market forecast methods

John Murphy states that the principal sources of information available to technicians are price, volume and open interest. Other data, such as indicators and sentiment analysis, are considered secondary.

However, many technical analysts reach outside pure technical analysis, combining other market forecast methods with their technical work. One advocate for this approach is John Bollinger, who coined the term *rational analysis* in the middle 1980s for the intersection of technical analysis and fundamental analysis. Another such approach, fusion analysis, overlays fundamental analysis with technical, in an attempt to improve portfolio manager performance.

Technical analysis is also often combined with quantitative analysis and economics. For example, neural networks may be used to help identify intermarket relationships. A few market forecasters combine financial astrology with technical analysis. Chris Carolan's article "Autumn Panics and Calendar Phenomenon", which won the Market Technicians Association Dow Award for best technical analysis paper in 1998, demonstrates how technical analysis and lunar cycles can be combined. Calendar phenomena, such as the January effect in the stock market, are generally believed to be caused by tax and accounting related transactions, and are not related to the subject of financial astrology.

2.2.3 EFFICIENT MARKET HYPOTHESIS (EMH)

An investment theory that states it is impossible to "beat the market" because stock market efficiency causes existing share prices to always incorporate and reflect all relevant information. According to the EMH, stocks always trade at their fair value on stock exchanges, making it impossible for investors to either purchase undervalued stocks or sell stocks for inflated prices. As such, it should be impossible to outperform the overall market through expert stock selection or market timing, and that the only way an investor can possibly obtain higher returns is by purchasing riskier investments

In finance, the **efficient-market hypothesis (EMH)**, popularly known as the Random Walk Theory, asserts that financial markets are "informationally efficient". That is, one cannot consistently achieve returns in excess of average market returns on a risk-adjusted basis, given the information publicly available at the time the investment is made.

The efficient markets hypothesis (EMH), is the proposition that current stock prices fully reflect available information about the value of the firm, and there is no way to earn excess profits, (more than the market overall), by using this information. It deals with one of the most fundamental and exciting issues in finance – why prices change in security markets and how those changes take place. It has very important implications for investors as well as for financial managers.

There are three major versions of the hypothesis: "weak", "semi-strong", and "strong". Weak EMH claims that prices on traded assets (e.g., stocks, bonds, or property) already reflect all past publicly available information. Semi-strong EMH claims both that prices reflect all publicly available information and that prices instantly change to reflect new public information. Strong EMH additionally claims that prices instantly reflect even hidden or "insider" information. There is evidence for and against the weak and semi-strong EMHs, while there is powerful evidence against strong EMH.

The validity of the hypothesis has been questioned by critics who blame the belief in rational markets for much of the financial crisis of 2007–2010.

Defenders of the EMH caution that conflating market stability with the EMH is unwarranted; when publicly available information is unstable, the market can be just as unstable

Historical Background

The efficient-market hypothesis was first expressed by Louis Bachelier, a French mathematician, in his 1900 PhD thesis, "The Theory of Speculation". His work was largely ignored until the 1950s; however beginning in the 30s scattered, independent work corroborated his thesis. A small number of studies indicated that US stock prices and related financial series followed a random walk model. Research by Alfred Cowles in the '30s and '40s suggested that professional investors were in general unable to outperform the market.

The efficient-market hypothesis was developed by Professor Eugene Fama at the University Of Chicago Booth School Of Business as an academic concept of study through his published Ph.D. thesis in the early 1960s at the same school. It was widely accepted up until the 1990s, when behavioral finance economists, who had been a fringe element, became mainstream. Empirical analyses have consistently found problems with the efficient-market hypothesis, the most consistent being that stocks with low price to earnings (and similarly, low price to cash-flow or book value) outperform other stocks. Alternative theories have proposed that cognitive biases cause these inefficiencies, leading investors to purchase overpriced growth stocks rather than value stocks.

The efficient-market hypothesis emerged as a prominent theory in the mid-1960s. Paul Samuelson had begun to circulate Bachelier's work among economists. In 1964 Bachelier's dissertation along with the empirical studies mentioned above were published in an anthology edited by Paul Cootner. In 1965 Eugene Fama published his dissertation arguing for the random walk hypothesis, and Samuelson published a proof for a version of the efficient-market hypothesis. In 1970 Fama published a review of both the theory and the evidence for the hypothesis. The paper extended and refined the theory,

included the definitions for three forms of financial market efficiency: weak, semi-strong and strong (see below).

Further to this evidence that the UK stock market is weak-form efficient, other studies of capital markets have pointed toward their being semi-strong-form efficient. A study by Khan of the grain futures market indicated semi-strong form efficiency following the release of large trader position information (Khan, 1986). Studies by Firth (1976, 1979, and 1980) in the United Kingdom have compared the share prices existing after a takeover announcement with the bid offer. Firth found that the share prices were fully and instantaneously adjusted to their correct levels, thus concluding that the UK stock market was semi-strong-form efficient. However, the market's ability to efficiently respond to a short term, widely publicized event such as a takeover announcement does not necessarily prove market efficiency related to other more long term, amorphous factors. David Dreman has criticized the evidence provided by this instant "efficient" response, pointing out that an immediate response is not necessarily efficient, and that the long-term performance of the stock in response to certain movements are better indications. A study on stocks response to dividend cuts or increases over three years found that after an announcement of a dividend cut, stocks underperformed the market by 15.3% for the three-year period, while stocks outperformed 24.8% for the three years afterward after a dividend increase announcement.

Many investors try to identify securities that are undervalued, and are expected to

Increase in value in the future, and particularly those that will increase more than others. Moreover, many investors, including investment managers, believe that they can select securities that will outperform the market.

They use a variety of forecasting and valuation techniques to aid them in their investment decisions. Obviously, any edge that an investor possesses can be translated into substantial profits. If a manager of a mutual fund with \$10 billion in assets can increase the fund's return, after transaction costs, by 1/10th of 1 percent, this would result in a \$10 million gain. The EMH asserts

that none these techniques are effective (i.e., the advantage gained does not exceed the transaction and research costs incurred), and therefore no one can predictably outperform the market.

Arguably, no other theory in economics or finance generates more passionate discussion between its challengers and proponents. For example, noted Harvard financial economist Michael Jensen writes “there is no other proposition in economics which has more solid empirical evidence supporting it than the Efficient Market Hypothesis,” while investment maven Peter Lynch claims “Efficient markets? That’s a bunch of junk, crazy stuff (Fortune, April 1995).

The efficient markets hypothesis (EMH) suggests that profiting from predicting price movements is very difficult and unlikely. The main engine behind price changes is the arrival of new information. A market is said to be “efficient” if prices adjust quickly and, on average, without bias, to new information. As a result, the current prices of securities reflect all available information at any given point in time. Consequently, there is no reason to believe that prices are too high or too low. Security prices adjust before an investor has time to trade on and profit from a new a piece of information.

The key reason for the existence of an efficient market is the intense competition among investors to profit from any new information. The ability to identify over- and underpriced stocks is very valuable (it would allow investors to buy some stocks for less than their “true” value and sell others for more than they were worth). Consequently, many people spend a significant amount of time and resources in an effort to detect priced" stocks. Naturally, as more and more analysts compete against each other in their effort to take advantage of over- and under-valued securities, the likelihood of being able to find and exploit such mispriced securities becomes smaller and smaller. In equilibrium, only a relatively small number of analysts will be able to profit from the detection of mispriced securities, mostly by chance. For the vast majority of investors, the information analysis payoff would likely not outweigh the transaction costs.

The most crucial implication of the EMH can be put in the form of a slogan: *Trust market prices!* At any point in time, prices of securities in efficient markets reflect all known information available to investors. There is no room for fooling investors, and as a result, all investments in efficient markets are *fairly priced*, i.e. on average investors get exactly what they pay for. Fair pricing of all securities does not mean that they will all perform similarly, or that even the likelihood of rising or falling in price is the same for all securities. According to capital markets theory, the expected return from a security is primarily a function of its risk. The price of the security reflects the present value of its expected future cash flows, which incorporates many factors such as volatility, liquidity, and risk of bankruptcy.

However, while prices are rationally based, changes in prices are expected to be random and unpredictable, because new information, by its very nature, is unpredictable.

Therefore stock prices are said to follow a **random walk**.

Theoretical background

Beyond the normal utility maximizing agents, the efficient-market hypothesis requires that agents have rational expectations, that on average the population is correct (even if no person is) and whenever new relevant information appears, the agents update their expectations appropriately. Note that it is not required that the agents be rational. EMH allows that when faced with new information, some investors may overreact and some may under react. All that is required by the EMH is that investors' reactions be random and follow a normal distribution pattern so that the net effect on market prices cannot be reliably exploited to make an abnormal profit, especially when considering transaction costs (including commissions and spreads). Thus, any person can be wrong about the market—indeed, everyone can be—but the market as a whole is always right. There are three common forms in which the efficient-market hypothesis is commonly stated: **weak-form efficiency**, **semi-strong-form efficiency** and **strong-form efficiency**, each of which has different implications for how markets work.

In **weak-form efficiency**, future prices cannot be predicted by analyzing prices from the past. Excess returns cannot be earned *in the long run* by using investment strategies based on historical share prices or other historical data. Technical analysis techniques will not be able to consistently produce excess returns, though some forms of fundamental analysis may still provide excess returns. Share prices exhibit no serial dependencies, meaning that there are no "patterns" to asset prices. This implies that future price movements are determined entirely by information not contained in the price series. Hence, prices must follow a random walk. This 'soft' EMH does not require that prices remain at or near equilibrium, but only that market participants not be able to *systematically* profit from market 'inefficiencies'. However, while EMH predicts that all price movement (in the absence of change in fundamental information) is random (i.e., non-trending), many studies have shown a marked tendency for the stock markets to trend over time periods of weeks or longer and that, moreover, there is a positive correlation between degree of trending and length of time period studied (but note that over long time periods, the trending is sinusoidal in appearance). Various explanations for such large and apparently non-random price movements have been promulgated. But the best explanation seems to be that the distribution of stock market prices is non-Gaussian (in which case EMH, in any of its current forms, would not be strictly applicable).

The problem of algorithmically constructing prices which reflect all available information has been studied extensively in the field of computer science. For example, the complexity of finding the arbitrage opportunities in pair betting markets has been shown to be NP-hard.

In **semi-strong-form efficiency**, it is implied that share prices adjust to publicly available new information very rapidly and in an unbiased fashion, such that no excess returns can be earned by trading on that information. Semi-strong-form efficiency implies that neither fundamental analysis nor technical analysis techniques will be able to reliably produce excess returns. To test for semi-strong-form efficiency, the adjustments to previously unknown news must be of a reasonable size and must be instantaneous. To test for

this, consistent upward or downward adjustments after the initial change must be looked for. If there are any such adjustments it would suggest that investors had interpreted the information in a biased fashion and hence in an inefficient manner.

In ***strong-form efficiency***, share prices reflect all information, public and private, and no one can earn excess returns. If there are legal barriers to private information becoming public, as with insider trading laws, strong-form efficiency is impossible, except in the case where the laws are universally ignored. To test for strong-form efficiency, a market needs to exist where investors cannot consistently earn excess returns over a long period of time. Even if some money managers are consistently observed to beat the market, no refutation even of strong-form efficiency follows: with hundreds of thousands of fund managers worldwide, even a normal distribution of returns (as efficiency predicts) should be expected to produce a few dozen "star" performers.

Criticism and behavioral finance

Investors and researchers have disputed the efficient-market hypothesis both empirically and theoretically. Behavioral economists attribute the imperfections in financial markets to a combination of cognitive biases such as overconfidence, overreaction, representative bias, information bias, and various other predictable human errors in reasoning and information processing. These have been researched by psychologists such as Daniel Kahneman, Amos Tversky, Richard Thaler, and Paul Slovic. These errors in reasoning lead most investors to avoid value stocks and buy growth stocks at expensive prices, which allow those who reason correctly to profit from bargains in neglected value stocks and the overreacted selling of growth stocks.

Empirical evidence has been mixed, but has generally not supported strong forms of the efficient-market hypothesis. According to Dreman, in a 1995 paper, low P/E stocks have greater returns. In an earlier paper he also refuted the assertion by Ray Ball that these higher returns could be attributed to

higher beta, whose research had been accepted by efficient market theorists as explaining the anomaly in neat accordance with modern portfolio theory.

One can identify "losers" as stocks that have had poor returns over some number of past years. "Winners" would be those stocks that had high returns over a similar period. The main result of one such study is that losers have much higher average returns than winners over the following period of the same number of years. A later study showed that beta (β) cannot account for this difference in average returns. This tendency of returns to reverse over long horizons (i.e., losers become winners) is yet another contradiction of EMH. Losers would have to have much higher betas than winners in order to justify the return difference. The study showed that the beta difference required to save the EMH is just not there.

Speculative economic bubbles are an obvious anomaly, in that the market often appears to be driven by buyers operating on irrational exuberance, who take little notice of underlying value. These bubbles are typically followed by an overreaction of frantic selling, allowing shrewd investors to buy stocks at bargain prices. Rational investors have difficulty profiting by shorting irrational bubbles because, as John Maynard Keynes commented, "Markets can remain irrational far longer than you or I can remain solvent." Sudden market crashes as happened on Black Monday in 1987 are mysterious from the perspective of efficient markets, but allowed as a rare *statistical event* under the Weak-form of EMH.

Burton Malkil, a well-known proponent of the general validity of EMH, has warned that certain emerging markets such as China are not empirically efficient; that the Shanghai and Shenzhen markets, unlike markets in United States, exhibit considerable serial correlation (price trends), non-random walk, and evidence of manipulation.

Behavioral psychology approaches to stock market trading are among some of the more promising alternatives to EMH (and some investment strategies seek to exploit exactly such inefficiencies). But Nobel Laureate co-founder of the program Daniel Kahneman announced his skepticism of investors beating

the market: "They're [investors] just not going to do it [beat the market]. It's just not going to happen." Indeed defenders of EMH maintain that Behavioral Finance strengthens the case for EMH in that BF highlights biases in individuals and committees and not competitive markets. For example, one prominent finding in Behavioral Finance is that individuals employ hyperbolic discounting. It is palpably true that bonds, mortgages, annuities and other similar financial instruments subject to competitive market forces do not. Any manifestation of hyperbolic discounting in the pricing of these obligations would invite arbitrage thereby quickly eliminating any vestige of individual biases. Similarly, diversification, derivative securities and other hedging strategies assuage if not eliminate potential mispricing from the severe risk-intolerance (loss aversion) of individuals underscored by behavioral finance. On the other hand, economists, behavioral psychologists and mutual fund managers are drawn from the human population and are therefore subject to the biases that behavioralists showcase. By contrast, the price signals in markets are far less subject to individual biases highlighted by the Behavioral Finance program. Richard Thaler has started a fund based on his research on cognitive biases. In a 2008 report he identified complexity and herd behavior as central to the global financial crisis of 2008.

Further empirical work has highlighted the impact transaction costs have on the concept of market efficiency, with much evidence suggesting that any anomalies pertaining to market inefficiencies are the result of a cost benefit analysis made by those willing to incur the cost of acquiring the valuable information in order to trade on it. Additionally the concept of liquidity is a critical component to capturing "inefficiencies" in tests for abnormal returns. Any test of this proposition faces the joint hypothesis problem, where it is impossible to ever test for market efficiency, since to do so requires the use of a measuring stick against which abnormal returns are compared - one cannot know if the market is efficient if one does not know if a model correctly stipulates the required rate of return. Consequently, a situation arises where either the asset pricing model is incorrect or the market is inefficient, but one has no way of knowing which the case is

A key work on random walk was done in the late 1980s by Profs. Andrew Lo and Craig MacKinlay; they effectively argue that a random walk does not exist, nor ever has. Their paper took almost two years to be accepted by academia and in 2001 they published "A Non-random Walk down Wall St." which explained the paper in layman's terms.

Economists Matthew Bishop and Michael Green claim that full acceptance of the hypothesis goes against the thinking of Adam Smith and John Maynard Keynes, who both believed irrational behavior, had a real impact on the markets.

Warren Buffet has also argued against EMH, saying the preponderance of value investors among the world's best money managers rebuts the claim of EMH proponents that luck is the reason some investors appear more successful than others.

Recent financial crisis

The recent global financial crisis has led to renewed scrutiny and criticism of the hypothesis. Market strategist Jeremy Grantham has stated flatly that the EMH is responsible for the current financial crisis, claiming that belief in the hypothesis caused financial leaders to have a "chronic underestimation of the dangers of asset bubbles breaking". Noted financial journalist Roger Lowenstein blasted the theory, declaring "The upside of the current Great Recession is that it could drive a stake through the heart of the academic nostrum known as the efficient-market hypothesis."

At the International Organization of Securities Commissions annual conference, held in June 2009, the hypothesis took center stage. Martin Wolf, the chief economics commentator for the Financial Times, dismissed the hypothesis as being a useless way to examine how markets function in reality. Paul McCulley, managing director of PIMCO, was less extreme in his criticism, saying that the hypothesis had not failed, but was "seriously flawed" in its neglect of human nature.

The financial crisis has led Richard Posner, a prominent judge, University of Chicago law professor, and innovator in the field of Law and Economics, to back away from the hypothesis and express some degree of belief in Keynesian economics. Posner accused some of his 'Chicago School' colleagues of being "asleep at the switch", saying that "the movement to deregulate the financial industry went too far by exaggerating the resilience - the self healing powers - of laissez-faire capitalism." Others, such as Fama himself, said that the hypothesis held up well during the crisis and that the markets were a casualty of the recession, not the cause of it.

COMMON MISCONCEPTIONS ABOUT THE EMH

Despite its relative simplicity, this hypothesis has also generated a lot of controversy. After all, the EMH questions the ability of investors to consistently detect mispriced securities. Not surprisingly, this implication does not sit very well with many financial analysts and active portfolio managers.

Arguably, in liquid markets with many participants, such as stock markets, prices should adjust quickly to new information in an unbiased manner. However, much of the criticism leveled at the EMH is based on numerous misconceptions, incorrect

Interpretations and myths about the theory of efficient markets. We present some of the most persistent "myths" about the EMH below.

Myth 1: EMH claims that investors cannot outperform the market. Yet we can see that some of the successful analysts (such as George Soros, Warren Buffett, or Peter Lynch) are able to do exactly that. Therefore, EMH must be incorrect

.

EMH does not imply that investors are unable to outperform the market. We know that the constant arrival of information makes prices fluctuate. It is possible for an investor to "make a killing" if newly released information causes the price of the security the investor owns to substantially increase. What EMH does claim, though, is that one should not be expected to outperform the market predictably or consistently.

It should be noted, though, that **some** investors could outperform the market for a very long time by chance alone, even if markets are efficient. Imagine, for the sake of simplicity, that an investor who picks stocks “randomly” has a 50% chance of “beating the market”. For such an investor, the chance of outperforming the market in each and every of the next ten years is then (0.5), or about one-tenth of one percent. However, the chance that there will be *at least one investor* outperforming the market in each of the next 10 years sharply increases as the number of investors trying to do exactly that rises.

In a group of 1,000 investors, the probability of finding one “ultimate winner” with a perfect 10-year record is 63%. With a group of 10,000 investors, the chance of seeing at least one who outperforms the market in every of next ten years is 99.99%, a virtual certainty. Each individual investor may have dismal odds of beating the market for the next 10 years. Yet the likelihood of, after the ten years, finding one very successful investor, even if he or she is investing purely randomly – is very high if there are a sufficiently large number of investors. This is the case with the state lottery, in which the probability of a *given* individual winning is virtually zero, but the probability that *someone* will win is very high. The existence of a handful of successful investors such as Messrs. Soros, Buffett, and Lynch is an expected outcome in a completely random distribution of investors. The theory would only be threatened if you could identify who those successful investors would be *prior* to their performance, rather than after the fact.

Myth 2: EMH claims that financial analysis is pointless and investors who attempt to research security prices are wasting their time. “Throwing darts at the financial page will produce a portfolio that can be expected to do as well as any managed by Professional security analysts”. Yet we tend to see that financial analysts are not “Driven out of market”, which means that their services are valuable. Therefore, EMH must be incorrect.

There are two principal counter-arguments against the equivalency of “dart-throwing” and professional analysis strategies. First, investors generally have different “tastes”, some may; for example, prefer to put their money in high-risk “hi-tech” firm portfolios, while others may like less risky investment strategies. Optimal portfolios should provide the investor with the combination of return and risk that the investor finds desirable. A randomly chosen portfolio may not accomplish this goal. Second, and more importantly, financial analysis is far from pointless in efficient capital markets. The competition among investors who actively seek and analyze new information with the goal to identify and take advantage of mispriced stocks is truly essential for the existence of efficient capital markets. In fact, one can say that financial analysis is actually the engine that enables incoming information to get quickly reflected into security prices.

So why don't all investors find it optimal to search for profits by performing financial Analysis? The answer is simple – financial research is very costly. As we have already discussed, financial analysts have to be able to gather, process, and evaluate vast amounts of information about firms, industries, scientific achievements, the economy, etc. They have to invest a lot of time and effort in sophisticated analysis, as well as many resources into data gathering, purchases of computers, software. In addition, analysts who frequently trade securities incur various transaction costs, including brokerage costs, bid ask spread, and market impact costs.

Therefore, any profits achieved by the analysts while trading on "mispriced" securities must be reduced by the *costs of financial analysis*, as well as the *transaction costs* involved. For mutual funds and private investment managers these costs are passed on to investors as fees, loads, and reduced returns. There is some evidence that some professional investment managers are able to improve performance through their analyses. However, this may be by pure chance. In general, the advantage gained is not sufficient to outweigh the cost of their advice.

In equilibrium, there will be only as many financial analysts in the market as optimal to insure that, on average, the incurred costs are covered by the achieved gross trading profits. For the majority of other investors, the chasing of "mispriced" stocks would indeed be pointless and they should stick with passive investment, such as with index mutual funds.

Myth 3: EMH claims that new information is always fully reflected in market prices. Yet one can observe prices fluctuating (sometimes very dramatically) every day, hour, and minute. Therefore, EMH must be incorrect.

The constant fluctuation of market prices can be viewed as an indication that markets *are* efficient. New information affecting the value of securities arrives constantly, causing continuous adjustment of prices to information updates. In fact, observing that prices *did not* change would be inconsistent with market efficiency, since we know that relevant information is arriving almost continuously.

Over-reaction and Under-reaction

The efficient markets hypothesis implies that investors react quickly and in an unbiased manner to new information. In two widely publicized studies, DeBondt and Thaler present contradictory evidence. They find that stocks with low long-term past returns tend to have higher future returns and vice versa - stocks with high long-term past returns tend to have lower future returns (long-term reversals).

These findings received significant publicity in the popular press, which ran numerous headlines touting the benefits of these so-called contrarian strategies. The results appear to be inconsistent with the EMH. However, they have not survived the test of time. Although the issues are complex, recent research indicates that the findings might be the result of methodological problems arising from the measurement of risk. Once risk is measured correctly, the findings tend to disappear.

One of the most enduring anomalies documented in the finance literature is the empirical observation that stock prices appear to respond to earnings for about a year after they are announced. Prices of companies experiencing positive earnings surprises tend to drift upward, while prices of stocks experiencing negative earnings surprises tend to drift downward. This “post-earnings-announcement drift” was first noted by Ball and Brown in 1968 and has since been replicated by numerous studies over different time periods and in different countries. After more than thirty years of research, this anomaly has yet to be explained.

Another study reported that stocks with high returns over the past year tended to have high returns over the following three to six months (short-term momentum in stock prices). This “momentum” effect is a fairly new anomaly and consequently significantly more research is needed on the topic. However, the effect is present in other countries and has persisted throughout the 1900s.

A variety of other anomalies have been reported. Some indicate market over-reaction to information, and others under-reaction. Some of these findings are simply related to chance: if you analyze the data enough, you will find some patterns. Dredging for anomalies is a rewarding occupation. Some apparent anomalies, such as the long-term reversals of DeBondt and Thaler, may be a by-product of rational (efficient) pricing.

This is not evident until alternative explanations are examined by appropriate analysis.

Value versus growth

A number of investment professionals and academics argue that so called “value strategies” are able to outperform the market consistently. Typically, value strategies involve buying stocks that have low prices relative to their accounting “book” values, dividends, or historical prices. In a provocative study, Lakonishok, Schleifer, and Vishny find evidence that the difference in

average returns between stocks with low price-to-book ratios (“value stocks”) and stocks with high price-to-book ratios (“glamour stocks”) was as high as 10 percent year. Surprisingly, this return differential cannot be attributed to higher risk (as measured by volatility) - value stocks are typically no riskier than glamour stocks. Rather, the authors argue, market participants consistently overestimate the future growth rates of glamour stocks relative to value stocks.

Consequently, these results may represent strong evidence against the EMH. It was also interesting that nearly the entire advantage of the value stocks occurred in January each year. However, current research indicates that the anomalous returns may be caused by a selection bias in a popular commercial database used by financial economists.

Small Firm Effect

Rolf Banz uncovered another puzzling anomaly in 1981. He found that average returns on small stocks were too large to be justified by the Capital Asset Pricing Model, while the average returns on large stocks were too low. Subsequent research indicated that most of the difference in returns between small and large stocks occurred in the month of January. The results were particularly surprising because for years financial economists had accepted that systematic risk or Beta was the single variable for predicting returns. Current research indicates that this finding is not evidence of market inefficiency, but rather indicates a failure of the Capital Asset Pricing Model.

Conclusions

The goal of all investors is to achieve the highest returns possible. Indeed, each year investment professionals publish numerous books touting ways to beat the market and earn millions of dollars in the process. Unfortunately for these so-called “investment gurus”, these investment strategies fail to perform as predicted. The intense competition between investors creates an efficient market in which prices adjust rapidly to new information. Consequently, on

average, investors receive a return that compensates them for the time value of money and the risks that they bear – nothing more and nothing less. In other words, after taking risk and transaction costs into account, active security management is a losing proposition.

Although no theory is perfect, the overwhelming majority of empirical evidence supports the efficient market hypothesis. The vast majority of students of the market agree that the markets are highly efficient. The opponents of the efficient markets hypothesis point to some recent evidence suggesting that there is under- and over-reaction in security markets. However, it's important to note that these studies are controversial and generally have not survived the test of time. Ultimately, the efficient markets hypothesis continues to be the best description of price movements in securities markets.

2.2.4 Elliott wave principle

The ***Elliott Wave Principle*** is a form of technical analysis that investors use to forecast trends in the financial markets by identifying extremes in investor psychology, highs and lows in prices, and other collective activities. Ralph Nelson Elliott (1871–1948), a professional accountant, developed the concept in the 1930s. He proposed that market prices unfold in specific patterns, which practitioners today call Elliott waves, or simply waves. Elliott published his theory of market behavior in the book *The Wave Principle* (1938), in a series of articles in *Financial World* magazine in 1939, and most fully in his final major work, *Nature's Laws the Secret of the Universe* (1946). Elliott said that "because man is subject to rhythmical procedure, calculations having to do with his activities can be projected far into the future with a justification and certainty heretofore unattainable."

The Elliott Wave Principle is a detailed description of how groups of people behave. It reveals that mass psychology swings from pessimism to optimism and back in a natural sequence, creating specific and measurable patterns. The Elliot Wave Theory represents a development of the well-known Dow Theory After he had retired and a serious illness had been discovered in his

organism, Elliott started to observe stock markets and their charts in the hope of understanding the market behavior. After he had performed a large work, he concluded that the market, being a product of predominant psychology of the masses, followed some laws.

One of the easiest places to see the Elliott Wave Principle at work is in the financial markets, where changing investor psychology is recorded in the form of price movements. If you can identify repeating patterns in prices, and figure out where we are in those repeating patterns today, you can predict where we are going.

Elliott Wave Principle measures *investor psychology*, which is the *real* engine behind the stock markets. When people are optimistic about the future of a given issue, they bid the price up.

Two observations will help us grasp this: First, for hundreds of years, investors have noticed that events external to the stock markets seem to have no consistent effect on their progress. The same news that today seems to drive the markets *up* are as likely to drive them *down* tomorrow. The only reasonable conclusion is that the markets simply do not react consistently to outside events. Second, when you study historical charts, you see that the markets continuously unfold in *waves*.

Using the Elliott Wave Principle is an exercise in probability. An Elliottician is someone who is able to identify the markets structure and anticipate the most likely next move based on our position within those structures. By knowing the wave patterns, we'll know what the markets are likely to do next and (sometimes most importantly) what they will *not* do next. By using the Elliott Wave Principle, we identify the highest probable moves with the least risk.

Market Predictions Based on Wave Patterns

Elliott made detailed stock market predictions based on unique characteristics he discovered in the wave patterns. An impulsive wave, which goes with the main trend, always shows five waves in its pattern. On a smaller scale, within

each of the impulsive waves, five waves can again be found. In this smaller pattern, the same pattern repeats itself ad infinitum. These ever-smaller patterns are labeled as different wave degrees in the Elliott Wave Principle. Only much later were fractals recognized by scientists.

In the financial markets we know that "every action creates an equal and opposite reaction" as a price movement up or down must be followed by a contrary movement. Price action is divided into trends and corrections or sideways movements. Trends show the main direction of prices while corrections move against the trend. Elliott labeled these "impulsive" and "corrective" waves.

