Anomalies Detection System for Stock Market

A dissertation submitted for the Degree of Master of Computer Science

U. A. U. Sumanaweera
University of Colombo School of Computing
2018

Declaration

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: U. A. U. Sumanaweera

Registration Number: 2015mcs074

Index Number: 15440748

Signature: Date:

This is to certify that this thesis is based on the work of Mr. U. A. U. Sumanaweera under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name: Mr. G. P. Senevirathne

Signature: Date:

Abstract

Stock market is the place to trade company stocks among market participants at an agreed price. Investors have to have a good knowledge about fluctuations of parameters of market and there is a possibility of novel investors get in trouble due to lack of awareness of fluctuations in market.

Rule based patterns are widely used in practice in the existing manipulation detection methods. However manipulators constantly change their strategies and they find new ways to manipulate markets. Therefore rule based or static detection methods fail to detect these new evolving manipulation attempts.

The main project objective is to research and implement a method to detect these evolving stock abusing patterns. Artificial immune system theories based on stock manipulation detection system is implemented as an advanced detection mechanism. This project approaches the problem by analyzing daily price,volume values of transactions along with the behavior of customer. Natural immune system techniques are used such as danger theory, negative selection, clonal selection and immune network theory which have approach to identify unknown signatures of anomalies in stock market transactions.

Novelty of this research is having a learning phase to train the system along with the usage of Artificial Immune System theories. Previous work does not have a learning phase based on AIS theories in detecting stock market anomalies