Theory Interpretation

The Elliott Wave Theory is interpreted as follows:

- Every action is followed by a reaction.
- Five waves move in the direction of the main trend followed by three corrective waves (a 5-3 move).
- A 5-3 move completes a cycle.
- This 5-3 move then becomes two subdivisions of the next higher 5-3 wave.
- The underlying 5-3 pattern remains constant, though the time span of each may vary.

Let's have a look at the following chart made up of eight waves (five up and three down) labeled 1, 2, 3, 4, 5, A, B and C.

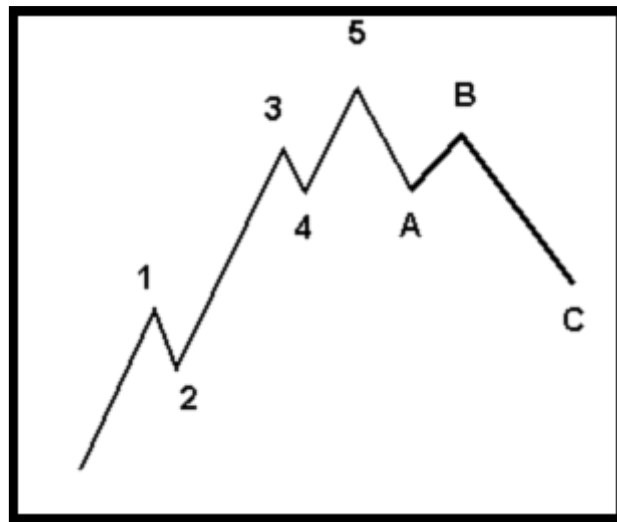


Figure 2.4

We can see that the three waves in the direction of the trend are impulses, so these waves also have five waves within them. The waves against the trend are corrections and are composed of three waves.

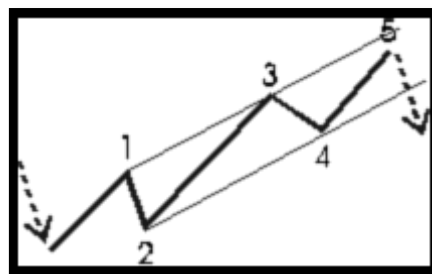


Figure 2.5

Theory Gained Popularity in the 1970s

In the 1970s, this wave principle gained popularity through the work of Frost and Prechter. They published a legendary book on the Elliott Wave entitled “*The Elliott Wave Principle – the Key to Stock Market Profits*”. In this book, the authors predicted the bull market of the 1970s, and Robert Prechter called

the crash of 1987. (For related reading, see *Digging Deeper into Bull and Bear Markets* and *The Greatest Market Crashes*.)

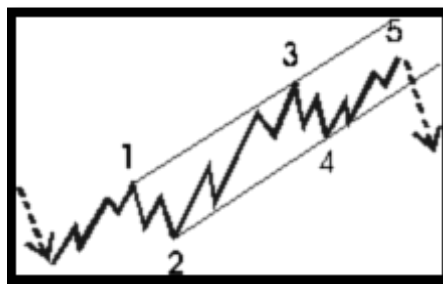


Figure 2.6

The corrective wave formation normally has three distinct price movements - two in the direction of the main correction (A and C) and one against it (B). Waves 2 and 4 in the above picture are corrections. These waves have the following structure:

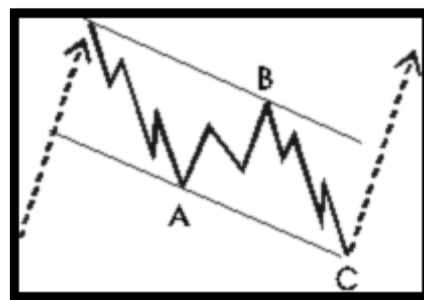


Figure 2.7

Note that waves A and C move in the direction of the shorter-term trend, and therefore are impulsive and composed of five waves, which are shown in the picture above. An impulse-wave formation, followed by a corrective wave, form an Elliott wave degree consisting of trends and countertrends. Although the patterns pictured above are bullish, the same applies for bear markets where the main trend is down.

Series of Wave Categories

The Elliott Wave Theory assigns a series of categories to the waves from largest to smallest. They are:

- Grand Supercycle
- Supercycle
- Cycle
- Primary
- Intermediate
- Minor
- Minute
- Minuette
- Sub-Minuette

To use the theory in everyday trading, the trader determines the main wave, or super cycle goes long and then sells or shorts the position as the pattern runs out of steam and a reversal is imminent.

Elliott proposes, as well, that all price moves on the market are divided into:

- five waves in the direction of the main trend (waves 1 to 5 in Fig. 1);
- Three corrective waves (waves A, B, C in Fig. 1).

The waves are divided into:

- impulses that create a directed trend (bull or bear) and cause the market to move very actively (waves 1, 3, 5, A, C in Fig. 1);
- Corrections (rollbacks) that are characterized by moving against the trend (waves 2, 4, B in Fig. 1).

In his Wave Theory, Elliott was based on the waves subdivision principle. This means that every wave is a part of a longer wave and is subdivided into shorter waves itself (Fig. 2). Every wave is subdivided into 3 or 5 waves. This subdivision depends on the direction of the longer wave.

The main principle in the Elliott's theory is that every impulse wave consists of five shorter waves and every corrective wave (against the trend) is composed of three waves, which can be well seen in Fig. 2. For example, Wave 1 in Fig. 2 is composed of 5 shorter waves since it is an impulse wave that creates the trend.

The longest cycle, according to Elliott, is called Grand Supercycle that is composing of 8 Supercycle waves. The latter ones are each composed of 8 Cycles, etc. For example, Fig. 2 shows 3 basic cycles. It can easily be seen that impulse waves and the subsequent corrective waves are proportional. The stronger impulse is, the stronger correction is, and vice versa.

The Elliott Wave Theory is criticized for there is not always a clear definition of when a wave starts or ends. Corrections are especially difficult in this regard.

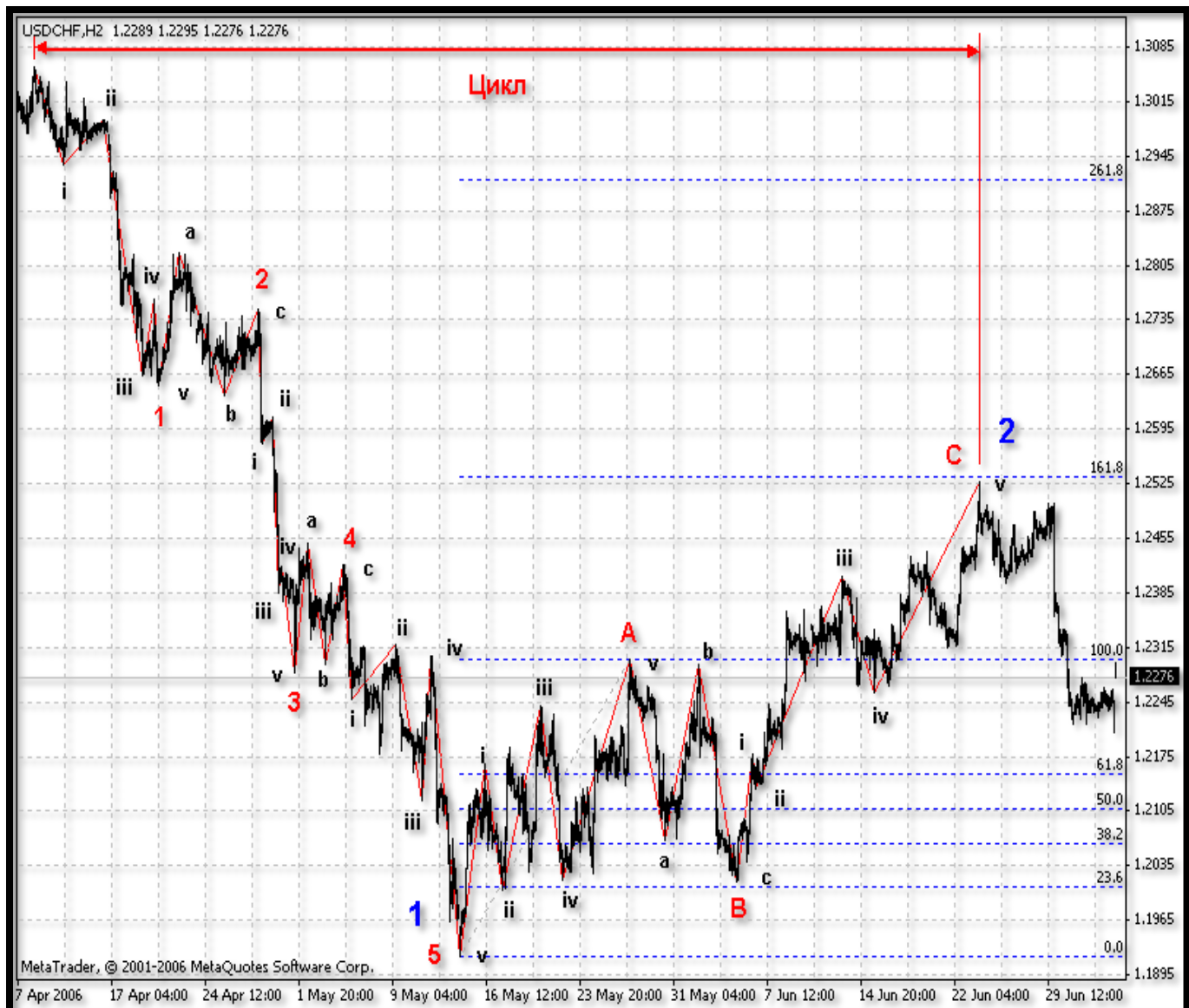
Elliott Wave Theory and Fibonacci Numbers

Fibonacci Numbers provide the mathematical foundation for the Elliott Wave Theory. Fibonacci numbers play an important role in the construction of the complete market cycle described with the Elliott's waves. Each of the cycles Elliott defined are comprised of a total wave count that falls within the Fibonacci number sequence.

Under closer examination of Fig. 2, one can notice that the complete market cycle is composed of two large waves, eight middle waves, and 34 small waves. Similarly, at a bull market, we can see that a bull Grand Supercycle is composed of one large wave, five middle waves, and 21 small waves. If we continue this subdivision, we will be able to observe the consequent 89 even smaller waves, etc.

Respectively, a bear Grand Supercycle is composed of one large wave, three middle waves, and 13 small waves. At the next sublevel, there are 55 very small waves, etc. This principle is normally used in the Elliott Wave Theory as follows: movement in a certain direction should continue until it reaches some point in concordance with the summational Fibonacci number sequence.

For example, if the time, during which the trend does not change, exceeds 3 days, this direction should not reverse until the 5th day begins. Similarly, the trend should continue up to 8 days if it has not changed the direction within 5 days. 9-day trend should not be completed until the 13th day begins, etc. This basic pattern of how the trend movements can be calculated equally applies for both hourly, daily, weekly, or monthly data. However, this is just an



"ideal model", and nobody can expect that prices' behavior will be so definite and predictable. Elliott noted that deviations could happen both in time and in amplitude and individual waves would hardly develop exactly in these regular forms.

Characteristics of Waves

Calculations within the Elliott Wave Theory resemble a road-map. Every wave has a set of characteristics. These characteristics are based on market behavior of masses.

In the Elliott Wave Theory, a special attention is paid to individual description of each wave. Besides, there are certain laws used for proportional formations of Elliott waves (Fig. 3). These laws enable proper definition of where the wave starts and how long it is. The wave lengths are measured from high to low of the corresponding wave.

Wave	Classical Relations between Waves
1	-
2	0.382, 0.5, or 0.618 of Wave 1 length
3	1.618, 0.618, or 2.618 of Wave 1 length
4	0.382 or 0.5 of Wave 1 length
5	0.382, 0.5, or 0.618 of Wave 1 length
A	1, 0.618 or 0.5 of Wave 5 length
B	0.382 or 0.5 of Wave A length
C	1.618, 0.618, or 0.5 of Wave A length

Figure 3

The above classical relations between waves are confirmed by actual ones with a 10%-error. Such error can be explained through short-term influences of some technical or fundamental factors. In whole, the data are rather relative. Important is that all relations between all waves can take values of 0.382, 0.50, 0.618, 1.618. Using this, we can calculate relations between both wave heights and wave lengths. Let us consider characteristics of each wave:

- Wave 1

Happens when the «market psychology» is practically bearish. News are still negative. As a rule, it is very strong if it represents a leap (change from bear trend to the bull trend, penetration into the might resistance level, etc.). In a state of tranquility, it usually demonstrates insignificant price moves in the background of general wavering.

- Wave 2

Happens when the market rapidly rolls back from the recent, hard-won profitable positions. It can roll back to almost 100% of Wave 1, but not below its starting level. It usually makes 60% of Wave 1 and develops in the background of prevailing amount of investors preferring to fix their profits.

- Wave 3

Is what the Elliott's followers live for. Rapid increase of investors' optimism is observed. It is the mightiest and the longest wave of rise (it can never be the shortest) where prices are accelerated and the volumes are increased. A typical Wave 3 exceeds Wave 1 by, at least, 1.618 times, or even more.

- Wave 4

Often difficult to identify. It usually rolls back by no more than 38% of Wave 3. Its depth and length are normally not very significant. Optimistic moods are still prevailing in the market. Wave 4 may not overlap Wave 2 until the five-wave cycle is a part of the end triangle.

- Wave 5

Is often identified using momentum divergences. The prices increase at middle-sized trade volumes. The wave is formed in the background of mass agiotage. At the end of the wave, the trade volumes often rise sharply.

- Wave A

Many traders still consider the rise to make a sharp come-back. But there appear some traders sure of the contrary. Characteristics of this wave are often very much the same as those of Wave 1.

- Wave B

Often resembles Wave 4 very much and is very difficult to identify. Shows insignificant movements upwards on the rests of optimism.

- Wave C

A strong decreasing wave in the background of general persuasion that a new, decreasing trend has started. In the meantime, some investors start buying cautiously. This wave is characterized by high momentum (five waves) and lengthiness up to 1.618-fold Wave 3.

Unfortunately, Elliott's waves are well observed in the "old" market, but they are rather dimmed for the future. This is why practical use of the Elliott Wave Theory is often difficult and requires special knowledge.

Criticism

The premise that markets unfold in recognizable patterns contradicts the efficient market hypothesis, which says that prices cannot be predicted from market data such as moving averages and volume. By this reasoning, if successful market forecasts were possible, investors would buy (or sell) when

the method predicted a price increase (or decrease), to the point that prices would rise (or fall) immediately, thus destroying the profitability and predictive power of the method. In efficient markets, knowledge of the Elliott wave principle among investors would lead to the disappearance of the very patterns they tried to anticipate, rendering the method, and all forms of technical analysis, useless.

Benoit Mandelbrot has questioned whether Elliott waves can predict financial markets: "But Wave prediction is a very uncertain business. It is an art to which the subjective judgment of the chartists matters more than the objective, replicable verdict of the numbers. The record of this, as of most technical analysis, is at best mixed."

Robert Prechter had previously said that ideas in an article by Mandelbrot "originated with Ralph Nelson Elliott, who put them forth more comprehensively and more accurately with respect to real-world markets in his 1938 book *The Wave Principle*." Critics also say the wave principle is too vague to be useful, since it cannot consistently identify when a wave begins or ends, and that Elliott wave forecasts are prone to subjective revision. Some who advocate technical analysis of markets have questioned the value of Elliott wave analysis. Technical analyst David Aronson wrote:

The Elliott Wave Principle, as popularly practiced, is not a legitimate theory, but a story, and a compelling one that is eloquently told by Robert Prechter. The account is especially persuasive because EWP has the seemingly remarkable ability to fit any segment of market history down to its most minute fluctuations. I contend this is made possible by the method's loosely defined rules and the ability to postulate a large number of nested waves of varying magnitude. This gives the Elliott analyst the same freedom and flexibility that allowed pre-Copernican astronomers to explain all observed planet movements even though their underlying theory of an Earth-centered universe was wrong.

CHAPTER 3

Artificial Neural Networks

3.1 THE BIOLOGICAL EXAMPLE

Traditionally, the term neural network has been used to refer to a system of biological neurons.

A neuron is an electrically excitable cell that processes and transmits information by electrical and chemical signaling. Chemical signaling occurs via synapses (specialized connections with other cells). Neurons are the core components of the nervous system, which includes the brain, spinal cord, and peripheral ganglia.

A number of specialized types of neurons exist: sensory neurons respond to touch, sound, light and numerous other stimuli affecting cells of the sensory organs that then send signals to the spinal cord and brain. Motor neurons receive signals from the brain and spinal cord and cause muscle contractions and affect glands. Interneurons connect neurons to other neurons within the same region of the brain or spinal cord.

A typical neuron possesses a cell body (often called the soma), dendrites, and an axon:

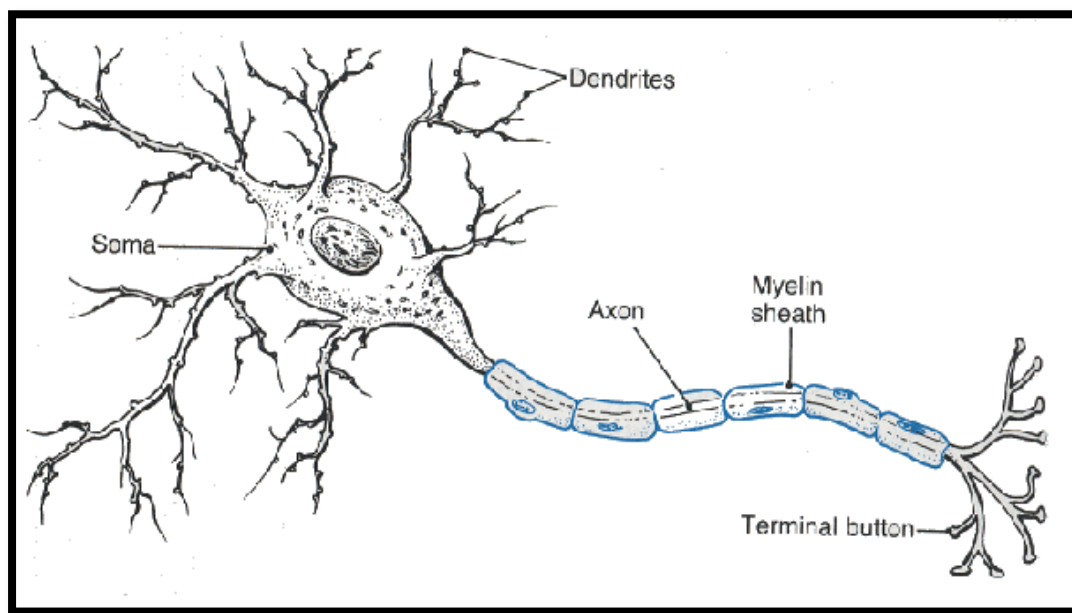


Figure 1: Typical neuron

Dendrites are filaments that arise from the cell body, often extending for hundreds of microns and branching multiple times, giving rise to a complex "dendritic tree". An axon is a special cellular filament that arises from the cell body at a site called the axon hillock and travels for a distance, as far as 1 m in humans or even more in other species. The cell body of a neuron frequently gives rise to multiple dendrites, but never to more than one axon, although the axon may branch hundreds of times before it terminates. At the majority of synapses, signals are sent from the axon of one neuron to a dendrite of another. There are, however, many exceptions to these rules: neurons that lack dendrites, neurons that have no axon, synapses that connect an axon to another axon or a dendrite to another dendrite, etc.

All neurons are electrically excitable, maintaining voltage gradients across their membranes by means of metabolically driven ion pumps, which combine with ion channels embedded in the membrane to generate intracellular-versus-extracellular concentration differences of ions such as sodium, potassium, chloride, and calcium. Changes in the cross-membrane voltage can alter the function of voltage-dependent ion channels. If the voltage changes by a large enough amount, an all-or-none electrochemical pulse

called an action potential is generated, which travels rapidly along the cell's axon, and activates synaptic connections with other cells when it arrives.

The number of neurons in the brain varies dramatically from species to species. One estimate puts the human brain at about 100 billion (10^{11}) neurons and 100 trillion (10^{14}) synapses. Another estimate is 86 billion neurons of which 16.3 billion are in the cerebral cortex and 69 billion in the cerebellum. By contrast, the nematode worm *Caenorhabditis elegans* has just 302 neurons making it an ideal experimental subject as scientists have been able to map all of the organism's neurons. The fruit fly *Drosophila melanogaster*, a common subject in biology experiments, has around 100,000 neurons and exhibits many complex behaviors. Many properties of neurons, from the type of neurotransmitters used to ion channel composition, are maintained across species, allowing scientists to study processes occurring in more complex organisms in much simpler experimental systems.

Neurons connect to each other to form networks: Simply, neural networks are groups of selected neurons that are connected with one another. Neural networks are functional circuits in the brain that process information and create useful activities by sending outputs to the body.

Some neurons transmit this neural network information over a distance much the same as information is transmitted over a telephone line. Electrical impulses, called action potentials, are used by neurons to transmit information over the distance from the cell body to the end of the fiber. Nerve fibers from one neuron in a network are connected to others in the network by synapses, the tiny gaps between the end of one neuron and the beginning of another. The fibers do not make direct contact with the other neuron at a synapse. Instead, each impulse triggers release of a chemical substance from the end of the fiber. The chemical, called a neurotransmitter, then carries the information across the synaptic space on to the next neuron. Neurons make and release over 50 different kinds of neurotransmitters. Some neurotransmitters, called excitatory neurotransmitters, start a new electrical impulse in the neuron at the other side of the synapse. Other

neurotransmitters prevent impulses from occurring in the neuron at the other side of the synapse. These are called inhibitory neurotransmitters.

Large numbers of neurons are interconnected by synapses to form a network. Depending on the job to be done, a network can have a few hundred to more than a million neurons. Each neuron in the network may receive a synaptic input from one hundred or more other neurons in the network. Moreover, one network may have connections for sharing information with other networks. For example, neural networks in the brain share information with networks in the spinal cord.

Neural networks are responsible for the basic functions of our nervous system. They determine how we behave as individuals.

Our emotions experienced as fear, anger, and what we enjoy in life come from neural networks in the brain.

Even our ability to think and store memories depends on neural networks. Neural networks in the brain and spinal cord, program all our movements including how fast we can type on a computer keyboard to how well we play sports. Our ability to see or hear is disturbed if something happens to the neural networks for vision or hearing in the brain. Neural networks also control important functions of our bodies. Keeping a constant body temperature and blood pressure are examples where neural networks operate automatically to make our bodies work without us knowing what the networks are doing. These are called autonomic functions of neural networks because they are automatic and occur continuously without us being aware of them.

3.2 HISTORICAL BACKGROUND

Research in the field of neural networks has been attracting increasing attention in recent years. Since 1943, when Warren McCulloch and Walter Pitts presented the first model of artificial neurons, new and more sophisticated proposals have been made from decade to decade. Mathematical analysis has solved some of the mysteries posed by the new models but has left many questions open for future investigations. Needless to say, the study of neurons, their interconnections, and their role as the

brain's elementary building blocks is one of the most dynamic and important research fields in modern biology. We can illustrate the relevance of this endeavor by pointing out that between 1901 and 1991 approximately ten percent of the Nobel Prizes for Physiology and Medicine were awarded to scientists who contributed to the understanding of the brain. It is not an exaggeration to say that we have learned more about the nervous system in the last fifty years than ever before.

Artificial Neural networks are adaptive statistical models based on an analogy with the structure of the brain. They are adaptive because they can learn to estimate the parameters of some population using a small number of exemplars (one or a few) at a time. They do not differ essentially from standard statistical models. For example, one can find neural network architectures akin to discriminant analysis, principal component analysis, logistic regression, and other techniques. In fact, the same mathematical tools can be used to analyze standard statistical models and neural networks. Neural networks are used as statistical tools in a variety of fields, including psychology, statistics, engineering, econometrics, and even physics. They are used also as models of cognitive processes by neuro and cognitive scientists.

Basically, neural networks are built from simple units, sometimes called neurons or cells by analogy with the real thing. These units are linked by a set of weighted connections (synapses). Learning is usually accomplished by modification of the connection weights. Each unit codes or corresponds to a feature or a characteristic of a pattern that we want to analyze or that we want to use as a predictor.

These networks usually organize their units into several layers. The first layer is called the input layer, the last one the output layer. The intermediate layers (if any) are called hidden layers. The information to be analyzed is fed to the neurons of the first layer and then propagated to the neurons of the second layer for further processing. The result of this processing is then propagated to the next layer and so on until the last layer. Each unit receives some information from other units (or from the external world through some devices)

and processes this information, which will be converted into the output of the unit.

The goal of the network is to learn or to discover some association between input and output patterns, or to analyze, or to find the structure of the input patterns. The learning process is achieved through the modification of the connection weights between units. In statistical terms, this is equivalent to interpreting the value of the connections between units as parameters (e.g., like the values of a and b in the regression equation $y = a + bx$) to be estimated.

The learning process specifies the “algorithm” used to estimate the parameters.

In the next paragraphs, we will try to describe the basic artificial network model, and the way it works.

3.3 THE BASIC ARTIFICIAL MODEL

An artificial network consists of a pool of simple processing units (neurons, cells), which communicate by sending signals to each other over a large number of weighted connections. Each unit performs a relatively simple job: receive input from neighbours or external sources and use this to compute an output signal, which is propagated to other units. Apart from this processing, a second task is the adjustment of the weights. The system is inherently parallel in the sense that many units can carry out their computations at the same time.

Within neural systems it is useful to distinguish three types of units:

- 1) **input units** (indicated by an index i) which receive data from outside the neural network
- 2) **output units** (indicated by an index o) which send data out of the neural network and
- 3) **hidden units** (indicated by an index h) whose input and output signals remain within the neural network



FUN FACT: The NASA Intelligent Flight Control System uses neural networks to autonomously fly an F-15.

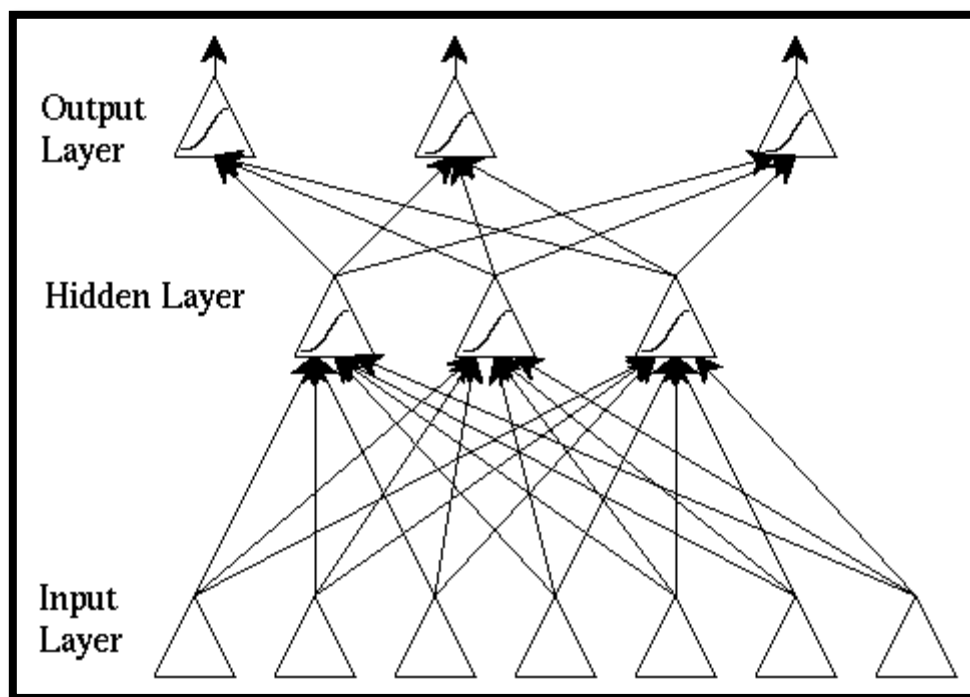


Figure 3: Typical neuron

During operation, units can be updated either synchronously or asynchronously. With synchronous updating, all units update their activation simultaneously while with asynchronous updating, each unit has a (usually fixed) probability of updating its activation at a time t and usually only one unit will be able to do this at a time.

Each input layer, (i), receives information from the network environment which comes via a connection that has a weight (or strength). These weights correspond to synaptic efficacy in a biological neuron. Each neuron also has a

single threshold value. The weighted sum of the inputs is formed, and the threshold subtracted, to compose the *activation* of the neuron (also known as the post-synaptic potential, or PSP, of the neuron). Once the neuron is activated, the activation signal is passed through an activation function, to produce the output of the neuron. If the step activation function is used (i.e., the neuron's output is 0 if the input is less than zero, and 1 if the input is greater than or equal to 0) then the neuron acts just like the biological neuron (subtracting the threshold from the weighted sum and comparing with zero is equivalent to comparing the weighted sum to the threshold). Note also that weights can be negative, which implies that the synapse has an inhibitory rather than excitatory effect on the neuron: inhibitory neurons are found in the brain. To completely understand the function of an Artificial Neural Network, we have to break it up, beginning with its simplest form, the neuron. The individual processing unit in ANNs (neuron) receives input from other sources or output signals of other units and produces an output as shown in Figure 4. The input signals (x_i) are multiplied with weights (w_{ji}) of connection strength between the sending unit " i " and receiving unit " j ". The sum of the weighted inputs is passed through an activation function. The output may be used as an input to the neighboring units or units at the next layer. Assuming the input signal by a vector \mathbf{x} (x_1, x_2, \dots, x_n) and the corresponding weights

to unit " j " by \mathbf{w}_j ($w_{j1}, w_{j2}, \dots, w_{jn}$), the net input to the unit " j " is given by Equation 1. The weight $w_{j0}(=b)$ is a special weight called bias whose input signal is always +1.

$$net_j = \sum w_{ja} X_a + w_{j0} = w_j X + b \text{ (eq.3.1)}$$

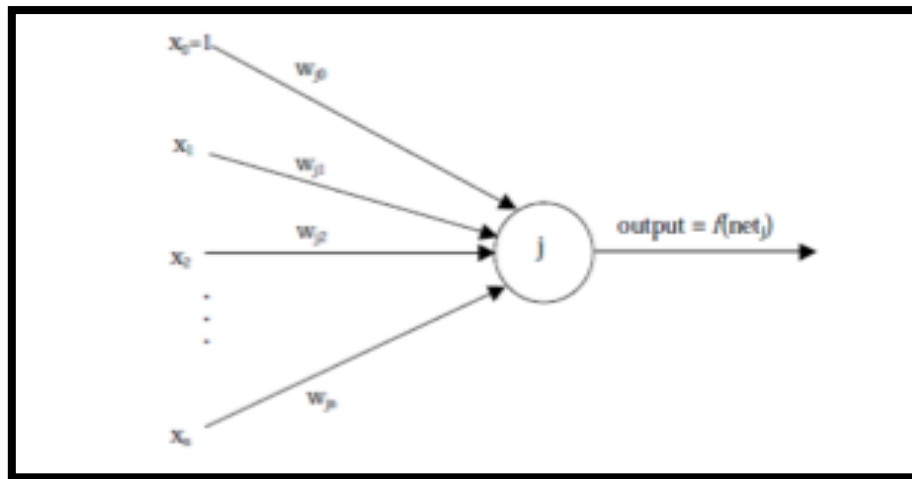


Figure 4: Individual unit in a neural network

The computed weighted sum of inputs is transformed into an output value by applying an activation function. In most cases, the activation function maps the net input between -1 to +1 or 0 to 1. This type of activation function is particularly useful in classification tasks. In cases where a neural network is required to produce any real value, linear activation function may be used at the final layer. A network with multiple layers using linear activation function at intermediate layers effectively reduces to a single-layer network. This type of network is incapable of solving nonlinearly separable problems and has limited capability. Since the most real-world problems are nonlinearly separable, nonlinearity in the intermediate layer is essential for modeling complex problems.

There are many different activation functions proposed in the literature that are often chosen to be monotonically increasing functions. Some of them are briefly presented in the table 1, below:

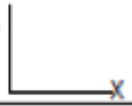
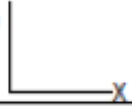
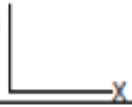

Activation Functions	Mathematical Expression	Graphical Expression
Linear	$f(x) = x$	
Logistic sigmoid	$f(x) = \frac{1}{1 + \exp(-x)}$	
Hyperbolic tangent	$f(x) = \tanh(x)$	
Gaussian	$f(x) = \exp(-x^2/2\sigma^2)$	

Table 1: Commonly used activation functions

Having defined an individual neuron, the next step is to connect them together. A neural network architecture represents a configuration indicating how the units are grouped together as well as the interconnection between them. There are many different architectures reported in the literature, however, most of these can be divided into two main broad categories: *feed-forward* and *feedback*. These architectures are shown in Figure 5. In feed-forward architecture, the information signal always propagates towards the forward direction while in feedback architecture the final outputs are again fed back at the input layer.

A multiple feed-forward layer can have one or more layers of hidden units. The number of units at the input layer and output layer is determined by the problem at hand. Input layer units correspond to the number of independent variables while output layer units correspond to the dependent variables or the predicted values. While the numbers of input and output units are determined by the task at hand, the numbers of hidden layers and the units in each layer may vary. There are no widely accepted rules for designing the configuration of a neural network. A network with fewer than the required

number of hidden units will be unable to learn the input-output mapping, whereas too many hidden units will generalize poorly of any unseen data.

Several researchers attempted to determine the appropriate size of hidden units. Kung and Hwang (1988) suggested that the number of hidden units should be equal to the number of distinct training patterns while Arai (1989) concluded that N input patterns required $N-1$ hidden units in a single layer. However, as remarked by Lee (1997), it is rather difficult to determine the optimum network size in advance. Other studies suggested that

ANNs generalize better when succeeding layers are smaller than the preceding ones (Kruschke, 1989; Looney, 1996). Although a two-layer network is commonly used in most problem solving approaches, the determination of an appropriate network configuration usually requires many trial and error methods. Another way to select network size is to use constructive approaches. In constructive approaches, the network starts with a minimal size and grows gradually during the training procedure (Fahlman & Lebiere, 1990; Lehtokangas, 2000).

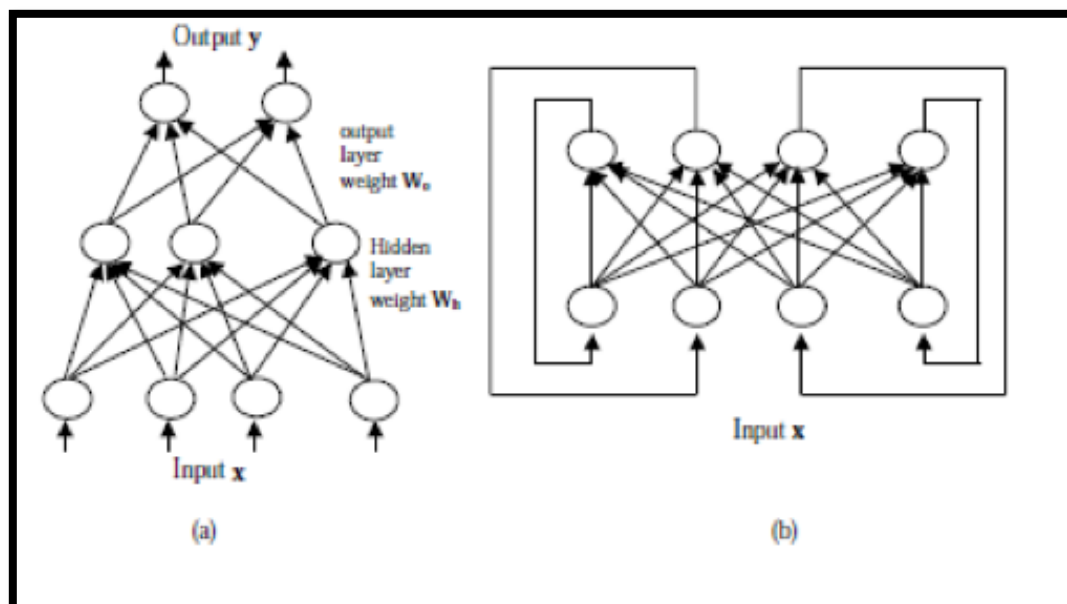


Figure 5: (a) Feed forward architecture (b) Feedback architecture



Figure 6: Aston Martin DB9 engines get neural network

Aston Martin has taken it upon themselves to add the first adaptive neural network to their DB9's V-12 engines to prevent misfiring, which is apparently rather common in those kinds of cars. The system listens to the noise the engine makes, and if everything sounds alright, you'll never even notice the system kicking in—unless things are going seriously wrong and some of the cylinders go into shutdown mode.

3.4 GENERAL FRAMEWORK

A large number of models exist, all different in details but share basic artificial neuron network concept. Here a general framework is to be introduced that possess most of the common features. (Rumelhart & McClelland, 1986):

Elements of the Model

Each model consists of eight principal aspects. They are introduced here with their possible analogy to the human brain and nervous system.

1. A set of *processing units*, or equivalent to the neurons in our nervous system.
2. A *state of activation* that in binary units may be “on” or “off” and in analogue units depends on input to the functional relationship.

3. An output function for each unit. In the neuron this may be translated as speed and/or frequency of impulse conduction.
4. A pattern of connectivity among units. Excitatory or inhibitory behavior of neurotransmitters that receive impulses from other postsynaptic neurons
5. A *propagation* rule for transferring patterns of activities through the network. It is believed that each neuron in the brain is connected to almost 10^4 other neurons.
6. An *activation rule* for combining the inputs to a single unit and based on that decide on the new state of the unit. *Spatial summation* and *threshold of stimulation* together make the *activation function* for the neuron, which decides the final state of the neuron.
7. Learning rule whereby patterns of connectivity are modified. Learning and memory in the brain, *habituation* and *sensitization*.
8. An environment within which the system must operate. Sensory (Afferent) and motor (Efferent) neurons that deal with the outside world in order to receive and send impulses.

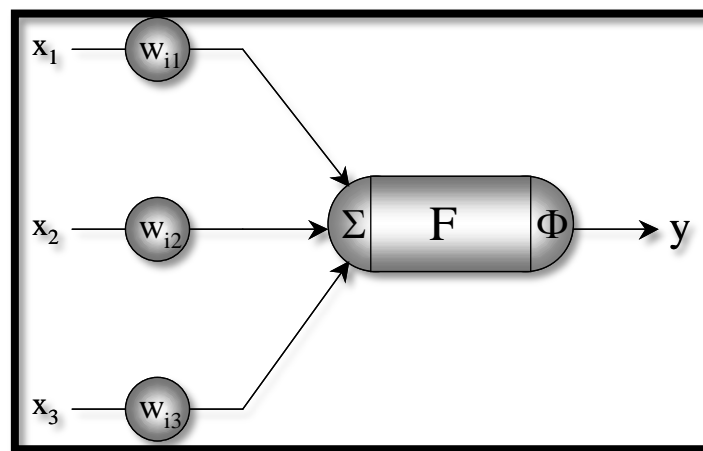


Figure 7: A processing unit

In figure 7 a processing unit is illustrated. The vector $[x_1(t), x_2(t), x_3(t), \dots]$ is called the *input vector*, which either carries information from outside or from the previous neurone. The vector $[w_{i1}, w_{i2}, w_{i3}, \dots]$ is the *weighting* or *strength* vector. Each input vector parameter will be inhibited or exhibited by corresponding member of the weighting vector. This vector (or matrix in networks with many neurones) is said to contain long-term memory of the network (LTM). Σ is the *input function*, which adds up all the incoming values.

F is called the *activation function*. Here, the data is processed and there are a number of possibilities for this function such as:

- The identity function:

$$\text{output} = \text{input}$$

- The threshold function:

$$\begin{cases} \text{if input} \geq \text{Threshold, output} = 1; \\ \text{otherwise output} = 0 \end{cases}$$

- The sigmoid function:

$$\text{output} = \frac{1}{1 + \exp^{-\text{input}}}$$

- The tanh function:

$$\text{output} = \tanh(\text{input})$$

Φ is the output function, which in most of the cases is considered as identity function.

$$\Phi(x) = x \quad (3.2)$$

In some cases it might be threshold function and in some others it is assumed to be a stochastic function in which the output of the unit depends in a probabilistic fashion on its activation values.

A unit's job is simply to receive input from its neighbors and as a function of inputs it receives, to compute an output value, which it sends to its neighbors:

$$y = F\left(\sum_{j=1}^n x_j w_{ij}\right) \quad (3.3)$$

It is also needed to represent the state of each unit. These may be discrete or continuous values. Depending on the unit's activation function, the state might be 0 or 1 (*on* or *off*) or restricted values of $\{-1, 0, +1\}$. In some models they may take any real value between some maximum and minimum. Such values, when represented in a vector format sometimes referred to as *short-term* memory of the network.

Processing units are connected to each other. We assume that each unit provides a contribution to the next one, which it is connected to. The weighted sum of the input signal is received by each element. The weights may be excitatory or inhibitory. The pattern of connectivity is represented by a weighting vector (or matrix) denoted by \mathbf{W} .

Each individual member of \mathbf{W} specifies the strength of the connection. It plays an important role since it represents the knowledge that is encoded in the network and because of this, it is said that matrix \mathbf{W} contains the *long-term* memory of the system.

How and which inputs are transmitted to the recipient unit is decided by the rule of propagation. This rule, in simple networks may be excitatory and inhibitory weighted of the input signal. For more complex patterns, more complex rules of propagation is required.

Activation rule is dependent on the activation function, F , and determines the state of activation. Sometimes, the new state of the activation depends on the old ones as well as the current input signal, the case, which will be discussed in dynamic networks. The activation function is assumed to be deterministic and would be sometimes useful to have another constraint of being a differentiable function.

Learning involves modifying the patterns of interconnectivity, which in principle can involve the development of new connections, loss of existing connections or modification of the strength of connections that already exist. This will be discussed in later sections in details.

The external environment interacts with the network to provide inputs to the network and to receive its outputs. Input units (or layer) receive the external signals and output units (or layer) send out the processed signal to the environment.

3.5 TRAINING OF ARTIFICIAL NEURAL NETWORKS

A neural network has to be configured in such a way, that the application of a set of inputs produces (either direct or via a relaxation process) the desired

set of outputs. Various methods to set the strengths of the connections exist. One way is to set the weights explicitly, using a priori knowledge. Another way is to train the neural network by feeding it teaching patterns and letting it change its weights according to some learning rule. The error of a particular configuration of the network can be determined by running all the training cases through the network, comparing the actual output generated with the desired or target outputs. The differences are combined together by an *error function* to give the network error. The most common error functions are the *sum squared error* (used for regression problems), where the individual errors of output units on each case are squared and summed together, and the cross entropy functions (used for maximum likelihood classification).

In traditional modeling approaches (e.g., linear modeling) it is possible to algorithmically determine the model configuration that absolutely minimizes this error. The price paid for the greater (non-linear) modeling power of neural networks is that although we can adjust a network to lower its error, we can never be sure that the error could not be lower still.

A helpful concept here is the error surface. Each of the N weights and thresholds of the network (i.e., the free parameters of the model) is taken to be a dimension in space. The $N+1$ th dimension is the network error. For any possible configuration of weights the error can be plotted in the $N+1$ th dimension, forming an *error surface*. The objective of network training is to find the lowest point in this many-dimensional surface.

In a linear model with sum squared error function, this error surface is a parabola (a quadratic), which means that it is a smooth bowl-shape with a single minimum. It is therefore "easy" to locate the minimum.

Neural network error surfaces are much more complex, and are characterized by a number of unhelpful features, such as local minima (which are lower than the surrounding terrain, but above the global minimum), flat-spots and plateaus, saddle-points, and long narrow ravines.

It is not possible to analytically determine where the global minimum of the error surface is, and so neural network training is essentially an exploration of the error surface. From an initially random configuration of weights and thresholds (i.e., a random point on the error surface), the training algorithms incrementally seek for the global minimum. Typically, the gradient (slope) of the error surface is calculated at the current point, and used to make a downhill move. Eventually, the algorithm stops in a low point, which may be a local minimum (but hopefully is the global minimum).

3.6 PARADIGMS OF LEARNING

We can categorize the learning situations in three distinct sorts. These are:

- A. **Supervised learning** or Associative learning in which the network is trained by providing it with input and matching output patterns. These input-output pairs can be provided by an external teacher, or by the system which contains the network (self-supervised)
- B. **Unsupervised learning** or Self-organization in which an (output) unit is trained to respond to clusters of pattern within the input. In this paradigm the system is supposed to discover statistically salient features of the input population. Unlike the supervised learning paradigm, there is no a priori set of categories into which the patterns are to be classified; rather the system must develop its own representation of the input stimuli.
- C. **Reinforcement Learning** This type of learning may be considered as an intermediate form of the above two types of learning. Here the learning machine does some action on the environment and gets a feedback response from the environment. The learning system grades its action good (rewarding) or bad (punishable) based on the environmental response and accordingly adjusts its parameters. Generally, parameter adjustment is continued until an equilibrium state occurs, following which there will be no more changes in its parameters. The self-organizing neural learning may be categorized under this type of learning

The best-known example of a neural network training algorithm is *back propagation* (see Patterson, 1996; Haykin, 1994; Fausett, 1994). Modern second-order algorithms such as *conjugate gradient descent* and *Levenberg-Marquardt* (see Bishop, 1995; Shepherd, 1997) (both included in *ST Neural Networks*) are substantially faster (e.g., an order of magnitude faster) for many problems, but *back propagation* still has advantages in some circumstances, and is the easiest algorithm to understand. We will introduce this now, and discuss the more advanced algorithms later. There are also heuristic modifications of *back propagation* which work well for some problem domains, such as *quick propagation* (Fahlman, 1988) and *Delta-Bar-Delta* (Jacobs, 1988) and are also included in *ST Neural Networks*.

In back propagation, the gradient vector of the error surface is calculated. This vector points along the line of steepest descent from the current point, so we know that if we move along it a "short" distance, we will decrease the error. A sequence of such moves (slowing as we near the bottom) will eventually find a minimum of some sort. The difficult part is to decide how large the steps should be.

Large steps may converge more quickly, but may also overstep the solution or (if the error surface is very eccentric) go off in the wrong direction. A classic example of this in neural network training is where the algorithm progresses very slowly along a steep, narrow, valley, bouncing from one side across to the other. In contrast, very small steps may go in the correct direction, but they also require a large number of iterations. In practice, the step size is proportional to the slope (so that the algorithms settle down in a minimum) and to a special constant: the *learning rate*. The correct setting for the learning rate is application-dependent, and is typically chosen by experiment; it may also be time-varying, getting smaller as the algorithm progresses.

The algorithm is also usually modified by inclusion of a momentum term: this encourages movement in a fixed direction, so that if several steps are taken in the same direction, the algorithm "picks up speed", which gives it the ability to (sometimes) escape local minimum, and also to move rapidly over flat spots and plateaus.

The algorithm therefore progresses iteratively, through a number of *epochs*. On each epoch, the training cases are each submitted in turn to the network, and target and actual outputs compared and the error calculated. This error, together with the error surface gradient, is used to adjust the weights, and then the process repeats. The initial network configuration is random, and training stops when a given number of epochs elapses, or when the error reaches an acceptable level, or when the error stops improving (you can select which of these stopping conditions to use).

Depending on the nature of the application and the strength of the internal data patterns you can generally expect a network to train quite well. This applies to problems where the relationships may be quite dynamic or non-linear. ANNs provide an analytical alternative to conventional techniques, which are often limited by strict assumptions of normality, linearity, variable independence etc. Because an ANN can capture many kinds of relationships it allows the user to quickly and relatively easily model phenomena which otherwise may have been very difficult or impossible to explain otherwise.

Artificial neural networks, although very promising in many fields and have advantageous in relation to traditional computing techniques, suffer from certain limitations (Richard & Lippmann, 1991). These problems could be “external” to the network, for example limitation in hardware technology or could be “internal” such as deficiencies in training algorithms etc. Some of the most significant problems could be:

- 1- Long training time. Quick convergence is not guaranteed in any training algorithm. Generally there will be a quick drop in total training error at the beginning of the training, but as the training goes on the error reduction rate will get smaller.
- 2- Network structure for new problems must be selected by trial and error. Although many attempts have been done, there is no clear prescription on how to choose the network topology for a successful generalisation or pattern classification.

- 3- Any *a priori* knowledge about the data cannot be transferred to the network parameters. Although that knowledge may help in the selection of network structure.
- 4- In many engineering applications, network does not provide any physical sense of its finding in relationships. All the data, no matter what they are, are treated the same, however their influence on the network output will inherently be taken into account.
- 5- Functional relationship provided by a trained network by no means is similar to the known and well-established mathematical models, such as regression techniques.
- 6- During training, the weights can be adjusted to very large values, causing the total input of a hidden or output unit to reach very high positive or negative values. Sigmoidal activation function will either saturate to one or zero respectively, meaning that the neurone does not “*respond or sense*” any more changes, a situation called “network paralysis”.

The ability of neural networks to discover nonlinear relationships in input data makes them ideal for modeling nonlinear dynamic systems such as the stock market. Various neural network configurations have been developed to model the stock market. Commonly, these systems are created in order to determine the validity of the EMH or to compare them with statistical methods such as regression. Often these networks use raw data and derived data from technical and fundamental analysis discussed previously.

Market forecasting involves projecting such things stock market indexes, like the Standard and Poor's (S&P) 500 stock index, Treasury bill rates, and net asset value of mutual funds. The role of SC in this case is to use quantitative inputs, like technical indices, and qualitative factors, like political effects, to automate stock market forecasting and trend analysis. This section provides an overview of representative SC studies in this area.

Apparently, White (1988) was the first to use NNs for market forecasting. He was curious as to whether NNs could be used to extract nonlinear regularities from economic time series, and thereby decode previously undetected regularities in asset price movements, such as fluctuations of common stock

prices. The purpose of his paper was to illustrate how the search for such regularities using a feed-forward NN (FFNN) might proceed, using the case of IBM daily common stock returns as an example. White found that his training results were over-optimistic, being the result of over-fitting or of learning evanescent features. He concluded, "*The present neural network is not a money machine.*" Chiang al. (1996) used a FFNN with back propagation (BP) to forecast the end-of-year net asset value (NAV) of mutual funds, where the latter was predicted using historical economic information. They compared those results with results obtained using traditional econometric techniques and concluded that NNs "significantly outperform regression models" when limited data is available. Kuo (1996), recognized that qualitative factors, like political effects, always play a very important role in the stock market environment, and proposed an intelligent stock market forecasting system that incorporates both quantitative and qualitative factors. This was accomplished by integrating a NN and a fuzzy Delphi model (Bojadziev and Bojadziev, 1997 p. 71); the former was used for quantitative analysis and decision integration, while the later formed the basis of the qualitative model. They applied their system to the Taiwan stock market.

Kim and Chun (1998) used a refined probabilistic NN (PNN), called an arrayed probabilistic network (APN), to predict a stock market index. The essential feature of the APN was that it produces a graded forecast of multiple discrete values rather than a single bipolar output. As a part of their study, they use a "mistake chart," which benchmarks against a constant prediction, to compare FFNN with BP models with a PNN, APN, recurrent NN (RNN), and case based reasoning. They concluded that the APN tended to outperform recurrent and BP networks, but that case base reasoning tended to outperform all the networks.

Aiken and Bsat (1999) use a FFNN trained by a genetic algorithm (GA) to forecast three-month U.S. Treasury Bill rates. They conclude that an NN can be used to accurately predict these rates.

Edelman et. al. (1999) investigated the use of an identically structured and independently trained committee of NNs to identify arbitrage opportunities in the Australian All-Ordinaries Index. Trading decisions were made based on the unanimous consensus of the committee predictions and the Sharpe Index was used to assess out-of-sample trading performance. Empirical results showed that technical trading based on NN predictions outperformed the buy-and-hold strategy as well as "naive prediction". They concluded that the reliability of the network predictions and hence trading performance was dramatically enhanced by the use of trading thresholds and the committee approach.

Thammano (1999) used a neuro-fuzzy model to predict future values of Thailand's largest government-owned bank. The inputs of the model were the closing prices for the current and prior three months, and the profitability ratios ROA, ROE and P/E. The output of the model was the stock prices for the following three months. He concluded that the neuro-fuzzy architecture was able to recognize the general characteristics of the stock market faster and more accurately than the basic backpropagation algorithm. Also, it could predict investment opportunities during the economic crisis when statistical approaches did not yield satisfactory results.

Trafalis (1999) used FFNNs with BP and the weekly changes in 14 indicators to forecast the change in the S&P 500 stock index during the subsequent week. In addition, a methodology for pre-processing of the data was devised, which involved differencing and normalizing the data, was successfully implemented. The text walked the reader through the NN process.

Tansel (1999) compared the ability of linear optimization, NNs, and GAs to model time series data using the criteria of modeling accuracy, convenience and computational time.

They found that linear optimization methods gave the best estimates, although the GAs could provide the same values if the boundaries of the parameters and the resolution were selected appropriately, but that the NNs resulted in the worst estimations. However, they noted that non-linearity could

be accommodated by both the GAs and the NNs and that the latter required minimal theoretical background.

Garliauskas (1999) investigated stock market time series forecasting using a NN computational algorithm linked with the kernel function approach and the recursive prediction error method. The main idea of NN learning by the kernel function is that the function stimulates to changes of the weights in order to achieve convergence of the target and forecast output functions. He concluded that financial times series forecasts by the NNs were superior to classical statistical and other methods.

Chan. (2000) investigated financial time series forecasting using a FFNN and daily trade data from the Shanghai Stock Exchange. To improve speed and convergence they used a conjugate gradient learning algorithm and used multiple linear regressions (MLR) for the weight initialization. They conclude that the NN can model the time series satisfactorily and that their learning and initialization approaches lead to improved learning and lower computation costs.

Kim and Han (2000) used a NN modified by a GA to predict the stock price index. In this instance, the GA was used to reduce the complexity of the feature space, by optimizing the thresholds for feature discretization, and to optimize the connection weights between layers. Their goal was to use globally searched feature discretization to reduce the dimensionality of the feature space, eliminates irrelevant factors, and to mitigate the limitations of gradient descent. They concluded that the GA approach outperformed the conventional models.

Romahi and Shen (2000) developed an evolving rule based expert system for financial forecasting. Their approach was to merge FL and rule induction so as to develop a system with generalization capability and high comprehensibility. In this way the changing market dynamics are continuously taken into account as time progresses and the rule base does not become outdated. They concluded that the methodology showed promise.

Abraham (2001) investigated hybridized SC techniques for automated stock market forecasting and trend analysis. They used principal component analysis to preprocess the input data, a NN for one-day-ahead stock forecasting and a neuro-fuzzy system for analyzing the trend of the predicted stock values. To demonstrate the proposed technique, they analyzed 24 months of stock data for the Nasdaq-100 main index as well as six of the companies listed therein. They concluded that the forecasting and trend prediction results using the proposed hybrid system were promising and warranted further research and analysis.

Cao and Tay (2001) used Support Vector Machines (SVMs) to study the S&P 500 daily price index. The generalization error with respect to the free parameters of SVMs were investigated and found to have little impact on the solution. They conclude that it is advantageous to apply SVMs to forecast the financial time series.

Hwarng (2001) investigated NN forecasting of time series with ARMA (p,q) structures. Using simulation and the performance of the Box-Jenkins model as a benchmark, it was concluded that FFNN with BP generally performed well and consistently for time series corresponding to ARMA (p,q) structures. Using the randomized complete block design of experiment, he concluded that overall, for most of the structures, FFNN with BP performed significantly better when a particular noise level was considered during network training

As a follow-up to Kuo (1996), Kuo (2001) developed a GA-based FNN (GFNN) to formulate the knowledge base of fuzzy inference rules, which can measure the qualitative effect (such as the political effect) in the stock market. The effect was further integrated with the technical indexes through the NN. Using the clarity of buying-selling points and buying-selling performance based on the Taiwan stock market to assess the proposed intelligent system, they conclude that a NN based on both quantitative (technical indexes) and qualitative factors is superior to one based only on quantitative factors.

Artificial Neural Networks or ANN has a multitude of real world applications in the business domain which have been classified as follows:

Accounting

- Identifying tax fraud
- Enhancing auditing by finding irregularities

Finance

- Signature and bank note verification
- Mortgage underwriting
- Foreign exchange rate forecasting
- Country risk rating
- Predicting stock initial public offerings
- Bankruptcy prediction
- Customer credit scoring
- Credit card approval and fraud detection
- Stock and commodity selection and trading
- Forecasting economic turning points
- Bond rating and trading
- Loan approvals
- Economic and financial forecasting
- Risk management

3.7 APPLICATION OF ANN IN REAL CASE SCENARIOS

3.7 .1 APPLICATION OF ANN IN BANKRUPTCY

Bankruptcy prediction has long been an important and widely studied topic. The main impact of such research is in bank lending. Banks need to predict the possibility of default of a potential counter-party before they extend a loan. This can lead to sounder lending decisions, and therefore result in significant savings.

The forecast of bankruptcies belong to classification problems. With input variables, generally financial and accounting data on a firm, we try to find out in which category the firm enters, bankrupt or not bankrupt. The availability of a large amount of accounting and financial data on computerized databases facilitates the use of artificial neural networks with quantitative data. They are tested as substitutes of traditional statistical tools such as multivariate discriminate analysis.

There are two main approaches to loan default/bankruptcy prediction. The first approach, the **structural approach**, is based on modeling the underlying dynamics of interest rates and firm characteristics and deriving the default probability based on these dynamics. The second approach is the **empirical** or the **statistical approach**. Instead of modeling the relationship of default with the characteristics of a firm, this relationship is learned from the data. In early empirical approaches, Altman used the classical multivariate discriminant analysis technique with following financial ratios as input variables:

- 1) Working capital/total assets
- 2) Retained earnings/total assets
- 3) Earnings before interest and taxes/total assets
- 4) Market capitalization/ total debt
- 5) Sales/total assets

These particular financial ratios have been widely used as inputs, even for NNs and other nonlinear models. Ohlson introduced the logistic regression approach (LR) to the bankruptcy prediction problem.

It is essentially a linear model with a sigmoid function at the output (it is thus similar to a single-neuron network). Because the output is in between 0 and 1, the model has a nice probabilistic interpretation.

Ohlson used a novel set of financial ratios as inputs. Both the MDA model and the LR model have been widely used in practice and in many academic studies. They have been standard benchmarks for the loan default prediction problem.

Research studies on using NN's for bankruptcy prediction started in 1990, and are still active now.

Currently, several of the major commercial loan default prediction products are based on NN's. For example, Moody's *Public Firm Risk Model* is based on NN's as the main technology. Many banks have also developed and are using proprietary NN default prediction models.

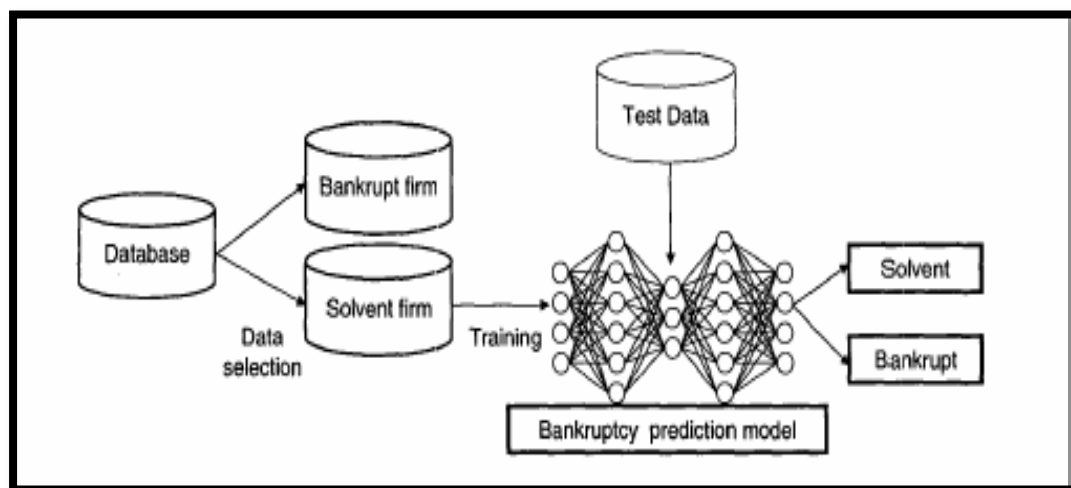


Figure 8: Bankruptcy prediction system framework

Odom and Sharda found that a back-propagation artificial neural network was superior to a Discriminant Analysis model in bankruptcy prediction of firms. In their survey of Savings and Loan

Associations, Tam and Kiang argue that empirical results have shown that ANNs have better predictive accuracy than Discriminant Analysis, Logit, k Nearest Neighbor (kNN) and Decision Tree (ID3) analysis.

From the many studies existing in the literature, it can be seen that NN's are generally more superior to other techniques. Also, the greatest part of the experimentations makes comparisons with traditional statistical forecast models such as ADM, Logistic Regression and Recursive Partitioning, but rare are the studies making a comparison between different neural networks. The ratios used for the implementations are never chosen specifically for a neural network application. The variables are extracted from traditional studies on bankruptcy forecast, or from the existing literature. The recognition of Artificial Neural Networks capacity in bankruptcy forecasting case appears through new experiments. With increasing research the ANN is not any more the object of comparison with traditional forecasting techniques, now it represents a tool of reference.

3.7.2 APPLICATION OF ANN IN CREDIT CARD FRAUD DETECTION

Fraud is increasing dramatically with the expansion of modern technology and the global superhighways of communication, resulting in the loss of billions of dollars worldwide each year. Although prevention technologies are the best way of reducing fraud, fraudsters are adaptive and, given time, will usually find ways to circumvent such measures. Methodologies for the detection of fraud are essential if we are to catch fraudsters once fraud prevention has failed. Statistics and machine learning provide effective technologies for fraud detection and have been applied successfully to detect activities such as credit card fraud.

One of the most interesting fields of prediction is the fraud of credit lines, especially credit card payments. For the high data traffic of 400,000 transactions per day, a reduction of 2.5% of fraud triggers a saving of one

million dollars per year. The extent of credit card fraud is difficult to quantify, partly because companies are often loath to release fraud figures in case they frighten the spending public, and partly because the figures change (probably grow) over time. Various estimates have been given. For example, Leonard suggested the cost of Visa/MasterCard fraud in Canada in 1989, 1990, and 1991 was 19, 29, and 46 million Canadian dollars, respectively. Ghosh and Reilly [20] suggest a figure of 850 million US dollars per year for all types of credit card fraud in the US, and Aleskerov *et al* cite estimates of \$700 million in the US each year for Visa/MasterCard, and \$10 billion worldwide in 1996.

Credit card fraud may be perpetrated in various ways, including simple theft, application fraud, and counterfeit cards. Use of a stolen card is perhaps the most straightforward type of credit card fraud. In this case, the fraudster typically spends as much as possible in as short a space of time as possible, before the theft is detected and the card stopped, so that detecting the theft early can prevent large losses.

Application fraud arises when individuals obtain new credit cards from issuing companies using false personal information. Traditional credit scorecards (Hand and Henley) are used to detect customers who are likely to default, and the reasons for this may include fraud. Such scorecards are based on the details given on the application forms, and perhaps also on other details, such as bureau information.

Statistical models, which monitor behavior over time, can be used to detect cards, which have been obtained from a fraudulent application (e.g. a first time card holder who runs out and rapidly makes many purchases should arouse suspicion).

Cardholder-not-present fraud occurs when the transaction is made remotely, so that only the card's details are needed, and a manual signature and card imprint are not required at the time of purchase. Such transactions include telephone sales and online transactions, and this type of fraud accounts for a high proportion of losses. Researchers who have used neural networks for credit card fraud detection include Ghosh and Reilly [20], Aleskerov [15],

Dorronsoró [18], and Brause [17], mainly in the context of supervised classification. HNC Software has developed **Falcon**, a software package that relies heavily on neural network technology to detect credit card fraud.

Supervised methods, using samples from the fraudulent/non-fraudulent classes as the basis to construct classification rules detecting future cases of fraud, suffer from the problem of unbalanced class sizes mentioned above: the legitimate transactions generally far outnumber the fraudulent ones. Simple misclassification rate cannot be used as a performance measure: with a bad rate of 0.1%, simply classifying every transaction as legitimate will yield an error rate of only 0.001. Instead, one must either minimize an appropriate cost-weighted loss or fix some parameter (such as the number of cases one can afford to investigate in detail) and then try to maximize the number of fraudulent cases detected subject to this. Further it can be seen that there is a dearth of published literature on fraud detection. Of that which has been published, much of it appears in the methodological data analytic literature, where the aim is to illustrate new data analytic tools by applying them to the detection of fraud, rather than being to describe methods of fraud detection per se. Furthermore, since anomaly detection methods are very context dependent, much of the published literature in the area concentrates on supervised classification methods. In particular, rule-based systems and neural networks have attracted interest.

3.7.3 APPLICATION OF ANN IN STOCK MARKET PREDICTION

Financial Market all over the globe is different from other sectors like HR etc. We could model any financial market as a complex feedback mechanism working on both external stimulus as well as past results. Prices are unstable and have a tendency to fall and rise by any magnitude. Typical example includes share markets all over the world. Stock Market involves trade risk; swap risk, and greater amount of uncertainty. Here the role of accurate prediction is highly appreciated for it was possible to predict it there would be no risk.

Neural networks have found ardent supporters among various avant-garde portfolio managers, investment banks and trading firms. Most of the major investment banks, such as Goldman Sachs and Morgan Stanley, have dedicated departments to the implementation of neural networks. Fidelity Investments has set up a mutual fund whose portfolio allocation is based solely on recommendations produced by an artificial neural network. The fact that major companies in the financial industry are investing resources in neural networks indicates that artificial neural networks may serve as an important method of forecasting.

In Stock Prediction, neural networks have been found useful in stock price prediction. Lee has talked about using back propagation algorithm based Artificial Neural Networks for stock prediction. D. Pham discusses various aspects of stock prediction processes and gives a general overview of it. Literature survey of last 5 years gives several solution models proposed for stock prediction problem. Some of the broader classification is Time Series method, recurrent neural network and Feed-forward neural network method. These models are all used to learn the relationships between different technical and economical indices and the decision to buy or sell stocks. The inputs to all the models are technical and economic indices. The output of the system is the decision to buy and sell and is mostly based on fuzzy logic where, by system gives the decision not as a binary signal but as fuzzy signal with a certain percentage of success.

Peifer advocated deploying an AI based neural network for stock prediction. This system should use sufficient amount of historical stock data as input and then train the network with this data. Once trained, the neural network can be used to predict stock behavior. Most of the papers we have studied in this area, advocated use of Back propagation algorithm in ANN for stock prediction.

Analysis of the problem domains of NN applications has shown that there are three main groups of problems that NN applications frequently deal with. First group consists of predicting stock performance by trying to classify stocks into

the classes such as: stocks with either positive or negative returns and stocks that perform well, neutrally, or poorly. Such NN applications give valuable support to making investment decisions, but do not specify the amount of expected price and expected profit. The next group of frequently used applications gives more information: NN's for stock price predictions. Such systems try to predict stock prices for one or more days in advance, based on previous stock prices and on related financial ratios. The third important group of NN applications in stock markets is concerned with modeling stock performance and forecasting.

An important trend in the applications is combining two or more NN's into a single NN system, or incorporating other artificial intelligence methods into a NN system, such as expert systems, genetic algorithms, natural language processing. The number of Kohonen's, Hopfield's, and other algorithms is relatively small in the stock market NN applications. Following table summarizes use of various solutions provided in Stock prediction market.

S	Problem Domain	Solutions given
1	Predicting stock performance	Backpropagation Boltzman machine
2	Stock price predictions	Backpropagation Perceptron, ADALINE /MADALINE
3	Modeling the stock performance (ANN combined	Backpropagation Hybrid approach (Backpropagation NN + Expert system)

Table2: Summary of various solutions provided on stock prediction market

After a brief overview of the articles and papers, it was evident that almost all applications of NN in stock markets are based on a different data model. The authors emphasize the necessity for including more data in the models, such as other types of asset; more financial ratios; and qualitative data.

Furthermore, the recommendation for the use of various time periods occurs frequently. Stocks are commonly predicted on the basis of daily data, although some researchers use weekly and monthly data.

Additionally, future research should focus on the examinations of other types of networks that were rarely applied, such as Hopfield's, Kohonen's, etc. Finally, almost all researchers emphasize the integration of NN's with other methods of artificial intelligence as one of the best solutions for improving the limitations.

Marijana in her paper says that Back propagation algorithm has the ability to predict with greater accuracy than other NN algorithms, no matter which data model was used. She also says that NN outperform classical forecasting and statistical methods, such as multiple regression analysis and discriminate analysis. The combination of the NN calculating ability based on heuristics and the ability of expert systems to process the rules for making a decision and to explain the results can be a very effective intelligent support in various problem domains.

After completing several simulations for predicting several stocks based on the past historical data using fuzzy neural network with the Back-Propagation learning algorithm, it is conclusive that the average error for simulations using lots of data is smaller than that using less amount of data. That is, the more data for training the neural network, the better prediction it gives. Further, it can be concluded that:

1. NN's are efficiency methods in the area of stock market predictions, but there is no "recipe" that matches certain methodologies with certain problems.
2. NN's are most implemented in forecasting stock prices and returns, although stock modeling is a very promising problem domain of its application.

3. Most frequent methodology in ANN's is the Back propagation algorithm, but many authors emphasize the importance of integration of NN with other artificial intelligence methods.
4. Benefits of NN are in their ability to predict accurately even in situations with uncertain data, and the possible combinations with other methods.
5. Limitations have to do with insufficient reliability tests, data design, and the inability to identify the optimal topology for a certain problem domain.

3.7.4. APPLICATION OF ANN IN FINANCE

Our study of ANN in the financial domain is how information technology developments affect the nature of the audit process and the audit skills. In this section we have reviewed several papers and articles and the review showed that the main application areas in auditing were material errors, management fraud, and support for backing concern decision. ANN's have also been find huge applications in control risk assessment, audit fee, and financial distress problems.

Very many things in our business and auditing environment are changing at an increasing rate. Increased competition and the need for faster and better information for decisions mark today's business environment. In addition, systems are complex and many times on-line. This complexity means that auditors have more and different kinds of work to do than they had earlier. In case of Indian financial sector, in early 90's most of the work is done by pen and paper way i.e. is use of electronic means is pretty less. But now things are changed. All major Audit firms like PWC, Vaish Associates, Kothari and Kothari etc. have all gone electronic in process. Following table summarizes different levels of software tools used by a modern auditor today.

STAGE	SOFTWARE APPLICATION	UTILIZATION
I	word-processing, spreadsheets	Documentation, auditor's report, financial analysis and calculations
II	graphics, external databases, electronic mail	Audit planning, comparison of financial information, company analysis
III	company models, audit databases, IS audit software applications	Testing of information systems, database inquiries
IV	expert systems, decision support systems, special software for continuous audit	Expert analysis for finding important tasks for audit
V	Advanced method, ANN-based systems	Assurance services

TABLE 3: Summary of different levels of software tools

The major ANN-application area in auditing is material errors. Material error applications direct auditors' attention to those financial account values where the actual relationships are not consistent with the expected relationships. An auditor has to decide whether and what kind of further audit investigation is required to explain the unexpected results. Auditors cannot assume that the management is honest or dishonest for most of the records here are fraudulent. Management fraud (MF) can be defined as deliberate fraud committed by the management that injures investors and creditors through materially misleading financial statements.

An auditor considers a huge amount of data when assessing the risk of the internal control (IC) structure of an entity failing to prevent or detect significant misstatements in financial statements. Curry and Peel provided an overview of the ANN modeling approach and the performance of ANN's, relative to conventional ordinary least squares (OLS) regression analysis, in predicting the cross-sectional variation in corporate audit fees (AF). As information technological changes occur at an increasing rate, auditors must keep pace with these emerging changes and their impact on their client's information processing systems as well as on their own audit procedures. This study pictures the current state of the ANN applications connected to auditing purpose. The review is comprehensive but by no means exhaustive, given the fast growing nature of the literature.

In brief, the main findings are summarized as follows: The main application areas were material errors, management fraud, and support for going concern decision. ANN's have also been applied to internal control risk assessment, audit fee, and financial distress problems.

Conclusions

Artificial Neural Networks offer qualitative methods for business and economic systems that traditional quantitative tools in statistics and econometrics cannot quantify due to the complexity in translating the systems into precise mathematical functions. Hence, the use of neural networks in finance is a promising field of research especially given the ready availability of large mass of data sets and the reported ability of neural networks to detect and assimilate relationships between a large numbers of variables.

CHAPTER 4

Fuzzy Logic

4.1) Definition and terminology

What Does *Fuzzy Logic* Mean?

Logic, according to Webster's dictionary, is ***the science of the normative formal principles of reasoning***. In this sense, fuzzy logic is concerned with the formal principles of approximate reasoning, with precise reasoning viewed as a limiting case. In more specific terms, what is central about fuzzy logic is that, unlike classical logical systems, it aims at modeling the imprecise modes of reasoning that play an essential role in the remarkable human ability to make rational decisions in an environment of uncertainty and imprecision. This ability depends, in turn, on our ability to infer an approximate answer to a question based on a store of knowledge that is inexact, incomplete, or not totally reliable.

Logic in general, deals with true and false. A *proposition* can be true on one occasion and false on another. "Apple is a red fruit" is such a proposition. If you are holding a Granny Smith apple that is green, the proposition that apple is a red fruit is false. On the other hand, if your apple is of a red delicious variety, it is a red fruit and the proposition in reference is true. If a proposition is true, it has a truth value of 1; if it is false, its truth value is 0. These are the only possible truth values. Propositions can be combined to generate other propositions, by means of logical operations.

When you say it will rain today or that you will have an outdoor picnic today, you are making statements with certainty. Of course your statements in this case can be either true or false. The truth values of your statements can be only 1, or 0. Your statements then can be said to be ***crisp***.

On the other hand, there are statements you cannot make with such certainty. You may be saying that you think it will rain today. If pressed further, you may

be able to say with a degree of certainty in your statement that it will rain today. Your level of certainty, however, is about 0.8, rather than 1. This type of situation is what *fuzzy logic* was developed to model. Fuzzy logic deals with propositions that can be true to a certain degree—somewhere from 0 to 1. Therefore, a proposition's truth value indicates the degree of certainty about which the proposition is true. The degree of certainty sounds like a probability (perhaps subjective probability), but it is not quite the same. Probabilities for mutually exclusive events cannot add up to more than 1, but their fuzzy values may. Suppose that the probability of a cup of coffee being hot is 0.8 and the probability of the cup of coffee being cold is 0.2. These probabilities must add up to 1.0. Fuzzy values do not need to add up to 1.0. The truth value of a proposition that a cup of coffee is hot is 0.8. The truth value of a proposition that the cup of coffee is cold can be 0.5. There is no restriction on what these truth values must add up to.

The following questions and the way of answering them can give us a more precise view on the term, fuzzy logic.

For example:

- (1) Usually it takes about an hour to drive from Athens to Megara and about half an hour to drive from Athens to Piraeus. How long would it take to drive from Athens to Megara via Piraeus?
- (2) Most of those who live in Switzerland have high incomes. It is probable that
Mary lives in Switzerland. What can be said about Mary's income?
- (3) Slimness is attractive. Carol is slim. Is Carol attractive?
- (4) Brian is much taller than most of his close friends. How tall is Brian?

There are two main reasons why classical logical systems cannot cope with problems of this type. First, they do not provide a system for representing the meaning of propositions expressed in a natural language when the meaning is imprecise; and second, in those cases in which the meaning can be represented symbolically in a meaning representation language, for example,

a semantic network or a conceptual-dependency graph, there is no mechanism for inference.

As will be seen, fuzzy logic addresses these problems in the following ways. First, the meaning of a lexically imprecise proposition is represented as an elastic constraint on a variable; and second, the answer to a query is deduced through a propagation of elastic constraints.

During the past several years, fuzzy logic has found numerous applications in fields ranging from finance to earthquake engineering. But what is striking is that it's most important and visible application today is in a realm not anticipated when fuzzy logic was conceived, namely, the realm of fuzzy-logic-based process control.

The basic idea underlying fuzzy logic control was suggested in notes published in **1968** and **1972**, and described in greater detail in **1973**. The first implementation was pioneered by Mamdani and Assilian in **1974** in connection with the regulation of a steam engine. In the ensuing years, once the basic idea underlying fuzzy logic control became well understood, many applications followed. In Japan, in particular, the use of fuzzy logic in control processes is being pursued in many application areas, among them automatic train operation (Hitachi), vehicle control (Sugeno Laboratory at Tokyo Institute of Technology), robot control (Hirota Laboratory at Hosei University), speech recognition (Ricoh), universal controller (Fuji), and stabilization Control (Yamakawa Laboratory at Kumamoto University).

Fuzzy logic is designed to solve problems in the same way that humans do: by considering all available information and making the best possible decision given the input.

With fuzzy logic, propositions can be represented with degrees of truthfulness and falsehood. For example, the statement, *today is sunny*, might be 100% true if there are no clouds, 80% true if there are a few clouds, 50% true if it's hazy and 0% true if it rains all day.

Fuzzy logic is often applied by advanced trading models/systems that are designed to react to changing markets. The goal of this type of system is to analyze thousands of securities in real time and to present the trader with the best available opportunity.

4.2 Basic concepts of fuzzy logic

A) Fuzzy set and membership functions.

We begin with an example. The interest rate 0.02 is considered as a prototype for “low interest rate” and 0.06 is considered as definitely being outside of the “low interest rate” set of rates. A particular rate (e.g., 0.04) can be compared with 0.02 and 0.06 and the result can be expressed as a number reflecting expert’s opinion about degree of membership of 0.04 to the set of “low interest rates”. This number is called a value of the membership function (MF) $m(x)$ of the fuzzy set “low interest rate”. Figure 7.5 gives an example of such membership function. Through this chapter, a financial demonstration example from [Von Altrock, 1997] is used. In particular, the membership function presented in Figure 4.1 is extracted from that demonstration. In this way, we show advantages and disadvantages of usage of fuzzy logic in finance. In particular Section 4.1 shows the inconsistency of standard “context-free” (truth-functional) operations. Then context spaces are used to fix this inconsistency. Let X be a set of all possible interest rates from 0.0 to 1.0. The set of interest rates $\{x\}$ is called the universe X and x is called a “Base variable”.

The support of the fuzzy set is defined as a set of x such that $m(x) > 0$. The degree of membership m covers the entire interval $[0, 1]$, where 0 corresponds to absolutely no membership in the set, 1 corresponds to complete membership and 0.5 usually corresponds to the most uncertain degree of membership.

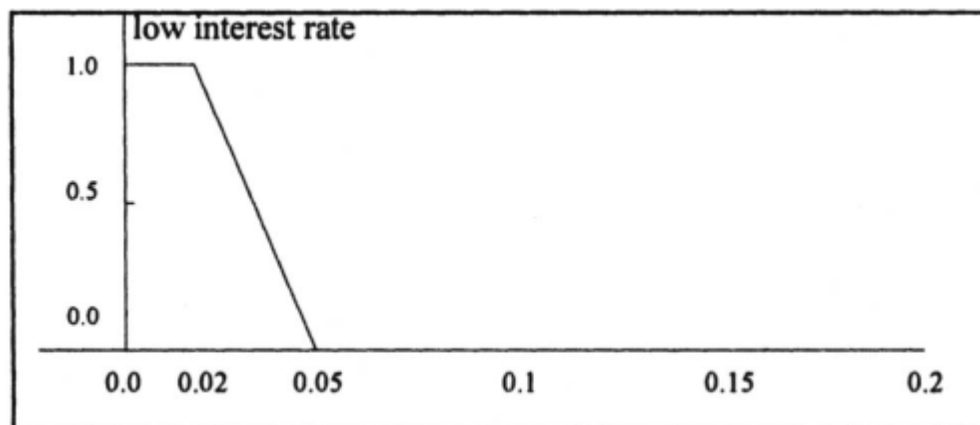


Figure 4.1: Fuzzy set “low interest rate” (“low rate”)

Some examples of membership values are presented below:

$$\begin{array}{lll}
 m_{\text{low-rate}}(1.0) = 0 & m_{\text{low-rate}}(0.05) = 0.022 & m_{\text{low-rate}}(0.025) = 0.9 \\
 m_{\text{low-rate}}(0.5) = 0 & m_{\text{low-rate}}(0.04) = 0.5 & m_{\text{low-rate}}(0.02) = 1 \\
 m_{\text{low-rate}}(0.06) = 0 & m_{\text{low-rate}}(0.03) = 0.72 & m_{\text{low-rate}}(0.01) = 1
 \end{array}$$

Fuzzy sets generalize conventional “crisp” sets with only two values, 1 and 0.

Formally the pair:

$$\langle X, \{m_{\text{low-rate}}(x): x \in X\} \rangle$$

is called a fuzzy set of low rates, i.e., to define a fuzzy set we need three components: linguistic term (“low rate”), universe i.e. ($X=[0,1]$, all possible rates), and a membership function. Fuzzy logic combines fuzzy sets to infer conclusions from fuzzy logic expressions using membership functions. For example, fuzzy logic assigns a truth value to the expression “the interest rate 0.03 is low AND the interest rate 0.24 is low”. It is based on the expression “the interest rate 0.03 is low” with the truth value 0.72 and the expression “the interest rate 0.024 is low” with the truth 0.82.

B) Linguistic Variables.

The most productive concept of fuzzy logic is the concept of linguistic variable [Zadeh, 1977]. Linguistic variable is a set of fuzzy sets defined on the same universe X for related linguistic terms like low, medium, high. See Figure 7.6 for membership functions for these terms [Von Altrock, 1997].

At first glance, any related fuzzy sets can and should be used to create a linguistic variable. However, assigning membership degree $m_{\text{medium}}(0.03)$ for rate 0.03 without coordinating this value with already assigned value $m_{\text{low}}(0.03)$ would be non-contextual. In Figure 4.2 it is not the case, here $m_{\text{medium}}(0.03)=1-m_{\text{low}}(0.03)$.

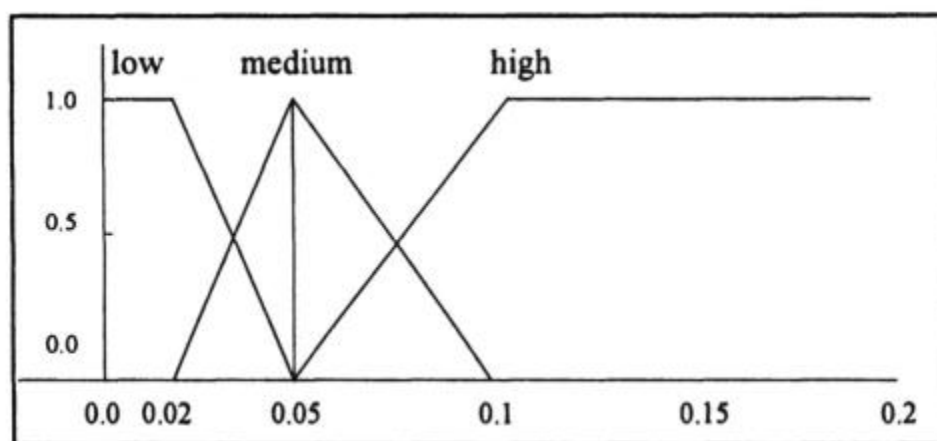


Figure 4.2: Membership functions “low”, “medium”, “high” for linguistic variable “interest rate”

C) Fuzzification.

The process of computing values of membership functions of fuzzy sets for given values of base variables is called fuzzification. Example 1: Let “interest rate” = 0.03. The result of fuzzification for 0.03 could be (Figure 4.2):

low	truth value = 0.72	$(m_{\text{low}}(0.03)=0.72)$
medium	truth value = 0.28	$(m_{\text{medium}}(0.03)=0.28)$
high	truth value = 0.0	$(m_{\text{high}}(0.03)=0.0)$

Example 2: Let "Trade fee" = 0.015. The result of fuzzification for 0.015 could be (Figure 4.3):

low
significant

truth value = 0.78
truth value = 0.22

The term set {low, medium, high} for the linguistic variable "interest rate" does not allow us to express linguistically an interest rate of 0.03. The closest term is "low" (0.72). A larger term set, which will include the term "almost low rate, just slightly greater" can be developed. In this way, 0.03 can be represented by a specific linguistic term. Otherwise fuzzification will represent the same idea numerically as three numbers (0.72, 0.28, and 0), respectively for low, medium and high rates. The result of fuzzification is used as input for the fuzzy rules.

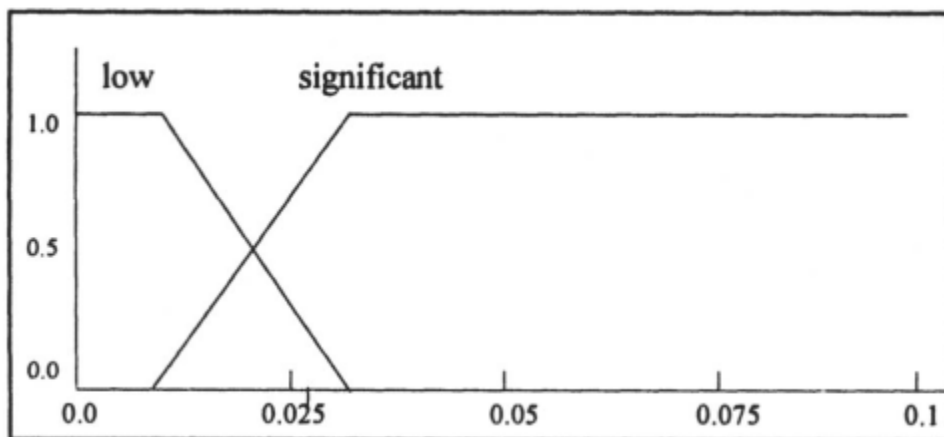


Figure 4.3: Linguistic variable "trade fee"

D) Fuzzy Sets

Fuzzy logic is best understood in the context of *set membership*. Suppose you are assembling a set of rainy days. Would you put today in the set? When you deal only with crisp statements that are either true or false, your inclusion of today in the set of rainy days is based on certainty. When dealing with fuzzy logic, you would include today in the set of rainy days via an *ordered pair*, such as (today, 0.8). The first member in such an ordered pair is a *candidate for inclusion* in the set, and the second member is a value between 0 and 1,

inclusive, called the *degree of membership* in the set. The inclusion of the degree of membership in the set makes it convenient for developers to come up with a set theory based on fuzzy logic, just as regular set theory is developed. Fuzzy sets are sets in which members are presented as ordered pairs that include information on degree of membership. A traditional set of, say, k elements, is a special case of a fuzzy set, where each of those k elements has 1 for the degree of membership, and every other element in the universal set has a degree of membership 0, for which reason you don't bother to list it.

E) Fuzzy Set Operations

The usual operations you can perform on ordinary sets are *union*, in which you take all the elements that are in one set or the other; and *intersection*, in which you take the elements that are in both sets. In the case of fuzzy sets, taking a union is finding the degree of membership that an element should have in the new fuzzy set, which is the union of two fuzzy sets.

If a , b , c , and d are such that their degrees of membership in the fuzzy set A are 0.9, 0.4, 0.5, and 0, respectively, then the fuzzy set A is given by the *fit vector* (0.9, 0.4, 0.5, 0). The components of this fit vector are called *fit values* of a , b , c , and d .

Union of Fuzzy Sets

Consider a union of two traditional sets and an element that belongs to only one of those sets. Earlier you saw that if you treat these sets as fuzzy sets, this element has a degree of membership of 1 in one case and 0 in the other since it belongs to one set and not the other. Yet you are going to put this element in the union. The criterion you use in this action has to do with degrees of membership. You need to look at the two degrees of membership, namely, 0 and 1, and pick the higher value of the two, namely, 1. In other words, what you want for the degree of membership of an element when listed in the union of two fuzzy sets, is the maximum value of its degrees of membership within the two fuzzy sets forming a union.

If a , b , c , and d have the respective degrees of membership in fuzzy sets A , B as $A = (0.9, 0.4, 0.5, 0)$ and $B = (0.7, 0.6, 0.3, 0.8)$, then $AB = (0.9, 0.6, 0.5, 0.8)$.

Intersection and Complement of Two Fuzzy Sets

Analogously, the degree of membership of an element in the intersection of two fuzzy sets is the *minimum*, or the smaller value of its degree of membership individually in the two sets forming the intersection. For example, if today has 0.8 for degree of membership in the set of rainy days and 0.5 for degree of membership in the set of days of work completion, then today belongs to the set of rainy days on which work is completed to a degree of 0.5, the smaller of 0.5 and 0.8.

Recall the fuzzy sets A and B in the previous example. $A = (0.9, 0.4, 0.5, 0)$ and $B = (0.7, 0.6, 0.3, 0.8)$.

$A[\cap]B$, which is the intersection of the fuzzy sets A and B , is obtained by taking, in each component, the smaller of the values found in that component in A and in B . Thus $A[\cap]B = (0.7, 0.4, 0.3, 0)$.

The idea of a universal set is implicit in dealing with traditional sets. For example, if you talk of the set of married persons, the universal set is the set of all persons. Every other set you consider in that context is a subset of the universal set. We bring up this matter of universal set because when you make the complement of a traditional set A , you need to put in every element in the universal set that is not in A . The complement of a fuzzy set, however, is obtained as follows. In the case of fuzzy sets, if the degree of membership is 0.8 for a member, then that member is not in that set to a degree of $1.0 - 0.8 = 0.2$. So you can set the degree of membership in the complement fuzzy set to the complement with respect to 1. If we return to the scenario of having a degree of 0.8 in the set of rainy days, then today has to have 0.2 membership degree in the set of nonrainy or clear days.

Continuing with our example of fuzzy sets A and B , and denoting the complement of A by A' , we have $A' = (0.1, 0.6, 0.5, 1)$ and $B' = (0.3, 0.4, 0.7, 0.2)$. Note that $A' [\cup] B' = (0.3, 0.6, 0.7, 1)$, which is also the complement of

A [cap] B. You can similarly verify that the complement of A [cup] B is the same as A' [cap] B'. Furthermore, A [cup] A' = (0.9, 0.6, 0.5, 1) and A [cap] A' = (0.1, 0.4, 0.5, 0), which is not a vector of zeros only, as would be the case in conventional sets. In fact, A and A' will be equal in the sense that their fit vectors are the same, if each component in the fit vector is equal to 0.5.

Commercial Applications

Many commercial uses of fuzzy logic exist today. A few examples are listed here:

- A subway in Sendai, Japan uses a fuzzy controller to control a subway car. This controller has outperformed human and conventional controllers in giving a smooth ride to passengers in all terrain and external conditions.
- Cameras and camcorders use fuzzy logic to adjust autofocus mechanisms and to cancel the jitter caused by a shaking hand.
- Some automobiles use fuzzy logic for different control applications. Nissan has patents on fuzzy logic braking systems, transmission controls, and fuel injectors. GM uses a fuzzy transmission system in its Saturn vehicles.
- FuziWare has developed and patented a fuzzy spreadsheet called *FuziCalc* that allows users to incorporate fuzziness in their data.
- Software applications to search and match images for certain pixel regions of interest have been developed. Avian Systems has a software package called *FullPixelSearch*.
- A stock market charting and research tool called *SuperCharts* from Omega Research, uses fuzzy logic in one of its modules to determine whether the market is bullish, bearish, or neutral.

FUZZY SETS AND OPERATIONS

Classical sets are also called 'crisp' sets so as to distinguish them from fuzzy sets. In fact, the crisp sets can be taken as special cases of fuzzy sets. Let A be a crisp set defined over the universe X. Then for any element x in X, either x is a member of A or *not*. In fuzzy set theory, this property is generalized.

Therefore, in a fuzzy set, it is not necessary that x is a *full member* of the set or *not a member*. It can be a *partial member* of the sets.

The generalization is performed as follows: For any crisp set A , it is possible to define a characteristic function $\mu_X = \{0,1\}$. i.e. the characteristic function takes either of the values 0 or 1 in the classical set. For a fuzzy set, the characteristic function can take any value between zero and one.

The membership function $\mu_A(x)$ of a fuzzy set A is a function $\mu_A: X \rightarrow [0, 1]$ --- (1)

So every element in x in X has membership degree: $\mu_A(x) \in [0,1]$ ----(2)

A is completely determined by the set of tuples: $A = \{(x, \mu_A(x)) \mid x \in X\}$ ----- (3)

Example1:

Suppose someone wants to describe the class of cars having the property of being expensive by considering BMW, Rolls Royce, Mercedes, Ferrari, Fiat, Honda and Renault. Some cars like Ferrari and Rolls Royce are definitely expensive and some like Fiat and Renault are not expensive in comparison and do not belong to the set. Using a fuzzy set, the fuzzy set of expensive cars can be described as:

$\{(Ferrari, 1), (Rolls\ Royce, 1), (Mercedes, 0.8), (BMW, 0.7), (Honda, 0.4)\}$. Obviously, Ferrari and Rolls Royce have membership value of 1 whereas BMW, which is less expensive, has a membership value of 0.7 and Honda 0.4.

A set of natural numbers 'close to 6' can be defined as a fuzzy set. This can be done, say, by including all numbers from 3 to 9 as follows:

$$\tilde{6} = \{(3,0.1), (4,0.2), (5,0.5), (6,1), (7,0.5), (8,0.2), (9,0.1)\}$$

The expression for the characteristic function or membership function can be written as:

$$\mu_{\tilde{6}}(x) = \frac{1}{(1+(x-6)^2)}$$

The fuzzy set $\tilde{6}$ contains the elements like (6, 1), (5.5, 0.8), and (100, 0.000113161) .etc.

Zadeh proposed an alternate representation for fuzzy sets, which is more convenient. Suppose A is a finite crisp set with elements $\{x_1, x_2, \dots, x_n\}$, then an alternative representation for C is:

$$C = \{x_1 + x_2 + x_3 + \dots + x_n\}$$

Here, +, denotes an enumeration or listing rather than addition. Yet another way is to include the characteristic function also into its fold is:

$$\frac{\mu(x)}{x}$$

where the line between the top and bottom entries is just a delimiter or separator.

The fuzzy set of expensive cars can be now written using this notation as:

$$\left\{ \frac{1}{\text{Ferrari}} + \frac{1}{\text{Rollsroyce}} + \frac{0.8}{\text{Mercedes}} + \frac{0.7}{\text{BMW}} + \frac{0.4}{\text{Honda}} + \frac{0}{\text{Fiat}} + \frac{0}{\text{Renault}} \right\}$$

The set $F = \{(x, \mu_F(x)) \mid x \in X\}$ can then be written as:

$$F = \frac{\mu_F(x_1)}{x_1} + \dots + \frac{\mu_F(x_n)}{x_n} = \sum_{i=1}^n \frac{\mu_F(x_i)}{x_i}$$

Where + satisfies the condition:

$$\frac{a}{x} + \frac{b}{x} = \max\left(\frac{a, b}{x}\right)$$

I.e. if the same element has two membership values, say 0.8 and 0.6, then its membership degree becomes 0.8, the larger of the two. Any countable or discrete universe U allows such a notation.

$$A = \sum_{x \in X} \frac{\mu_F(x)}{x}$$

But when the set is uncountable, or continuous, it can be written as:

$$A = \int \frac{\mu_A(x)}{x}$$

Here the symbol \int denotes a listing or collection rather than integration.

The set 'close to 6' can now be re-written as:

$$\tilde{6} = \left\{ \frac{0.1}{3}, \frac{0.2}{4}, \frac{0.5}{5}, \frac{1}{6}, \frac{0.5}{7}, \frac{0.2}{8}, \frac{0.1}{9} \right\}$$

Example 2:

X can be the set of integers described by the membership function:

$$\mu_F(x) = \begin{cases} 1 - \sqrt{\left| \frac{x-6}{3} \right|} & 3 \leq x \leq 6 \\ 0 & \text{otherwise} \end{cases}$$

Then F can be expressed as:

$$A = \int \frac{\mu_F(x)}{x}$$

Where the operation is on the set of all real integers.

Example 3:

Consider the set of old people belonging to the universe of people in the age of 0 to 120. $X = [0, 120]$

We can define the membership function as:

$$\mu_{old} = \begin{cases} 0 & ; & 0 \leq x \leq 60 \\ (x - 60) / 20 & ; & 60 \leq x \leq 80 \\ 1 & ; & x \geq 80 \end{cases}$$

Then, we can call the set as:

$$\int_0^{120} \frac{\mu_{old}(x)}{x}$$

PROPERTIES OF FUZZY SETS

Fuzzy sets follow the same properties as crisp sets. Since membership values of crisp sets are a subset of the interval [0, 1], classical sets can be thought of as generalization of fuzzy sets:

$$\text{Commutativity : } \tilde{A} \cup \tilde{B} = \tilde{B} \cup \tilde{A}$$

$$\tilde{A} \cap \tilde{B} = \tilde{B} \cap \tilde{A}$$

$$\text{Associativity: } \tilde{A} \cup (\tilde{B} \cup \tilde{C}) = (\tilde{A} \cup \tilde{B}) \cup \tilde{C}$$

$$\tilde{A} \cap (\tilde{B} \cap \tilde{C}) = (\tilde{A} \cap \tilde{B}) \cap \tilde{C}$$

$$\text{Distributivity: } \tilde{A} \cup (\tilde{B} \cap \tilde{C}) = (\tilde{A} \cup \tilde{B}) \cap (\tilde{A} \cup \tilde{C})$$

$$\text{Idempotency: } \tilde{A} \cup \tilde{A} = \tilde{A}$$

$$\tilde{A} \cap \tilde{A} = \tilde{A}$$

$$\text{Identity : } \tilde{A} \cup \emptyset = \tilde{A} \quad \tilde{A} \cap X = \tilde{A}$$

$$\tilde{A} \cap \emptyset = \emptyset \quad \tilde{A} \cup X = X$$

$$\text{Transitivity: } \tilde{A} \subseteq \tilde{B} \subseteq \tilde{C} \text{ then } \tilde{A} \subseteq \tilde{C}$$

$$\text{Involution: } \overline{\overline{\tilde{A}}} = \tilde{A}$$

4.3. OPERATIONS ON FUZZY SETS

Let $\tilde{A}, \tilde{B}, \tilde{C}$ be three fuzzy sets defined on the universe of discourse X. For a given element x of the universe, the following function-theoretic operations of Union, Intersection and Complements are defined as follows:

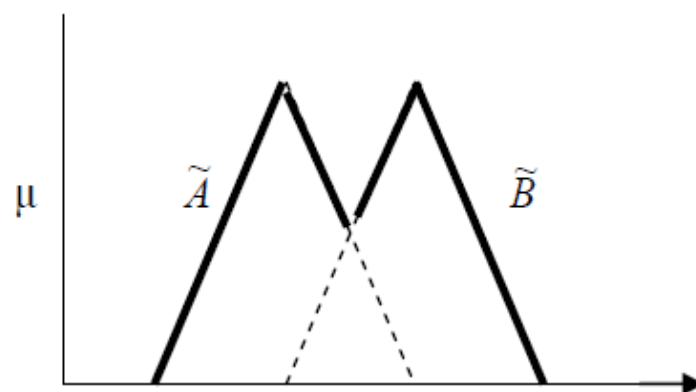


Figure 4.4: Union of fuzzy sets.

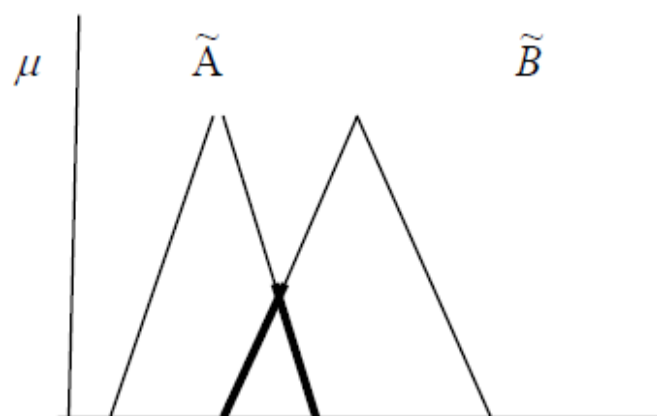


Figure 4.5 : Intersection of fuzzy sets

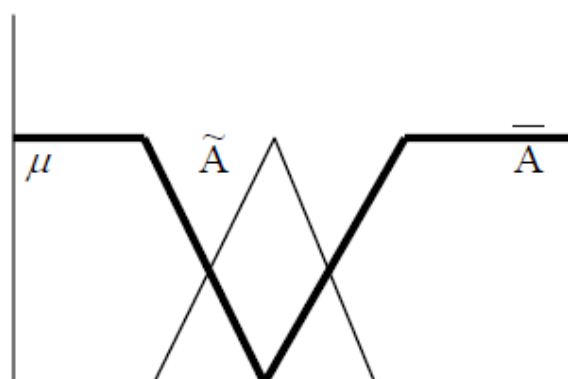


Figure 4.6: Complement of fuzzy set

Any fuzzy set $A \sim$ defined on the universe X is a subset of the universe. Also by definition null set has a membership 0 and x in X has a membership 1. Note that the null set and the whole set are not fuzzy sets.

Example4: A simple hollow shaft is 1-m radius and has a wall thickness of $(1/2\pi)$ m. The shaft is built up stacking a ductile section and a brittle section. A downward force P and a torque T are simultaneously applied to the shaft. The failure properties of the two sections can be described by the following fuzzy sets A and B for the ductile and brittle sections as follows:

$$\tilde{A} = \left\{ \frac{1}{2} + \frac{0.5}{3} + \frac{0.3}{4} + \frac{0.2}{5} \right\} \quad \text{and} \quad \tilde{B} = \left\{ \frac{0.5}{2} + \frac{0.7}{3} + \frac{0.2}{4} + \frac{0.4}{5} \right\}$$

We can see the following:

1. The set of loadings for which **either** material B **or** material D will be “safe” can be obtained by getting $A \sim \cup B \sim$
2. The set of loadings for which one expects that **both** material B **and** material D are “Safe” can be obtained by forming $A \sim \cap B \sim$.
3. The complements $A \sim$ and $B \sim$ represents the set of loadings for material D and B are unsafe.
4. $A \sim | B \sim$ gives the set of loadings for which the ductile material is safe but the brittle is not.
5. $B \sim | A \sim$ gives the set of loadings for which the brittle material is safe but the ductile not.
6. De Morgans laws can be used to find $A \sim \cap B \sim = A \sim \cup B \sim$ which asserts that the loadings that are not safe with respect to both materials are the union of the those that are unsafe with respect to the brittle material with those that are unsafe for with respect to the ductile material.
7. De Morgans law $A \sim \cup B \sim = A \sim \cap B \sim$ asserts that the loads that are safe for neither material D nor material B are the intersection of those that are unsafe for material D with those that are unsafe for material B.

Consequently, we can find the following:

$$\tilde{A} = \left\{ \frac{1}{2} + \frac{0.5}{3} + \frac{0.3}{4} + \frac{0.2}{5} \right\} \quad \text{and} \quad \tilde{B} = \left\{ \frac{0.5}{2} + \frac{0.7}{3} + \frac{0.2}{4} + \frac{0.4}{5} \right\}$$

Complements:

$$\overline{\tilde{A}} = \left\{ \frac{1}{1} + \frac{0}{2} + \frac{0.5}{3} + \frac{0.7}{4} + \frac{0.8}{5} \right\} \quad \text{and} \quad \overline{\tilde{B}} = \left\{ \frac{0.5}{2} + \frac{0.3}{3} + \frac{0.8}{4} + \frac{0.6}{5} \right\}$$

Union :

$$\tilde{A} \cup \tilde{B} = \left\{ \frac{1}{2} + \frac{0.7}{3} + \frac{0.3}{4} + \frac{0.4}{5} \right\}$$

Intersection :

$$\tilde{A} \cap \tilde{B} = \left\{ \frac{0.5}{2} + \frac{0.5}{3} + \frac{0.2}{4} + \frac{0.2}{5} \right\}$$

Difference :

$$\tilde{A} \setminus \tilde{B} = \tilde{A} \cap \overline{\tilde{B}} = \left\{ \frac{0.5}{2} + \frac{0.3}{3} + \frac{0.3}{4} + \frac{0.2}{5} \right\}$$

etc.

4.4 Membership functions

The **membership function** of a fuzzy set is a generalization of the indicator function in classical sets. In fuzzy logic, it represents the degree of truth as an extension of valuation. Degrees of truth are often confused with probabilities, although they are conceptually distinct, because fuzzy truth represents membership in vaguely defined sets, not likelihood of some event or condition. Membership functions were introduced by Zadeh in the first paper on fuzzy sets (1965).

For any set X , a membership function on X is any function from X to the real unit interval $[0, 1]$. Membership functions on X represent fuzzy subsets of X . The membership function which represents a fuzzy set \tilde{A} is usually denoted by μ_A . For an element x of X , the value $\mu_A(x)$ is called the *membership degree* of x in the fuzzy set \tilde{A} . The membership degree $\mu_A(x)$ quantifies the grade of membership of the element x to the fuzzy set \tilde{A} . The value 0 means that x is not a member of the fuzzy set; the value 1 means that x is fully a member of the fuzzy set. The values between 0 and 1 characterize fuzzy members, which belong to the fuzzy set only partially.

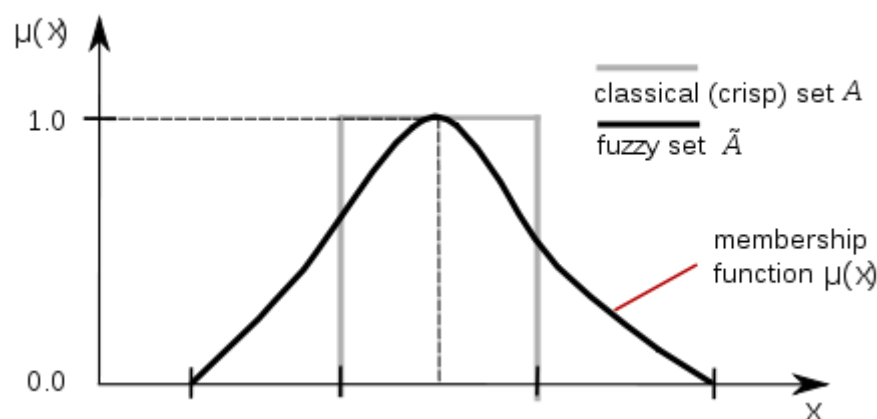


Figure 4.7 : Membership function of a fuzzy set

Sometimes, a more general definition is used, where membership functions take values in an arbitrary fixed algebra or structure L ; usually it is required that L be at least a poset or lattice. The usual membership functions with values in $[0, 1]$ are then called $[0, 1]$ -valued membership functions.

The membership function of Fuzzy set can be defined as a curve that defines how each point in universe of discourse maps to a membership value (or degree of membership) between 0 and 1. A Fuzzy set is fully defined by its membership function but for most Fuzzy logic control problems, the assumption is that the membership functions are piecewise linear and usually triangular in shape. This means that the values to be determined are the parameters defining the triangles; the parameters, in turn, are usually based on the control engineer's experience. However, for many applications, especially where linguistic terms are involved, triangular membership functions are not the most appropriate as they do not represent accurately the linguistic terms being modeled. In such applications, the Gaussian and exponential membership functions provide better representations.

Determination of Membership Functions

According to Watanabe (1979), the determination of membership function can be either manual or automatic. Manual statistical techniques for determining membership functions fall into two broad categories: use of frequencies and

direct estimation. However, such manual techniques can be deficient since they usually rely on subjective interpretation of words, are subject to the inadequacies of human experts, and generally suffer from other documented problems associated with the knowledge acquisition process. The automatic generation of membership function differs significantly from the manual methods in that either the expert is completely removed from the process, or the membership functions are 'fine-tuned' based on an initial guess by the expert. The emphasis in this case is on the use of Genetic Algorithm (GA) and Artificial Neural Networks (ANN).

As Takagi and Hayashi (1991) first pointed out, Fuzzy reasoning presents two specific challenges, namely (1) the lack of a definite method for determining the membership function, and (2) the lack of a learning function. Takagi and Hayashi then went ahead to describe an approach for using ANNs to overcome these problems. The method investigates *If/Then* rules by using neural networks to determine the membership functions of the antecedent and then determine the consequent component at the output for each rule. The approach used is to take some raw data sample (say, in a control problem), apply a conventional clustering algorithm to group the data into clusters, and thereafter apply an ANN to this clustered data to determine the membership of a pattern within a particular Fuzzy set. Wang (1994) in an alternative approach builds on the information provided by an expert and uses ANNs to fine-tune the membership function. In other words, the pairs (x, y) that describes the relationship between X and Y is presented to the neural network, which fits a function to the points.

In an interesting contribution, Meredith et al (1992) applied GA to the fine-tuning of membership functions in a Fuzzy logic controller for a helicopter. An initial guess for the membership function is made by the Control Engineer, and then the GA adjusts the parameters that define the functions by using them to minimize the movement of a hovering helicopter. For this case, triangular membership functions were used. Karr (1991) also applied GA to the design of Fuzzy logic controller for the "Cart Pole" problem. However, the membership function used here is Gaussian in nature, and the objective is to

minimize an objective function that minimizes the squared difference between a Cart and the centre of the track that the Cart is on, while keeping the pole balanced at the same time. Lee and Takagi (1993) also tackled the Cart problem, adopting a holistic approach by using GA to design the whole system (i.e. determine the optimal number of rules as well as the membership functions).

Ross (2007) reported on six popular methods for developing membership functions namely: intuition, inference, rank ordering, neural networks, genetic algorithms and inductive reasoning.

Olunloyo and Ajofoyinbo (2008), have, in a set of recent papers suggested an alternative approach for treatment of union and intersection of Fuzzy sets based on Fourier series representation of the membership functions.

The membership function of Fuzzy sets can take on any shape. A literature survey indicates that the following shapes of Fuzzy membership functions are commonly used: triangular, trapezoidal, exponential, Gaussian, and cosine (Badiru Adedeji, 2002). Although various functional profiles of membership functions could be used, the triangular and trapezoidal give approximations of the other forms and will thus receive further consideration in this work. The trapezoidal form can also be approximated by the triangular forms since the end-points of the 'tolerance' interval have the same grade of membership and could therefore be assigned a point value that represents the peak of the triangular profile. In any case, our methodology can also be adopted for other functional forms; as will be shown in a future paper. In general, we assume that the Fuzzy sets are triangular and symmetric (Fig 4.8)

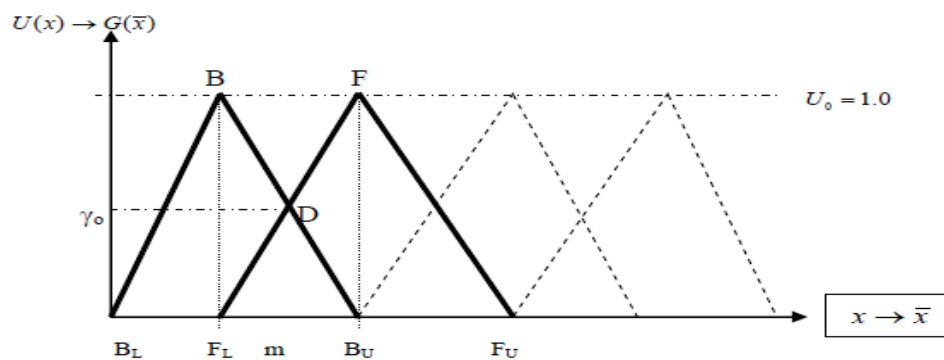


Figure 4.8: Triangular Fuzzy sets

In particular, the overlap domains are triangular and symmetric. For purposes of modeling, we adopted the following relations: The grade of membership function of Fuzzy set X , μ_X , is mapped to $f(x)$, and the data values (i.e. the universe of discourse) are mapped to x for the intersection and union as illustrated in Fig. 4.9 and Fig. 4.10, respectively, below:

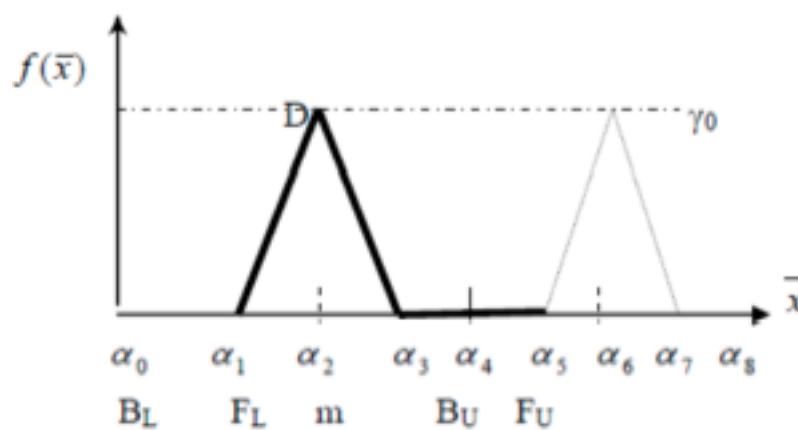


Figure 4.9: Triangular Pulses

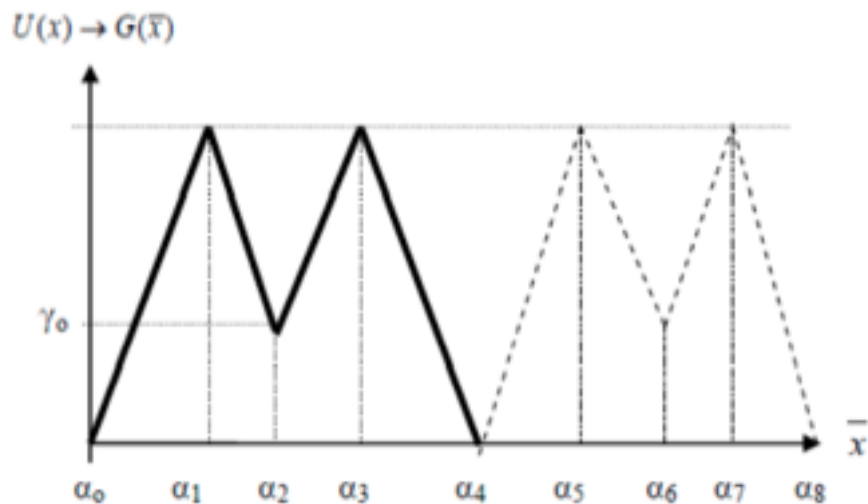


Figure 4.10 : Union of triangular fuzzy sets

4.5 Fuzzy rules

Fuzzy rules may be expressed in terms such as "If the room gets hotter, spin the fan blades faster" where the temperature of the room and speed of the fan's blades are both imprecisely (fuzzily) defined quantities, and "hotter" and "faster" are both fuzzy terms. Fuzzy logic, with fuzzy rules, has the potential to add human-like subjective reasoning capabilities to machine intelligences, which are usually based on bivalent Boolean logic

For the most part, the fuzzy rules that are used in control systems are hand-crafted by the designers of the systems, and machine learning is rarely employed. As such, it can be argued that this human input into the system's design constitutes a homunculus, and that such systems can never be independently intelligent.

The compactness of the rules is desirable because there exists evidence suggesting smaller rules perform better Holte, 1993, with reasons essentially the same as those for over fitting in decision trees - the underlying structure of the process generating the data is captured rather than the superficial structure of the training data. Fuzzy logic appears to be very well suited to the creation of small rules, as fuzzy rules have a higher "information density", so to speak - each rule encapsulates a richness of information and meaning.

The syntax of the rules is convenient for control purposes, but much too restrictive for fuzzy reasoning; defuzzification and defuzzification are automatic and inescapable. There are several development environments available for constructing fuzzy control systems. A typical fuzzy control rule might be. IF input1 is high AND input2 is Low THEN output is Zero.

Rules for fuzzy reasoning cannot be described so compactly. The application domain of fuzzy control systems is well defined; they work very satisfactorily with input and output restricted to numbers. But the domain of fuzzy reasoning systems is not well defined; by definition, fuzzy reasoning systems attempt to emulate human thought, with no a priori restrictions on that thought. Fuzzy control systems deal with numbers; fuzzy reasoning systems can deal with both numeric and non-numeric data. Inputs might be temperature and pulse, where temperature might 38.58C and pulse might be 110 and “thready”, where “thready” is clearly non-numeric. Output might be “CBC” and “Admit” and “transfer MICU” (not very realistic, but illustrates non-numeric data input and output). Accordingly rules for fuzzy reasoning do not make fuzzification and defuzzification automatic and inescapable; they may be broken out as separate operations that may or may not be performed as the problem requires.

The syntax of fuzzy reasoning rules accepts a wide variety of rule types. Here are two. 1. IF symptom is Depressive and duration is about 6) THEN diagnosis is Major depression; This rule resembles a fuzzy control rule, but it is actually quite different. In a fuzzy control rule, “symptom” would be a scalar number; in the fuzzy reasoning rule, “symptom is a (fuzzy) set of linguistic terms of which “depressive” is a member. Similarly, in a fuzzy control rule “diagnosis” would be a scalar number; in the fuzzy reasoning rules, “diagnosis” is a (fuzzy) set of diagnoses of which “depressive” is a member.

Background knowledge of any fuzzy system is expressed in fuzzy logic in the form of “If-Then” rules. These rules connect linguistic terms of several linguistic variables to infer decision (outputs). The **If-part** of a rule is called the precondition or antecedent and the **Then-part** is called conclusion or

consequent. The If-part can consist of several preconditions joined by linguistic connectors like AND and OR.

Fuzzy Rule Inference

Let us begin from examples.

Rule 1:

IF "interest rate" = low AND "trade fee" = low

THEN "environment (trade environment) = positive

Rule 2:

IF "interest rate" = low AND "trade fee" = significant

THEN "environment" = positive

Rule 3:

IF "interest rate" = high AND "trade fee" = significant

THEN "environment" = negative

Rule 4:

IF "interest rate" = medium AND "trade fee" = significant

THEN "environment" = indifferent

These rules use the AND operator in their IF-parts. Consider a trade in which the interest rate is 0.03 and the trade fee is 0.015. Fuzzification of this data would establish for rule 1 that for the first precondition and for the second precondition. $\mu_{\text{low-interest-rate}}(0.03) = 0.72$ $\mu_{\text{trade-fee}}(0.015) = 0.78$

Fuzzy rule inference consists of two parts: **aggregation** (computing the truth value of the IF-part of each rule) and composition (computing the truth value of the **conclusion** of the set of rules).

The typical fuzzy logic assumption in aggregation is truth-functionality: "the truth of complex sentences can be computed from the truth of the components. Probability combination does not work this way, except under strong independence assumptions" [Russell, Norvig, 1995]. Several operators were suggested for combining components, for instance, Min and Product are used as AND operation.

The minimum (MIN) operator used to compute the truth value of the entire condition in rule 1 produces the following truth value:

$$\text{MIN} \{ \text{Truth value ("interest rate" = low)}, \text{Truth value ("trade fee" = low)} \} = \text{MIN} \{ 0.72, 0.78 \} = 0.72.$$

Similarly, truth value of rule 2 can be calculated:

$$\text{MIN} \{ \text{Truth value ("interest rate" = low)}, \text{Truth value ("trade fee" = significant)} \} = \text{MIN}\{0.72, 0.22\} = 0.22.$$

If more than one rule produce the same conclusion (e.g. "environment" = positive), the maximum of truth values of the conclusions is selected for all further processing. This is actually representation for logical OR operator.

Example:

Rules 1 and 2 yield the same result for "environment", but with different truth values of the conditions:

$$\begin{aligned} m_{\text{positive-environment}}(0.03, 0.015) &= \min(0.72, 0.78) = 0.72 & (\text{rule 1}) \\ m_{\text{positive-environment}}(0.03, 0.015) &= \min(0.72, 0.22) = 0.22 & (\text{rule 2}) \end{aligned}$$

The composition step will produce the truth value $\text{MAX}(0.72, 0.22) = 0.72$ for the pair $((0.03, 0.015))$ of interest rate and trade fee. The above used combination of max-min operators is called **MAX-MIN inference**.

Another standard for fuzzy logic is **MAX-PROD**. This method produces the following output for the previous example:

$$\begin{aligned} m_{\text{positive-environment}}(0.03, 0.015) &= 0.72 * 0.78 = 0.56 & (\text{rule 1}) \\ m_{\text{positive-environment}}(0.03, 0.015) &= 0.72 * 0.22 = 0.16 & (\text{rule 2}) \end{aligned}$$

The third operation is called **BSUM (bounded sum)**. The result is equal to 1, if the sum exceeds 1; else it is equal to the sum. For the same example this method delivers:

$$\begin{aligned} m_{\text{positive-environment}}(0.03, 0.015) &= \text{BSUM}(0.72, 0.78) = 1 & (\text{rule 1}) \\ m_{\text{positive-environment}}(0.03, 0.015) &= 0.72 + 0.22 = 0.94 & (\text{rule 2}) \end{aligned}$$

Many practical applications have shown that these inference methods can be used more or less **interchangeably**, depending on which defuzzification method is used [Von Altrock, 1997]. This empirical observation has some

theoretical explanation [Kovalerchuk, Dalabaev, 1994]. It was found under some assumptions that different inference methods produce close final orderings of alternatives (x,y) with a difference about 10%. In the example above, we need to order alternatives as a pair (interest rate, trade fee) with respect to the ordering relation “better trade environment”

For instance,

$$(0.06, 0.02) <_{\text{trade-environment}} (0.03, 0.015)$$

means that (0.03, 0.015) corresponds to a better trade environment than (0.06, 0.02)

Fuzzy Operators

All listed operators are truth-functional as we already mentioned. However, in real financial and many other applications, this assumption may not be true. Therefore, several other operators were suggested.

These operators like the GAMMA operator have adjustable parameters to fit a particular decision-making task. These operators are called **compensatory operators**. **GAMMA** and **MIN-AVG** belong to this class of operators.

For instance, the following compensatory operator $g(x,y)$ can be considered:

$$G_{\alpha,\beta}(x,y) = \alpha * \text{MIN}(x,y) + \beta * \text{MAX}(x,y)$$

Linguistic result design

At the end of fuzzy rule inference, all output variables are associated with a fuzzy value. To exemplify this, the following truth values are assigned to the trade environment:

$$\begin{aligned} m_{\text{Positive-trade-environment}}(0.03, 0.015) &= 0.72 \\ m_{\text{Indifferent-trade-environment}}(0.03, 0.015) &= 0.2 \\ m_{\text{Negative-trade-environment}}(0.03, 0.015) &= 0.0 \end{aligned}$$

Next, an extended set of linguistic terms can be developed to capture linguistically trade environment expressed with three numbers (0.72; 0.2; 0).

There can be the term “positive, but slightly indifferent trade environment”.

Rule Definition

There are two major stages in defining rules:

A) Formulate pure linguistic rule (prototypes) and

B) Formulate the prototype as a fuzzy rule.

Example:

Consider a statement:

IF the trade period has low interest rate and low trade fee THEN the trade environment is positive.

This can be formulated as a fuzzy rule:

IF "trade rate" = low AND "trade fee" = low THEN "environment" = positive.

Rule development is an iterative process, which involves tuning rules and the development of similar rules for other terms. For example, the following rule could be defined: "IF trade period has medium interest rate and it has the trade fee higher than the low trade fee then the trade environment is indifferent".

This can be formulated as a fuzzy rule too:

IF "interest rate" = medium AND "trade fee" = significant
THEN "environment" = indifferent

The next important concept in fuzzy logic is a **matrix of linguistic rules (Rule matrix)**. This matrix is presented in Table 4.1. This matrix is coded with numbers in Table 4.2. Code low interest rate as 1, medium as 0 and high as 1. For trade fee use code 1 for low and 0 for significant. For environment use 1 for positive, 0 for indifferent and -1 for negative. It seems that this coding is inconsistent.

If		Then
Interest Rate	Trade Fee	Environment
Low	Low	Positive
Medium	Low	Positive
High	Low	Indifferent
Low	Significant	Positive
Medium	Significant	Indifferent
High	Significant	Negative

Table 4.1: Matrix of linguistic rules

We do not code highest degrees for interest and trade fee with highest number (1). This is done deliberately. The suggested coding scheme allows

us to keep the meaningful property of monotonicity -- larger numbers reflects better interest rate, trade fee and trade environment.

The information from the table above, is visualized in Table 4.2. Here 1-1 on the left side represent the IF-part of rule (the interest rate is low and the trade fee is low); 01 means that the interest rate is medium and the trade fee is low. The right lattice shows in each node the IF-part of rules along with their THEN-part (environment value) too.

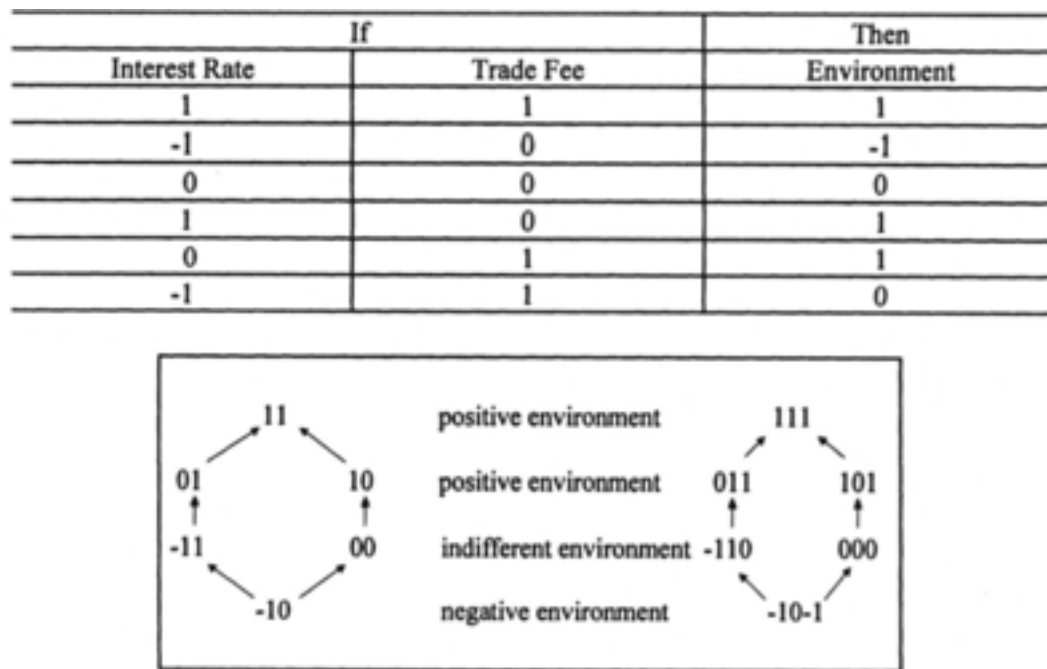


Table 4.2 Numerical Rule Table

Interest rate	Trade fee	
	1 (low)	0 (significant)
1 (low)	1	1
0 (medium)	1	0
-1 (high)	0	-1

Table 4.3: Rule matrix: Relation form

Defuzzification. Fuzzy rules, in contrast with rules in classical logic, produce fuzzy output. It can be a set of values of the membership functions values or a linguistic term. For example, the result could be equivalent to a complex linguistic statement "Trading environment is mostly positive for buying, but it is also slightly indifferent for buying". Obviously, program trade systems or traders cannot interpret such linguistic commands in the same way as crisp

buy/hold/sell signals. In fuzzy logic, membership functions are used to retranslate the fuzzy output into a crisp value. This **retranslating** is known as **defuzzification** summarized in Table 4.4 [Nauck et al, 1997; Passino, Yurkovoch, 1998, Von Altrrock, 1997].

Abbreviation	Name	Algorithm	Comment
CoA, CoG	Center-of-Area, Center-of-Gravity	Compute x such that areas on both sides of x are equal.	Slow computation
CoM	The Center of Maximum	A weighted mean of the term membership maxima, weighted by the inference results.	Fast, but neglects overlapping approximating CoA (CoG).
CoA BSUM		Boundary sum variant of CoA	Optimized for efficient VLSI implementation
MoM	The Mean-of-Maximum	Computes the mean of the truth values for the term with highest resulting truth value.	
MoM BSUM		BSUM Variant of MoM	Optimized for efficient VLSI implementation

Table 4.4 Procedures of difuzzification

4.6 FUZZY INFERENCE SYSTEMS

Fuzzy inference is the process of formulating the mapping from a given input to an output using fuzzy logic. The mapping then provides a basis from which decisions can be made, or patterns discerned. The process of fuzzy inference involves all of the pieces that are described in the previous sections: Membership Functions, Logical Operations, and If-Then Rules. We can implement two types of fuzzy inference systems: Mamdani-type and Sugeno-type. These two types of inference systems vary somewhat in the way outputs are determined.

Fuzzy inference systems have been successfully applied in fields such as automatic control, data classification, decision analysis, expert systems, and computer vision. Because of its multidisciplinary nature, fuzzy inference systems are associated with a number of names, such as fuzzy-rule-based systems, fuzzy expert systems, fuzzy modeling, fuzzy associative memory, fuzzy logic controllers, and simply (and ambiguously) fuzzy systems.

Mamdani's fuzzy inference method is the most commonly seen fuzzy methodology. Mamdani's method was among the first control systems built using fuzzy set theory. It was proposed in 1975 by Ebrahim Mamdani as an attempt to control a steam engine and boiler combination by synthesizing a set of linguistic control rules obtained from experienced human operators. Mamdani's effort was based on Lotfi Zadeh's 1973 paper on fuzzy algorithms for complex systems and decision processes. Although the inference process described in the next few sections differs somewhat from the methods described in the original paper, the basic idea is much the same.

4.6.1: Mamdani-type inference, expects the output membership functions to be fuzzy sets. After the aggregation process, there is a fuzzy set for each output variable that needs defuzzification. It is possible, and in many cases much more efficient, to use a single spike as the output membership function rather than a distributed fuzzy set. This type of output is sometimes known as a *singleton* output membership function, and it can be thought of as a pre-defuzzified fuzzy set. It enhances the efficiency of the defuzzification process because it greatly simplifies the computation required by the more general Mamdani method, which finds the centroid of a two-dimensional function. Rather than integrating across the two-dimensional function to find the centroid, you use the weighted average of a few data points. Sugeno-type systems support this type of model. In general, Sugeno-type systems can be used to model any inference system in which the output membership functions are either linear or constant.

Overview of Fuzzy Inference Process

In the following example we examine the services in a restaurant. The variables under examination are, the quality of service and the quality of food, which are used as the input to our fuzzy system. The fuzzy rules examine the quality of the variables, and produce as an output the tip that we are willing to give. The basic structure of this example is shown in the following diagram:

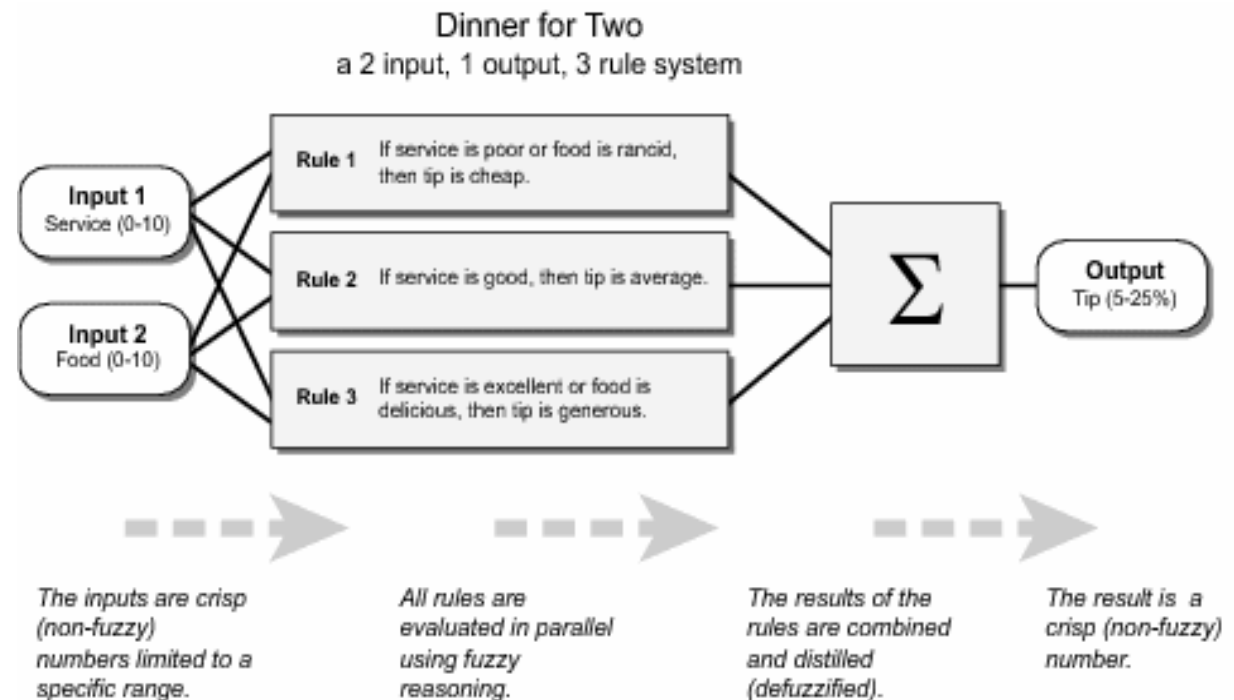


Figure 4.11 Example of 3-rule fuzzy system

Information flows from left to right, from two inputs to a single output. The parallel nature of the rules is one of the more important aspects of fuzzy logic systems. Instead of sharp switching between modes based on breakpoints, logic flows smoothly from regions where the system's behavior is dominated by either one rule or another.

Fuzzy inference process comprises of five parts: fuzzification of the input variables, application of the fuzzy operator (AND or OR) in the antecedent, implication from the antecedent to the consequent, aggregation of the consequents across the rules, and defuzzification. These sometimes cryptic and odd names have very specific meaning that is defined in the following steps.

Step 1. Fuzzify Inputs

The first step is to take the inputs and determine the degree to which they belong to each of the appropriate fuzzy sets via membership functions. The input is always a crisp numerical value limited to the universe of discourse of the input variable (in this case the interval between 0 and 10) and the output

is a fuzzy degree of membership in the qualifying linguistic set (always the interval between 0 and 1). Fuzzification of the input amounts to either a table lookup or a function evaluation.

This example is built on three rules, and each of the rules depends on resolving the inputs into a number of different fuzzy linguistic sets: service is poor, service is good, food is rancid, food is delicious, and so on. Before the rules can be evaluated, the inputs must be fuzzified according to each of these linguistic sets. For example, to what extent is the food really delicious? The following figure shows how well the food at the hypothetical restaurant (rated on a scale of 0 to 10) qualifies, (via its membership function), as the linguistic variable delicious. In this case, we rated the food as an 8, which, given your graphical definition of delicious, corresponds to $\mu = 0.7$ for the delicious membership function.

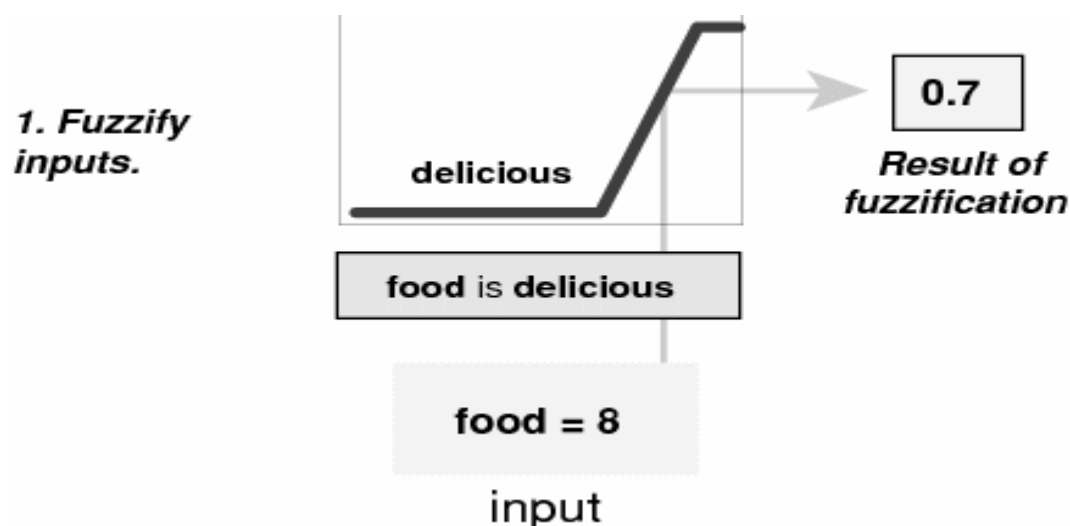


Figure 4.12: membership function

In this manner, each input is fuzzified over all the qualifying membership functions required by the rules.

Step 2. Apply Fuzzy Operator

After the inputs are fuzzified, we know the degree to which each part of the antecedent is satisfied for each rule. If the antecedent of a given rule has more than one part, the fuzzy operator is applied to obtain one number that

represents the result of the antecedent for that rule. This number is then applied to the output function. The input to the fuzzy operator is two or more membership values from fuzzified input variables. The output is a single truth value.

The following figure shows the OR operator *max* at work, evaluating the antecedent of the rule 3 for the tipping calculation. The two different pieces of the antecedent (service is excellent and food is delicious) yielded the fuzzy membership values 0.0 and 0.7 respectively. The fuzzy OR operator simply selects the maximum of the two values, 0.7, and the fuzzy operation for rule 3 is complete. The probabilistic OR method would still result in 0.7.

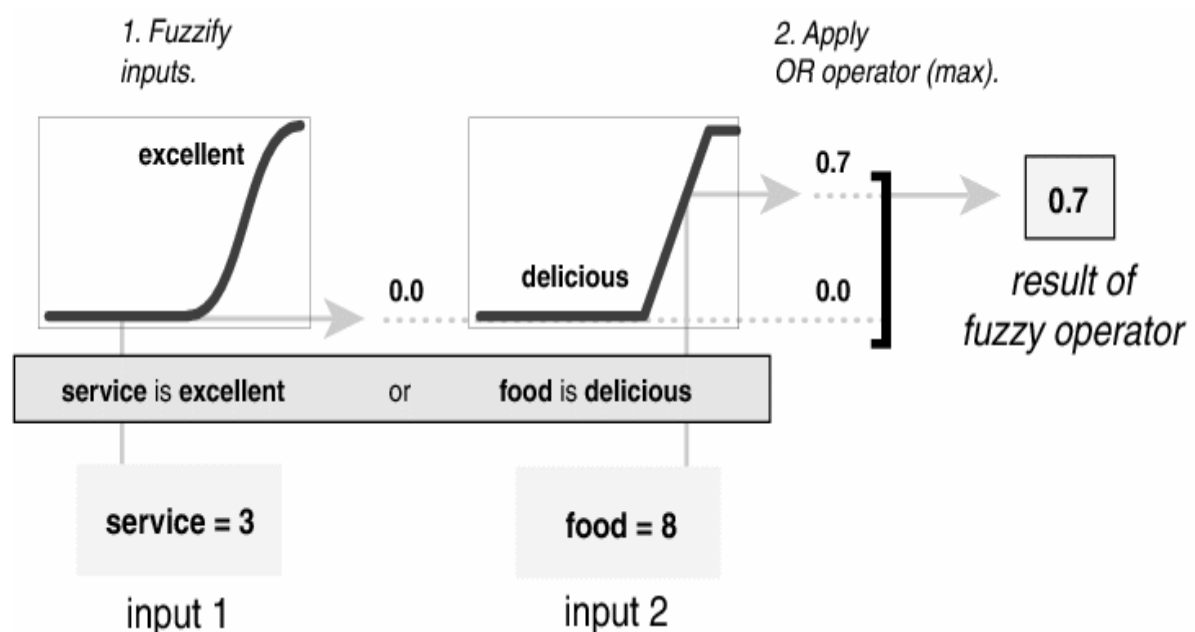
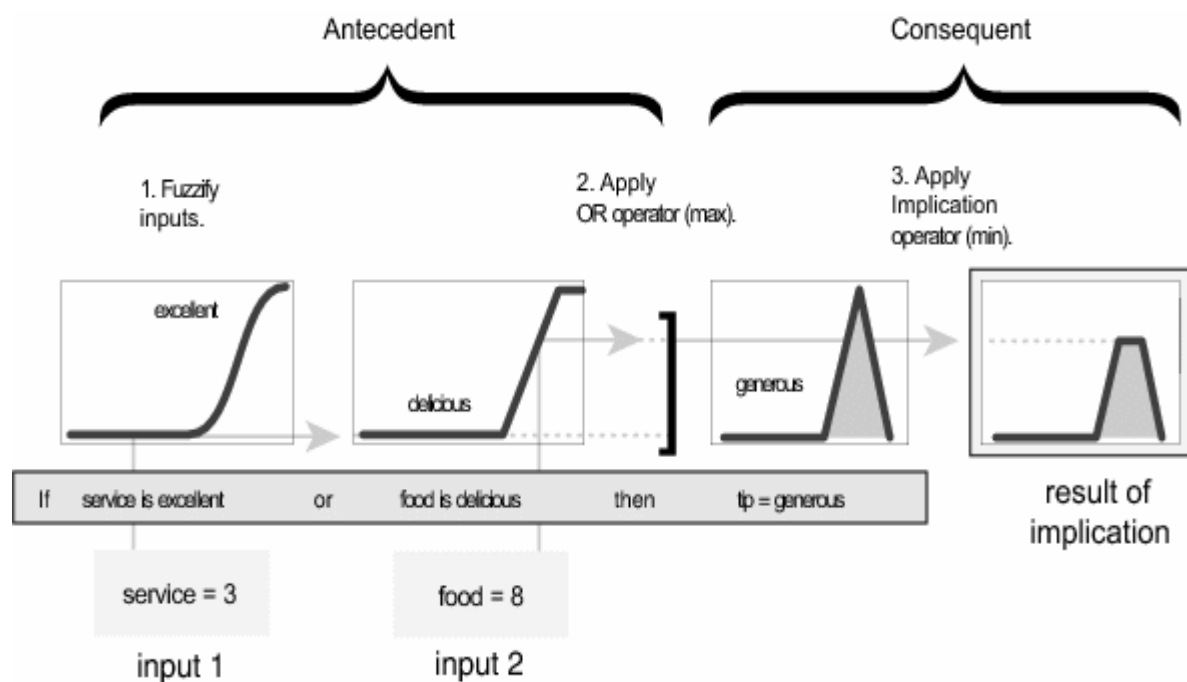


Figure 4.13: Membership functions

Step 3. Apply Implication Method

Before applying the implication method, we must determine the rule's weight. Every rule has a *weight* (a number between 0 and 1), which is applied to the number given by the antecedent. Generally, this weight is 1 (as it is for this example) and thus has no effect at all on the implication process. From time to time we may want to weight one rule relative to the others by changing its weight value to something other than 1.

After proper weighting has been assigned to each rule, the implication method is implemented. A consequent is a fuzzy set represented by a membership function, which weights appropriately the linguistic characteristics that are attributed to it. The consequent is reshaped using a function associated with the antecedent (a single number). The input for the implication process is a single number given by the antecedent, and the output is a fuzzy set. Implication is implemented for each rule.



Step 4. Aggregate All Outputs

Because decisions are based on the testing of all of the rules in a FIS, the rules must be combined in some manner in order to make a decision. Aggregation is the process by which the fuzzy sets that represent the outputs of each rule are combined into a single fuzzy set. Aggregation only occurs once for each output variable, just prior to the fifth and final step, defuzzification. The input of the aggregation process is the list of truncated output functions returned by the implication process for each rule. The output of the aggregation process is one fuzzy set for each output variable.

As long as the aggregation method is commutative (which it always should be), then the order in which the rules are executed is unimportant. Three methods are presented here:

- max (maximum)
- probor (probabilistic OR)
- sum (simply the sum of each rule's output set)

In the following diagram, all three rules have been placed together to show how the output of each rule is combined, or aggregated, into a single fuzzy set whose membership function assigns a weighting for every output (tip) value.

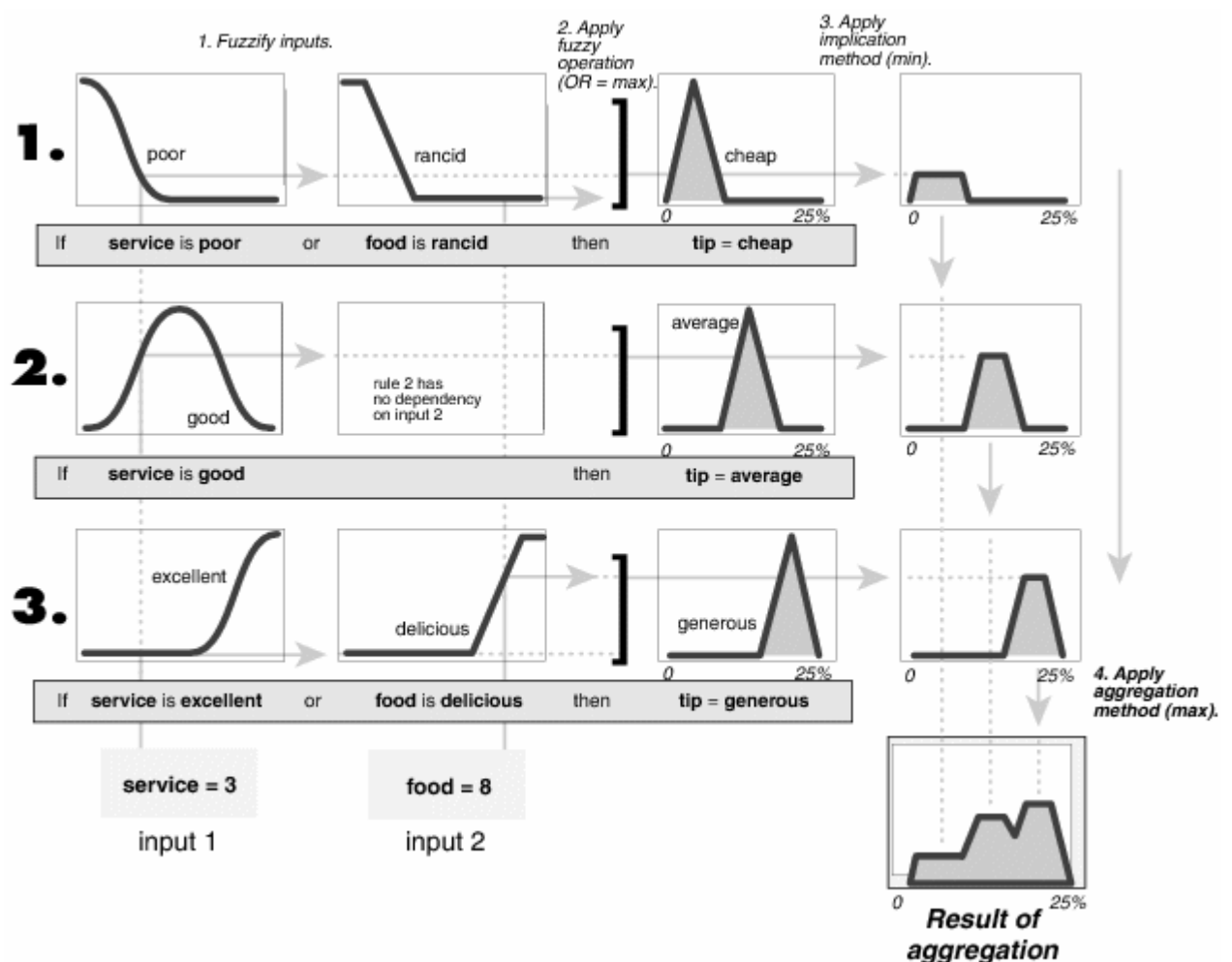


Figure 4.15: Membership functions and aggregation effect

Step 5. Defuzzify

The input for the defuzzification process is a fuzzy set (the aggregate output fuzzy set) and the output is a single number. As much as fuzziness helps the rule evaluation during the intermediate steps, the final desired output for each variable is generally a single number. However, the aggregate of a fuzzy set encompasses a range of output values, and so must be defuzzified in order to resolve a single output value from the set.

Perhaps the most popular defuzzification method is the centroid calculation, which returns the center of area under the curve. There are five well known methods for that: centroid, bisector, middle of maximum (the average of the maximum value of the output set), largest of maximum, and smallest of maximum.

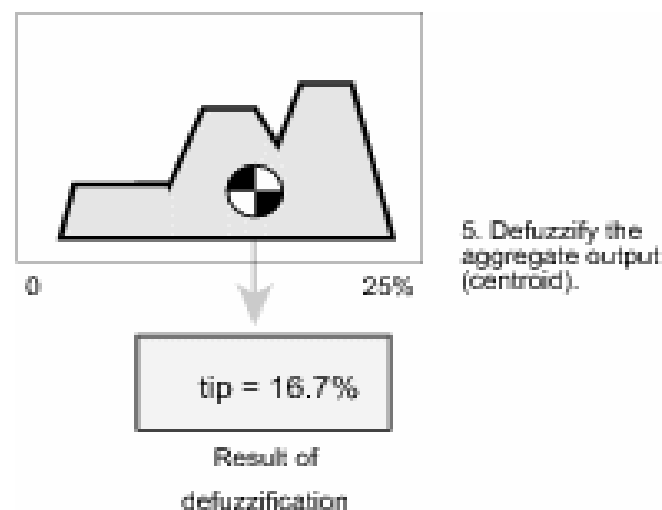


Figure 4.16: Defuzzification effect

The Fuzzy Inference Diagram

The fuzzy inference diagram is the composite of all the smaller diagrams presented so far in this section. It simultaneously displays all parts of the fuzzy inference process we have examined. Information flows through the fuzzy inference diagram as shown in the following figure.

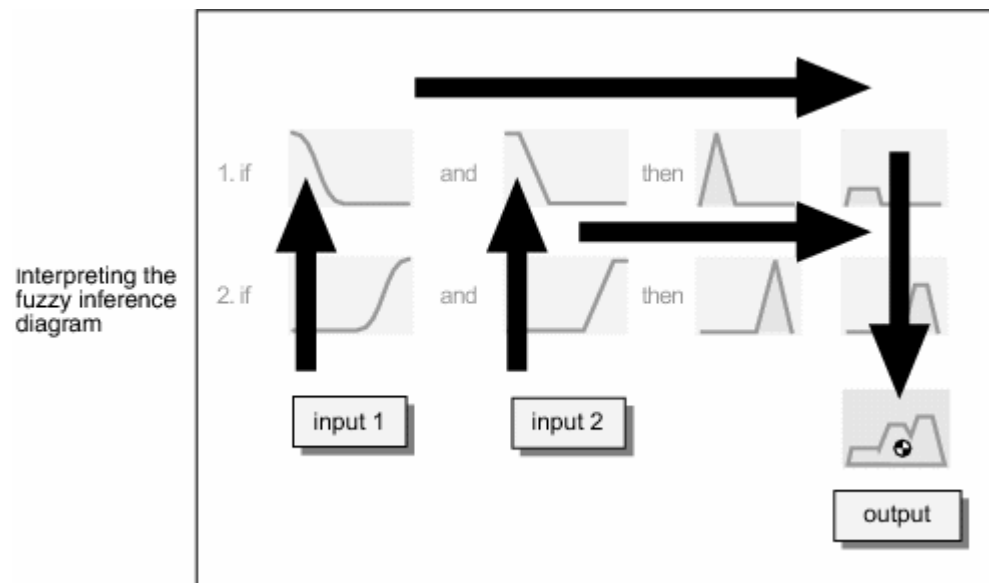
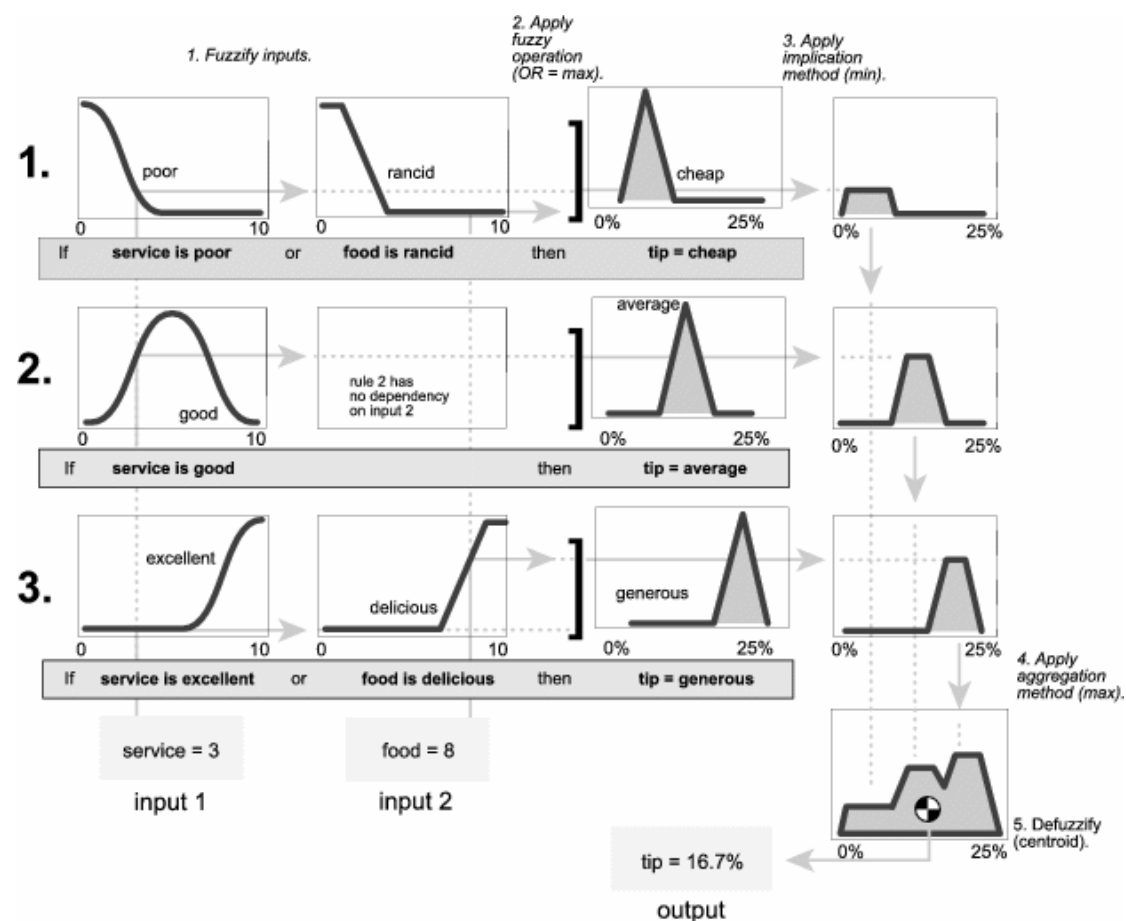


Figure 4.17: Fuzzy inference diagram

In this figure, the flow proceeds up from the inputs in the lower left, then across each row, or rule, and then down the rule outputs to finish in the lower right. This compact flow shows everything at once, from linguistic variable fuzzification all the way through defuzzification of the aggregate output.

The following figure shows the actual full-size fuzzy inference diagram. There is a lot to see in a fuzzy inference diagram, but after we become accustomed to it, we can learn a lot about a system very quickly. For instance, from this diagram with these particular inputs, we can easily see that the implication method is truncation with the *min* function. The *max* function is being used for the fuzzy OR operation. Rule 3 (the bottom-most row in the diagram shown

previously) is having the strongest influence on the output. And so on.



4.7. Applications of fuzzy logic in finance

This section covers a number of successful applications of fuzzy logic in finance. For example, Yamaichi Secures of Tokyo uses fuzzy logic to make decisions for an investment fund, and Nikko Secures of Yokohama uses a NeuroFuzzy system for a bond rating program [Houlder, 1994]. Several authors noticed that many financial institutions consider their systems based on fuzzy logic a proprietary technology and do not publicize details, or even the fact of implementation and use [Lee, Smith, 1995; Von Altrock, 1997]. However, many uses of fuzzy logic in finance are published in the literature, including those listed below:

- 1) Foreign-exchange trade support system in Japan with approximately 5000 fuzzy rules derived from a back propagation neural network [Rao,

Rao, 1993]. Fuzzy logic has been used in a foreign exchange trading System to predict the Yen-Dollar exchange rates [Yuize, 1991].

2) Analysis of market psychology using a fuzzy expert system with data Fuzzification and evaluation [Deny, 1993].

3) Insider Trading Surveillance [Moulder, 1994].

4) Fuzzy logic and variables in investing and trading [Caldwell, 1994].

5) Neural network and fuzzy logic hybrid system in finance [Derry, 1994, WongF, 1994].

6) Interpretation of neural network outputs using fuzzy logic: a fuzzy expert system is applied to the task of interpreting multiple outputs from a neural network designed to generate signals for trading the S&P 500 index [Caldwell, 1994b].

7) A portfolio insurance strategy of Japanese stocks based on Nikkei Stock Index Futures using fuzzy logic. The system assists in deciding when to Rebalance the replicating portfolio [Kay-Hwang and Woon-Seng Gan, 1996].

8) Financial modeling and forecasting using a hybrid Neural-Fuzzy system. The model's performance is compared with a random walk model, an ARIMA model and a variety of regression models [Pan et al, 1997].

9) Fuzzy Scoring for Mortgage Applicants.

10) Creditworthiness Assessment and Fraud Detection.

11) Investor Classification [Von Altrock, 1997].

12) Cash Supply Optimization [Von Altrock, 1997].

13) Prediction of stock market direction using fuzzy logic [Von Altrock, 1997]

14) Rating bonds [Loofbourrow, 1995].

Some of them are briefly presented below:

11) Investor Classification: Many investment institutions classify customers and investments into three risk groups:

a) Conservative and security-oriented (risk shy),
b) growth-oriented and dynamic (risk neutral), and
c) chance-oriented and progressive (risk happy). The fuzzy logic system was designed to evaluate how well a customer fits into these three groups. Each customer was represented by the set of 25 answers. Each question represents an attribute with five values from 1 to 5. The questions include personal background (age, marital state, number of children, job type, education type, etc.) the customer's expectation from an investment (capital protection, tax shelter, liquid assets, etc.) and others [Von Altrock, 1997].

3) Insider Trading Surveillance. Houlder [1994] describes a system developed for the London Stock Exchange. The goal of the system is to automatically detect insider dealing and market manipulation using a combination of fuzzy logic, neural nets, and genetic algorithms. The system tries to detect suspicious rings of individuals with several accounts in a vast amount of electronic camouflage.

1) Foreign Exchange Trading. Fuzzy logic has been used to predict the Yen-Dollar exchange rates [Yuize, 1991]. The system uses fuzzy logic rules to make inferences based on economic **news** events that may affect the currency market. This news is "translated" into the fuzzy logic system's input format by **domain experts**.

12) Cash Supply Optimization [Von Altrock, 1997]. Banks are interested in reaching two contradictory goals for each branch and ATM:

1. Minimize **unused cash** and
2. Minimize the rate of **out of cash** situations.

Cash surplus could be used for other profitable operations or/and decrease the cost of cash supply for branches and ATM. On the other hand, if the bank is able to minimize out of cash situations, it can better compete with other banks. The traditional expert solution is to set the **minimum amount of cash** for each branch and ATM. However, this minimum is not static, bank business conditions are changed dynamically for each individual branch and ATM due to:

- Seasonal factors (week, month, year) and
- Environmental factors (new shops, offices, banks nearby and so on)

Suppose the bank takes into account five such factors with only two values for each of them. This means that the bank should analyze regularly $10000 \times 32 = 32000$ alternatives to set up minimum amount of cash for its 1000 units. “In a project of a European bank, fuzzy logic was used to recompute the minimum cash amount of each branch and ATM **daily**. The system was able to reduce the average cash supply in the branches and ATMs by 7.1% without increasing the rate of situations where the branch or ATM ran out of cash. For a bank with about 450 branches and 1270 ATMs, this results in an average total of \$3.8M less in cash supply” [Von Altrock, 1997].

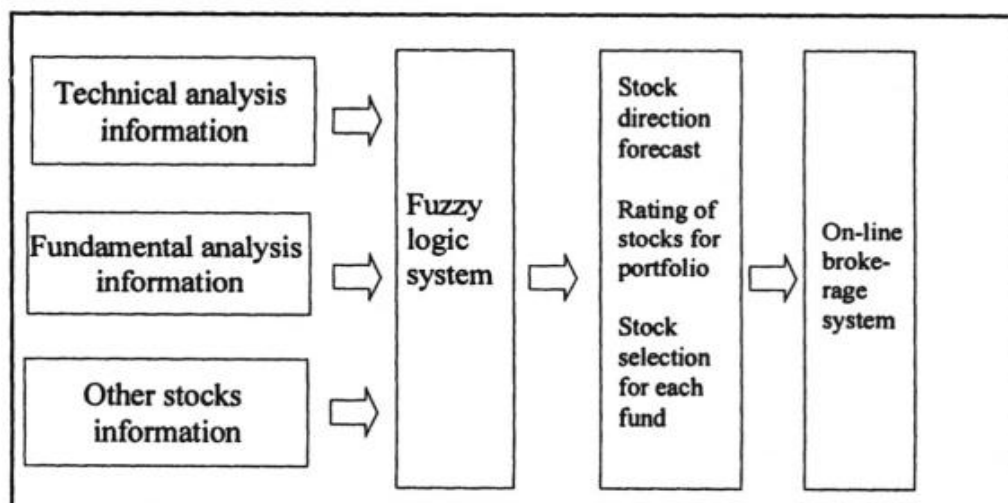


Figure 4.19 Fuzzy logic based system for trading

The system is based on three sources of information:

- 1) The past cash flow of the branches and ATMs,
- 2) The lowest cash amount suggested by the bank experts for each unit,
- 3) Classification of ATMs and branches according to the properties of the neighborhoods.

CHAPTER 5

Neuro-Fuzzy Systems

5.1 Basics

Since the moment that fuzzy systems become popular in industrial and financial application, the community perceived that the development of a fuzzy system with good performance is not an easy task. The problem of finding membership functions and appropriate rules is frequently a tiring process of attempt and error. This led to the idea of applying learning algorithms to the fuzzy systems. The neural networks, that have efficient learning algorithms, had been presented as an alternative to automate or to support the development of tuning fuzzy systems. The first studies of the neuro-fuzzy systems date of the beginning of the 90's decade, with Jang, Lin and Lee in 1991, Berenji in 1992 and Nauck from 1993, etc. The majority of the first applications were in process control. Gradually, their applications spread to all the areas of knowledge, like, data analysis, data classification, imperfections detection, support to decision-making, etc.

Neural networks and fuzzy systems can be combined to join their advantages and to cure their individual illnesses. Neural networks introduce their computational characteristics of learning in the fuzzy systems and receive from them the interpretation and clarity of systems representation. Thus, the disadvantages of the fuzzy systems are compensated by the capacities of the neural networks. These techniques are complementary, which justifies their use together. The aim of neuro-fuzzy systems (Fig. 5.1) is to combine collectively the benefits of both neural networks and fuzzy logic. Simply, the operation of the system is expressed as linguistic fuzzy expressions

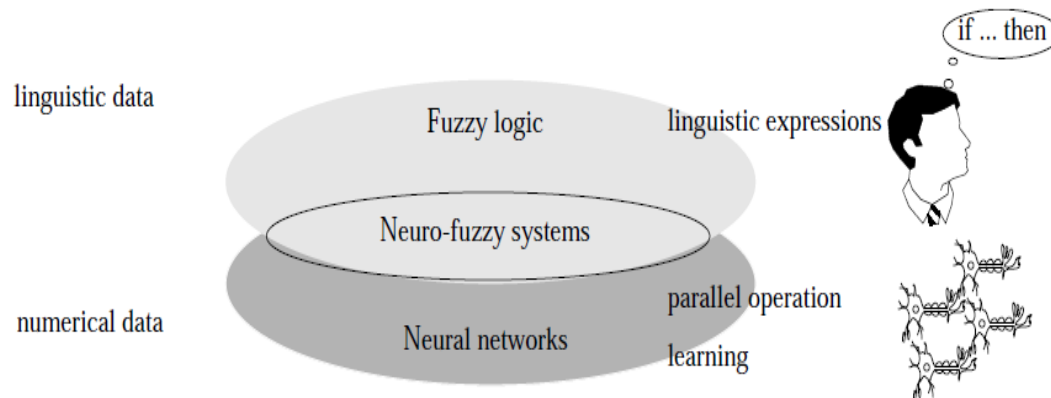


Figure 5.1: Neuro-fuzzy systems integrate fuzzy logic and neural networks.

and learning schemes of neural networks are used to learn the system. In addition, neuro-fuzzy systems allow incorporation of both numerical and linguistic data into the system. The neuro-fuzzy system is also capable of extracting fuzzy knowledge from numerical data.

The neuro-fuzzy systems can be divided into two main groups:

1. Neural fuzzy inference systems and
2. Fuzzy neural networks.

The origin of neural fuzzy inference systems is to incorporate neural concepts, such as learning and parallelism, into fuzzy logic inference systems (fuzzy controllers, in the context of control applications). Neural fuzzy inference systems realize fuzzy inference. The architecture of the systems are parallel, and they exploit the same learning algorithms, which are used with neural networks. The fuzzy inference can be implemented in two ways. The most common approach is to use one network which realizes the whole fuzzy inference. In the second approach, each fuzzy rule is realized using a neural network, when the fuzzy inference is the result of several neural networks. In the fuzzy neural networks, the fuzzy ideas are incorporated into neural networks.

Lee and Lee are the first who worked on neuro-fuzzy systems, especially on fuzzy neural networks. Their approach replaced the weighted sum of the McCulloch-Pitts neuron by a corresponding fuzzy operation. The operation of

their neuro-fuzzy system was exactly the same as the McCulloch-Pitts neural network. Unfortunately, they did not give any training rules for the network. The fuzzy neural networks consist of two components in same system: a fuzzy system and a neural network. The fuzzy system can be either a fuzzy inference block which converts linguistic information for the neural network or the neural network can drive the fuzzy inference block.

A common approach to neuro-fuzzy systems is to fuzzify the learning algorithms of different neural networks paradigms. For example, Huntsberger and Ajjimarangsee have proposed that the learning rate coefficient of Kohonen's self-organizing map is treated as a fuzzy membership value of the current input sample in the class of each neuron. The membership values are computed with the fuzzy *c*-means algorithm. Tsao *et al.* have extended their ideas to a family of Kohonen's SOM algorithms. Similarly, Bedzek *et al.* and Chung and Lee have extended this approach to Kohonen's learning vector quantization algorithm.

5.2. Types of Neuro-Fuzzy Systems

In general, all the combinations of techniques based on neural networks and fuzzy logic can be called neuro-fuzzy systems. The different combinations of these techniques can be divided, in accordance with, in the following classes:

Cooperative Neuro-Fuzzy System: In the cooperative systems there is a pre-processing phase where the neural network mechanisms of learning determine some sub-blocks of the fuzzy system. For instance, the fuzzy sets and/or fuzzy rules (fuzzy associative memories or the use of clustering algorithms to determine the rules and fuzzy sets position). After the fuzzy sub-blocks are calculated the neural network learning methods are taken away, executing only the fuzzy system. In a cooperative system the neural networks are only used in an initial phase. In this case, the neural networks determine sub-blocks of the fuzzy system using training data. After this, the neural networks are removed and only the fuzzy system is executed. In the cooperative neuro-fuzzy systems, the structure is not totally interpretable which can be considered as a disadvantage.

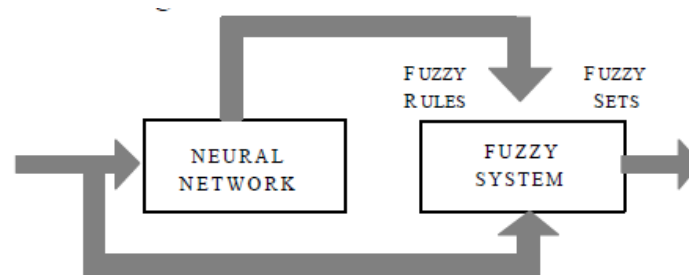


Figure 5.2: Cooperative neuro fuzzy system

Concurrent Neuro-Fuzzy System: In the concurrent systems the neural network and the fuzzy system work continuously together. In general, the neural networks pre-processes the inputs (or post-processes the outputs) of the fuzzy system. A concurrent system is not a neuro-fuzzy system in the strict sense, because the neural network works together with the fuzzy system. This means that the inputs entered in the fuzzy system, are pre-processed and then the neural network processes the outputs of the concurrent system or vice versa. In the concurrent neuro-fuzzy systems, the results are not completely interpretable, fact that can be considered as their major disadvantage.

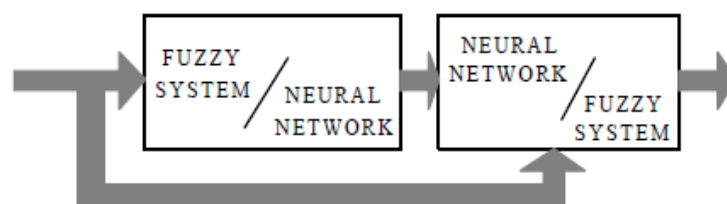


Figure 5.3: Concurrent neuro-fuzzy system

Hybrid Neuro-Fuzzy System: In this category, a neural network is used to learn some parameters of the fuzzy system (parameters of the fuzzy sets, fuzzy rules and weights of the rules) of a fuzzy system in an iterative way. The majority of the researchers use the neuro-fuzzy term to refer only to hybrid neuro-fuzzy systems. According to Nauck's definition: "A hybrid neuro-fuzzy system is a fuzzy system that uses a learning algorithm based on gradients or inspired by the neural networks theory (heuristic learning strategies) to

determine its parameters (fuzzy sets and fuzzy rules) through the patterns processing (input and output)". A neuro-fuzzy system can be interpreted as a set of fuzzy rules. This system can be totally created from input-output data or initialized with the *à priori* knowledge in the same way as fuzzy rules.

A neuro-fuzzy system represents the knowledge-base of a rule-based expert system.

Most of the neuro-fuzzy systems have mainly three layers:

1) Fuzzification Layer:

Fuzzification as discussed in previous chapters is a process to convert the discrete values of the input variables to the corresponding fuzzy sets. Each node in Fuzzification layer is a membership function denoting a fuzzy linguistic variable such as SMALL, MEDIUM, LARGE. This layer represents the variables in the antecedent part of an expert rule.

2) Inference Layer:

This part produces the strength of the fuzzy rule. In other words, the strength of the antecedent part is computed by using an inference mechanism. AND operator is used to form the relationship in the antecedent part in the previous rule example. The AND represents the multiplication of the membership values of the variables to some linguistic values given in the fuzzification layer.

3) Defuzzification Layer:

Defuzzification is the inverse process of Fuzzification. It deduces an output value from the rule antecedents. In this layer, the result or the value of consequent part is computed.

There are several examples of neuro-fuzzy systems in literature. The most important ones are briefly presented below, while this thesis focuses only on the ANFIS (Adaptive Neuro Fuzzy Inference Systems), whose applications in

finance and stock market prediction have been significant over the last few years.

A) **FALCON Architecture:**

The **Fuzzy Adaptive Learning Control Network**, FALCON's architecture, consists of five layers as it is shown in figure 5.4. There are two linguistic nodes for each output. One is for the patterns and the other is for the real output of the FALCON. The first hidden layer is responsible for the mapping of the input variables relatively to each membership function. The second hidden layer defines the antecedents of the rules followed by the consequents in the third hidden layer. FALCON uses a hybrid learning algorithm composed by unsupervised learning to define the initial membership functions and initial rule base, and a learning algorithm based on the gradient descent to optimize/adjust the final parameters of the membership functions to produce the desired output.

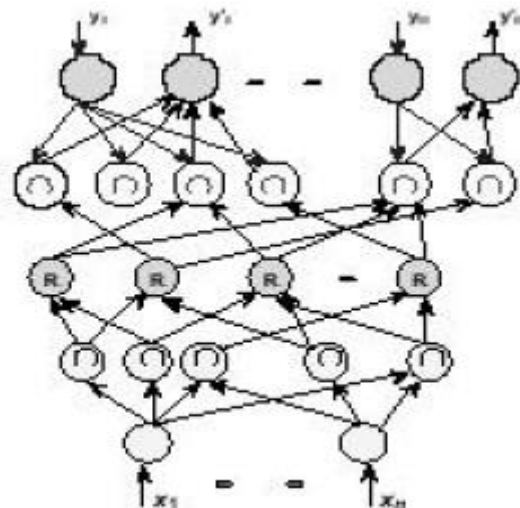


Figure 5.4: FALCON architecture

B) **GARIC Architecture:**

The **Generalized Approximate Reasoning based Intelligence Control** GARIC, implements a neuro-fuzzy system using two neural network modules, ASN (Action Selection Network) and AEN (Action Evaluation Network). The

AEN is an adaptive evaluator of ASN actions. The ASN of the GARIC is an advanced network of five layers. Figure 5.5 illustrates GARIC-ASN structure. The connections between the layers are not weighted. The first hidden layer stores the linguistics values of all input variables. Each input can only connect to the first layer, which represents its associated linguistics values. The second hidden layer represents the fuzzy rule nodes that determine the compatibility degree of each rule using a softmin operator. The third hidden layer represents the linguistic values of the output variables. The conclusions of each rule are calculated depending on the strength of the rules antecedents calculated in the rule nodes. GARIC uses the local mean of maximum method to calculate the output of the rules. This method produces a numerical value in the exit of each rule. Thus, the conclusions are transformed from fuzzy values to numerical values before being accumulated in the final output value of the system. GARIC uses a mixture of gradient descending and reinforcement learning algorithms to achieve a fine adjustment of its internal parameters.

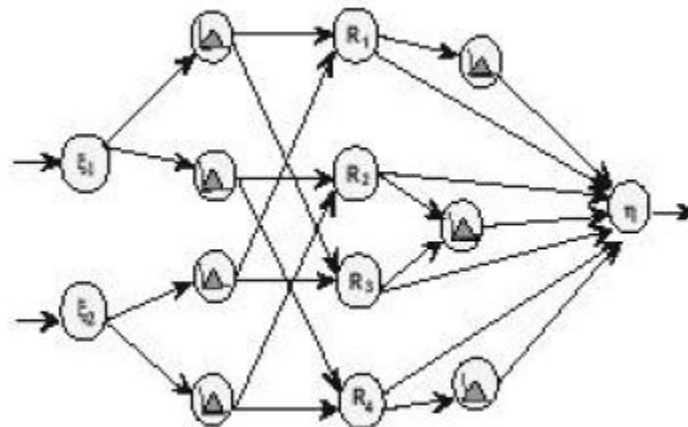


Figure 5.5: GARIC Architecture

C) NEFCON Architecture:

The **Neural Fuzzy Controller** NEFCON was initially designed to implement a Mamdani type inference fuzzy system as illustrated in figure 5.6. The connections in this architecture are weighted with fuzzy sets and rules using

the same antecedents (called shared weights), which are represented by the drawn ellipses. They assure the integrity of the base of rules. The input units assume the function of fuzzification interface. The logical interface is represented by the propagation function and the output unit is responsible for the defuzzification interface. The process of learning in architecture NEFCON is based on a fixture of reinforcement learning with back propagation algorithm. This architecture can be used to learn the rule base from the beginning, if there is no *à priori knowledge* of the system, or to optimize an initial manually defined rule base. NEFCON has two variants NEFPROX (for function approximation) and NEFCLASS (for classification tasks).

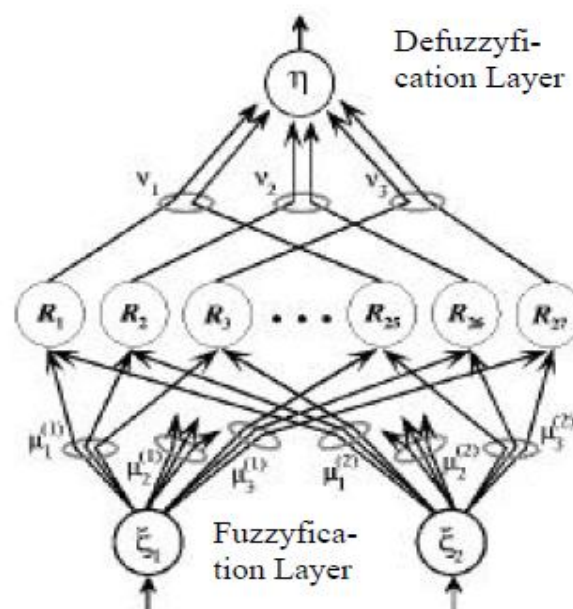


Figure 5.6: NEFCON architecture

D) EFuNN Architecture:

In **Evolving Neural Fuzzy Network** EFuNN all nodes are created during the learning phase. The first layer passes data to the second layer that calculates the degrees of compatibility in relation to the predefined membership functions. The third layer contains fuzzy rule nodes representing prototypes of input-output data as well as an association of hyper-spheres from the fuzzy input and fuzzy output spaces. Each rule node is defined by two sectors of

connection weights, which are adjusted through a hybrid learning technique. The fourth layer calculates the degree to which output membership functions are matched with the input data and the fifth layer carries out the defuzzification and calculates the numerical value for the output variable.

Dynamic Evolving Neural Fuzzy Network (dmEFuNN) is a modified version of the EFuNN with the difference that not only the winning rule node's activation is propagated but also a group of rule nodes that is selected for every new input vector and their activation values are used to calculate the dynamical parameters of the output function. While EFuNN implements Mamdani type fuzzy rules, dmEFuNN implements Takagi Sugeno fuzzy rules.

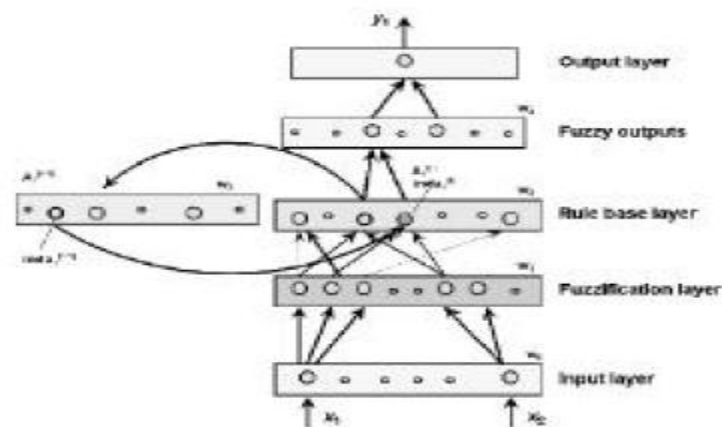


Figure 5.7: EfuNN architecture

5.3) ANFIS (Adaptive Neuro Fuzzy Inference System)

ANFIS (Adaptive Neuro-Fuzzy Inference System) is a neuro-fuzzy system developed by Roger Jang. It has a feed-forward neural network structure where each layer is a neuro-fuzzy system component (Figure 5.8). It simulates TSK (Takagi-Sugeno-Kang) fuzzy rule of type-3 where the consequent part of the rule is a linear combination of input variables and a constant.

The final output of the system is the weighted average of each rule's output. The form of the type-3 rule simulated in the system is as follows:

$$\begin{aligned} &\text{IF } x_1 \text{ is } A_1 \text{ AND } x_2 \text{ is } A_2 \text{ AND...AND } x_p \text{ is } A_p \\ &\text{THEN } y = c_0 + c_1x_1 + c_2x_2 + \dots + c_px_p \end{aligned}$$

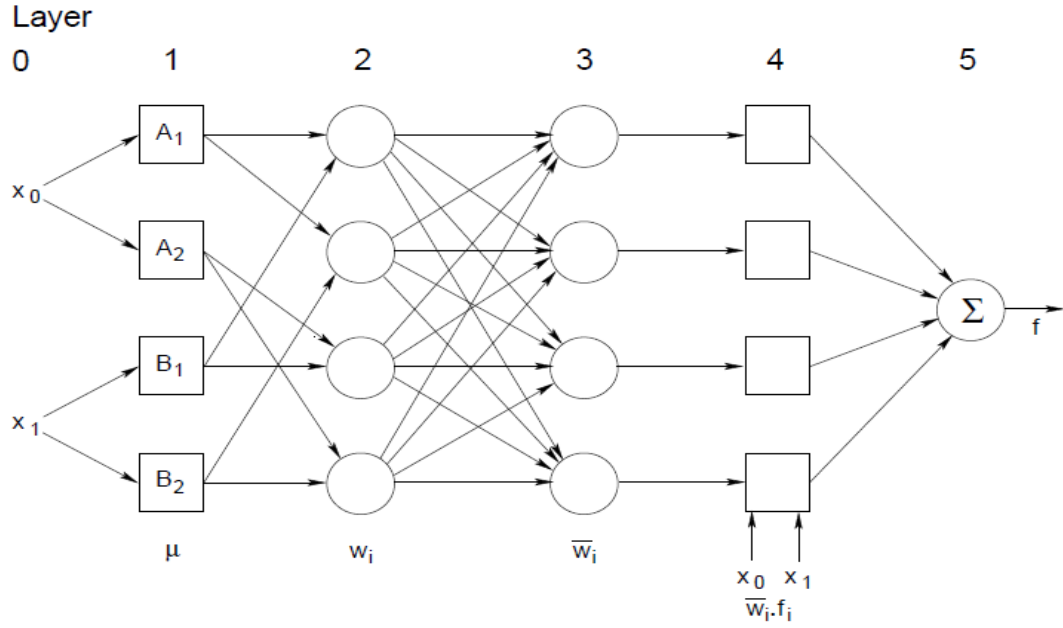


Figure 5.8: Basic ANFIS structure

The neural network structure contains 5 layers excluding input layer:

- 1) Layer 0 is the input layer. It has n nodes where n is the number of inputs to the system.
- 2) Layer 1 is the fuzzification layer in which each node represents a membership value to a linguistic term as a Gaussian function with the mean

$$\mu_{A_i}(x) = \frac{1}{1 + \left[\left(\frac{x - c_i}{a_i} \right)^2 \right]^{b_i}}$$

Where a_i , b_i , c_i are parameters of the function. These are adaptive parameters. Their values are adapted by means of the back-propagation algorithm during the learning stage. As the values of the parameters change, the membership function of the linguistic term A_i changes. These parameters are called premise parameters.

In that layer there exists $n * p$ nodes where n is the number of input variables and p is the number of membership functions. For example, if size is an input variable and there exist two linguistic values for size which are SMALL and LARGE then two nodes are kept in the first layer and they denote the

membership values of input variable size to the linguistic values SMALL and LARGE.

3) Each node in Layer 2 provides the strength of the rule by means of multiplication operator. It performs AND operation:

$$W_i = \mu_{A_i}(x_0) * \mu_{B_i}(x_1)$$

Every node in this layer computes the multiplication of the input values and gives the product as the output as in the above equation. The membership values represented by $\mu_{A_i}(x_0)$ and $\mu_{B_i}(x_1)$ are multiplied in order to find the firing strength of a rule where the variable x_0 has linguistic value A_i and x_1 has linguistic value B_i in the antecedent part of Rule i .

There are p^n nodes denoting the number of rules in Layer 2. Each node represents the antecedent part of the rule. If there are two variables in the system namely x_1 and x_2 that can take two fuzzy linguistic values SMALL and LARGE, there exist four rules in the system whose antecedent parts are as follows:

IF x_1 is SMALL AND x_2 is SMALL
IF x_1 is SMALL AND x_2 is LARGE
IF x_1 is LARGE AND x_2 is SMALL
IF x_1 is LARGE AND x_2 is LARGE

4) Layer 3 is the normalization layer which normalizes the strength of all rules according to the equation:

$$\bar{w}_i = \frac{w_i}{\sum_{j=1}^R w_j}$$

where w_i is the firing strength of the i th rule which is computed in Layer 2. Node i computes the ratio of the i th rule's firing strength to the sum of all rules' firing strengths. There are p^n nodes in this layer.

5) Layer 4 is a layer of adaptive nodes. Every node in this layer computes a linear function where the function coefficients are adapted by using the error function of the multi-layer feed forward neural network:

$$\bar{w}_i f_i = \bar{w}_i (p_0 x_0 + p_1 x_1 + p_2)$$

p_i 's are the parameters where $i = n + 1$ and n is the number of inputs to the system (i.e. number of nodes in Layer 0). In this example, since there exist two variables (x_1 and x_2), there are three parameters (p_0 , p_1 and p_2) in Layer 4. w_i is the output of Layer 3. The parameters are updated by a learning step. Least squares approximation is used in ANFIS. In the temporal model, back-propagation algorithm is used for training.

6) Layer 5 is the output layer whose function is the summation of the net outputs of the nodes in Layer 4. The output is computed as:

$$\sum_i \bar{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i}$$

where $w_i f_i$ is the output of node i in Layer 4. It denotes the consequent part of rule i . The overall output of the neuro-fuzzy system is the summation of the rule consequents.

ANFIS uses a hybrid learning algorithm in order to train the network. For the parameters in the layer 1, back-propagation algorithm is used. For training the parameters in the Layer 4, a variation of least squares approximation is used.

The subsequent to the development of ANFIS approach, a number of methods have been proposed for learning rules and for obtaining an optimal set of rules. For example, Mascioli et al. have proposed to merge Min-Max and ANFIS model to obtain neuro-fuzzy network and determine optimal set of fuzzy rules. Jang and Mizutani have presented application of Lavenberg-Marquardt method, which is essentially a nonlinear least-squares technique, for learning in ANFIS network. In another paper, Jang has presented a scheme for input selection and used Kohonen's map to training.

Jang has introduced four methods to update the parameters of ANFIS structure, as listed below according to their computation complexities:

1. Gradient decent only: all parameters are updated by the gradient descent.
2. Gradient decent only and one pass of LSE: the LSE is applied only once at the very beginning to get the initial values of the consequent parameters and then the gradient decent takes over to update all parameters.
3. Gradient decent only and LSE: this is the hybrid learning.
4. Sequential LSE: using extended Kalman filter to update all parameters.

The strength of neuro-fuzzy systems involves two contradictory requirements in fuzzy modeling: interpretability versus accuracy. In practice, one of the two properties prevails. The neuro-fuzzy in fuzzy modeling research field is divided into two areas: linguistic fuzzy modeling that is focused on interpretability, mainly the Mamdani model; and precise fuzzy modeling that is focused on accuracy, mainly the Takagi-Sugeno-Kang (TSK) model.

5.4. Trainable neuro-fuzzy systems and training procedures

Two major approaches of trainable neurofuzzy models can be distinguished. The network based Takagi-Sugeno fuzzy inference system and the locally linear neurofuzzy model. The locally linear model is equivalent to Takagi-Sugeno fuzzy inference system under certain conditions, and can be interpreted as an extension of normalized RBF network as well. Therefore, the mathematical description of Takagi Sugeno neurofuzzy model which is the most general formulation will be described in this section.

The Takagi-Sugeno fuzzy inference system is constructed by fuzzy rules of the following type:

$$\begin{aligned} \text{Rule}_i : & \text{ If } u_1 = A_{i1} \text{ And } \dots \text{ And } u_p = A_{ip} \\ & \text{ then } \hat{y} = f_i(u_1, u_2, \dots, u_p) \end{aligned} \quad (1)$$

Where $i=1 \dots M$ and M is the number of fuzzy rules. u_1, \dots, u_p are the inputs of network, each A_{ij} denotes the fuzzy set for input u_j in rule i and $f_i(\cdot)$ is a crisp function which is defined as a linear combination of inputs in most applications

$$\hat{y} = \omega_{i0} + \omega_{i1}u_1 + \omega_{i2}u_2 + \dots + \omega_{ip}u_p \quad (2)$$

$$\text{Matrix form} \quad \hat{y} = a^T(\underline{u}) \cdot W$$

Thus the output of this model can be calculated by

$$\hat{y} = \frac{\sum_{i=1}^M f_i(\underline{u}) \mu_i(\underline{u})}{\sum_{i=1}^M \mu_i(\underline{u})} ; \quad \mu_i(\underline{u}) = \prod_{j=1}^p \mu_{ij}(u_j) \quad (3)$$

Where $\mu_{ij}(u_j)$ the membership function of j th is input in the i th rule and $\mu_i(\underline{u})$ is the degree of validity of the i th rule. This system can be formulated in the basis function realization which clarifies the relation between Takagi-Sugeno fuzzy inference system and the normalized RBF network. The basis function will be

$$\phi_i(\underline{u}) = \frac{\mu_i(\underline{u})}{\sum_{j=1}^M \mu_j(\underline{u})} \quad (4)$$

and as a result,

$$\sum_{j=1}^M \phi_j(\underline{u}) = 1 \quad (5)$$

This neurofuzzy model has two sets of adjustable parameters; first the antecedent parameters, which belong to the input membership functions such as centers and deviations of Gaussians; second the rule consequent parameters such as the linear weights of output in equation (2). It is more common to optimize only the rule consequent parameters. This can be simply done by linear techniques like least squares. A linguistic interpretation to determine the antecedent parameters is usually adequate. However, one can opt to use a more powerful nonlinear method to optimize all parameters together. Gradient based learning algorithms can be used in the optimization

of consequent linear parameters. Supervised learning is aimed to minimize the following loss function (mean square error of estimation):

$$J = \frac{1}{N} \sum_{i=1}^N (y(i) - \hat{y}(i))^2 \quad (6)$$

where N is the number of data samples.

According to the matrix form of (2) this loss function can be expanded in the quadratic form

$$J = W^T R W - 2W^T P + Y^T Y / N \quad (7)$$

Where $R = (1/N) A^T A$ is the autocorrelation matrix, A is the $N \times p$ solution matrix whose i th row is $a(\underline{u}(i))$ and $P = (1/N) A^T y$ is the p dimensional cross correlation vector. From

$$\frac{\partial J}{\partial W} = 2RW - 2P = 0 \quad (8)$$

the following linear equations are obtained to minimize J :

$$RW = P \quad (9)$$

and W is simply defined by pseudo inverse calculation. One of the simplest local nonlinear optimization techniques is the steepest descent. In this method the direction of changes in parameters will be opposite to the gradient of cost function

$$\Delta W(i) = -\frac{\partial J}{\partial W(i)} = 2P - 2RW(i) \quad (10)$$

and

$$W(i+1) = W(i) + \eta \cdot \Delta W(i) \quad (11)$$

where η is the learning rate.

Other nonlinear local optimization techniques can be used for this purpose, e.g. the conjugate gradient or Levenberg-Marquardt which are faster than steepest descent. All these methods have the possibility of getting stuck at local minima. Some of the advanced learning algorithms, that have been proposed for the optimization of parameters in Takagi-Sugeno fuzzy inference

system, include ASMOD (Adaptive B-Spline modeling of observation data), ANFIS (Adaptive network based fuzzy inference system and FUREGA (fuzzy rule extraction by genetic algorithm). ANFIS is one of the most popular algorithms that has been used for different purposes, such as system identification, control, prediction and signal processing. ASMOD is an additive constructive algorithm based on k -d tree partitioning. It reduces the problems of derivative computation, because of the favorable properties of B-spline basis functions. Although ASMOD has a complicated procedure, it has advantages like high generalization and accurate estimation.

One of the most important problems in learning is the prevention of over fitting. It can be done by observing the error index of test data at learning iterations. The learning algorithm will be terminated, when the error index of test data starts to increase, in an average sense. Prevention of over fitting is the most common way of providing high generalization.

5.4.1. Emotional Learning

Satisfying approaches to decision making have, in recent years, been widely adopted for dealing with complex engineering problems. New learning algorithms like reinforcement learning, Q-learning, and the method of temporal differences are characterized by their fast computation and in some cases lower error in comparison with classical learning methods. They can be interpreted as approximations to dynamic programming, which although furnishes a well known computational algorithm, via recursive solution of the Bellman-Jacobean-Hamilton equation and perhaps the best example of fully rational approach to decision making, is notorious for its computational complexity, sometimes referred to as the “curse of dimensionality”. Fast training is a notable consideration in control applications. Prediction applications also belong to the class of decision making problems where two desired characteristics are accuracy and low computational complexity.

The Emotional learning method is a psychologically motivated algorithm which is developed to reduce the complexity of computations in prediction problems with particular goals. In this method the reinforcement signal is replaced by an emotional cue, which can be interpreted as a

cognitive assessment of the present state in light of goals and intentions. The main reason of using emotion in a prediction problem is to lower the prediction error in some regions or according to some features. For example predicting the sunspot number is more important in the peak points of the eleven-year cycle of solar activity, or accurate prediction of the peaks and valleys in the price of securities may be desired. This method is based on an emotional signal which shows the emotions of a critic about the overall performance of prediction. The emotional signal can be produced by any combination of objectives or goals which improve estimation or prediction. The loss function will be defined just as a function of emotional signal and the training algorithm will be simply designed to decrease this loss function. So the predictor will be trained to provide the desired performance in a holistic manner. If the critic emphasizes on some regions or some properties, this can be observed in his emotions and simply affects the characteristics of predictor. Thus the definition of emotional signal is absolutely problem dependent. It can be a function of error, rate of error change and many other features. Finding an appropriate formulation for emotion is not usually possible; in contrast a linguistic fuzzy definition of it is absolutely intuitive and plausible.

A loss function is defined on the base of emotional signal. A simple form is

$$J = \frac{1}{2} K \sum_{i=1}^N es(i)^2 \quad (12)$$

where $es(i)$ is the of emotional signal to the i th sample of training data, and K is a weighting matrix, which can be simply replaced by unity.

Learning is adjusting the weights of model by means of a nonlinear optimization method, e.g. the steepest descent or conjugate gradient. With steepest descent, the weights are adjusted by the following variations:

$$\Delta\omega = -\eta \frac{\partial J}{\partial \omega} \quad (13)$$

where η is the learning rate of the corresponding neurofuzzy controller and the right hand side can be calculated by chain rule:

$$\frac{\partial J}{\partial \omega} = \frac{\partial J}{\partial es} \cdot \frac{\partial es}{\partial y} \cdot \frac{\partial y}{\partial \omega} \quad (14)$$

$$\text{According to (12): } \frac{\partial J}{\partial es} = K \cdot es$$

and $\frac{\partial y}{\partial \omega}$ is accessible from (3) where $f_i(.)$ is a linear function of weights.

Calculating the remaining part, $\frac{\partial es}{\partial y}$, is not straightforward in most cases. This is the price to be paid for the freedom to choose any desired emotional cue as well as not having to impose presuppose any predefined model. However, it can be approximated via simplifying assumptions. If, for example error is defined by

$$e = y_r - y \quad (15)$$

where y_r is the output to be estimated, then

$$\frac{\partial es}{\partial y} = -\frac{\partial es}{\partial e} \quad (16)$$

can be replaced by its sign (-1) in (14). The algorithm is after all, supposed to be satisfying rather than optimizing.

Finally the weights will be updated by the following formula:

$$\Delta \omega = -K \cdot \eta \cdot es \cdot \frac{\partial y}{\partial \omega} = -K \cdot \eta \cdot es \cdot \frac{\sum_{i=1}^M u_i \mu_i(\underline{u})}{\sum_{i=1}^M \mu_i(\underline{u})} \quad (17)$$

5.5. Prediction of the stock market and other securities using neuro fuzzy networks

As discussed in previous chapters, there have been various attempts to predict the stock market using the classic theories (fundamental analysis, technical analysis etc.), as well neural networks and fuzzy logic. The last few years, there has been significant progress in the field of stock market prediction, with the use of neuro-fuzzy networks (especially ANFIS). The

method used, will be described with a specific paradigm in this section. If there is a predictor that predicts the future exactly; then the best investment is on the maximum rate of return. For this reason, the performance of prediction is significant. Some researchers have used neural networks e.g. MLP and RBF for the prediction of securities. In this research, the emotional learning algorithm is applied to the network initially trained by ANFIS to predict the stock price of General Electric (GE) in S&P index 500. For this case one can use various definitions of emotional signal, as a function of prediction error and differential of error, or even any significant event like crossing of spot price with some well monitored moving average. Here the emotional signal is taken as the output of a linguistic fuzzy inference system with the error and the rate of error change as inputs. Five and three Gaussian membership functions are used for the inputs respectively. Figure 5.9 shows the surface generated by the fuzzy rules of emotional critic.

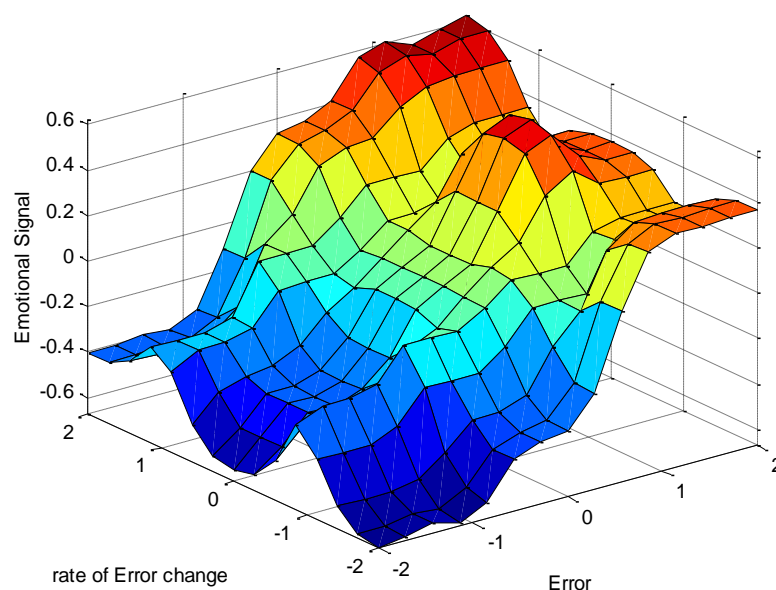


Figure 5.9: The output surface of a linguistic fuzzy inference system for producing emotional signal

In this research, the daily closed price for the stock is considered. The model parameters, number of regressors and number of neurons are optimized to prevent over fitting. The stock price of 800 days and the price of 400 following days are used for train data and test data respectively. The result of predicting

the stock price by ELFIS (Emotional learning fuzzy interference system) is presented in figure 5.10.

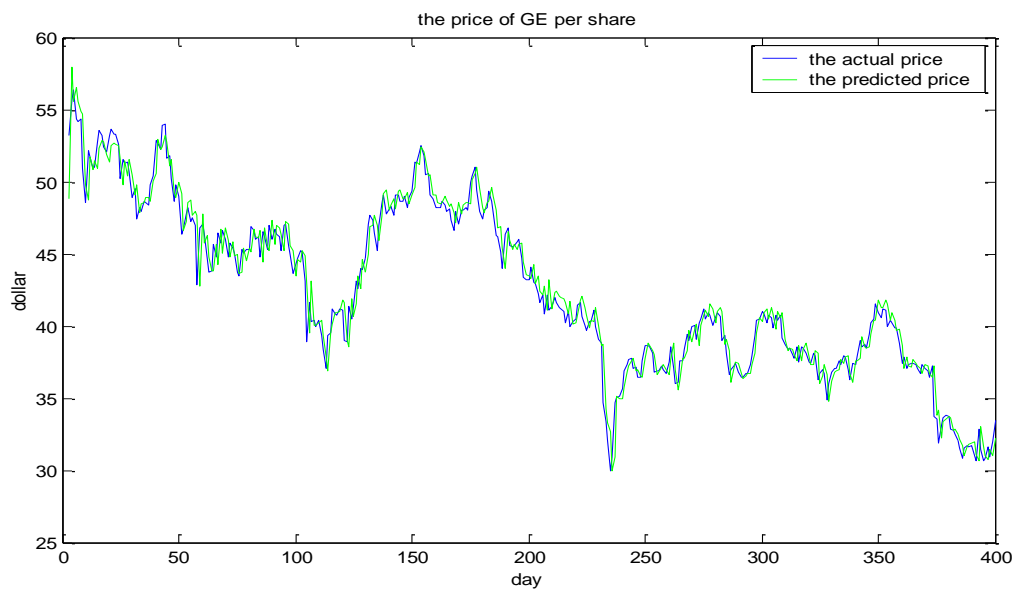


Figure 5.10: Predicting the security price using emotional learning plus ANFIS

Table 1 presents a comparison of the quality of predicting the daily closed stock price of General Electric (GE) by ELFIS with some other networks such as ANFIS, RBF, and MLP. These results are obtained by considering the over fitting and the optimal neurons in the hidden layer (on a 1.8 GHz Celeron processor). As this practical example shows, the emotional learning algorithm provides more accurate predictions with lower computational complexity.

	Specifications	Computation Time	NMSE
MLP	37 neurons in hidden layer	6.5600 sec.	0.0347
ANFIS	12 rules and 257 epochs	13.8390 sec.	0.0370
RBF	31 neurons in	7.2000 sec.	0.0395

	hidden layer			
ELFIS	12	Sugeno	1.8320 sec.	0.0358
	type	fuzzy		
	rules			

Table 1: A comparison of various neural and neurofuzzy models in the stock price prediction

Training a system to make decision in the presence of uncertainties is a difficult problem especially when computational resources are limited. Supervised training cannot be used because the desired values for the decision variables are unknown. However, the desirability of past decisions can usually be assessed after the outcomes of their implementations are observed. Therefore, unsupervised training methods that do not utilize those assessments cannot take full advantage of the available knowledge. Several approximate methods like back propagation through plant, and identification of the plant or (pseudo) inverse plant model have been successfully used in the past couple of decades. Behavioral and emotional approaches to control and decision making can also be classified in this category. Besides providing biological plausibility, they have the extra advantage of not being confined to cheap control problems like set point tracking. The emotional approach is a step higher in the cognitive ladder and can be more useful in goal-aware or context-aware applications (e.g. dealing with multiple objectives in decision problems even when the objectives are fuzzy or cannot be differentiated or directly evaluated with simple mathematical expressions).

Although prediction is easier to deal with because we do not have the further complexity of unknown plant, and so the proposed learning methods should also be compared to error minimization methodologies, model free prediction has also become of great importance in the past few decades, and there have been many past efforts to train neuro- and/or fuzzy predictors with alternative loss functions.

REFERENCES / BIBLIOGRAPHY

Forecasting stock market movement direction with support vector machine
Wei Huang, Yoshiteru Nakamura, Shou-Yang Wang; (2004)

Neuro-Fuzzy Systems: A Survey Jose Vieira, Fernando Mogado Dias,
Alexandre Mota (2000)

Web-based Fuzzy Neural Networks for Stock Prediction, Yu Tang, Fujun Xu,
Xuhui Wan and Yan-Qing Zhang (2002)

Union and Intersection of Triangular Fuzzy Sets Vincent, O. S. Olunloyo
(2010)

The parametric definition of membership functions in XML3, F. J. Moreno-
Velo, I. Baturone, S. Sánchez-Solano, A. Barriga (1998)

Predicting stocks average price, a neuro fuzzy approach, José Manuel
Andújar Márquez, Patricio Salmerón Revuelta, Omar Sánchez Pérez, and
Juan José de la Vega Jiménez. (2001)

Application of Artificial Neural Networks in Business Applications, *Nikhil
Bargava, student of Master of Technology, IIT Delhi, Manik Gupta, student of
Master of Technology, IIT Delhi (2002)*

Stock Price Forecast by Using Neuro-Fuzzy Inference System, Ebrahim
Abbasi, and Amir Abouec (2009)

An introduction to neural networks. Patrick van der Smagt, Ben krose (1996)

Consumer Choice Prediction: Artificial Neural Networks versus Logistic Models, Christopher Gan, Visit Limsombunchai, Mike Clemes and Amy Weng, Lincoln University, Canterbury, New Zealand (2005)

Using Neural Networks to Forecast Stock Market Prices, Ramon Lawrence Department of Computer Science (1997)

The Determinants of Foreign Listing Decision: Neural Networks versus Traditional Approaches, (2006) Piotr Staliński

The application of fuzzy logic to the construction of the ranking function of informal retrieval systems (2006) N.O Rubens (2006) N.O Rubens

Methods for the Construction of Membership Functions, (1998) A. Sancho Royo, J. L. Verdegay2.

Fuzzy Logic and Investment Strategy, (2009) Fatma Khcherem and Abdelfettah Bouri

A temporal neuro-fuzzy approach for times analysis (2006) Nuran Arzu Isman

Neural Networks and Fuzzy Logic *by Valluru B. Rao (1995)*

Fuzzy expert systems and fuzzy reasoning, James J. Buckley (2005)

WEBSITES

www.investopedia.com (some full-text documents available free)

www.statsoft.com (some free textbooks available)

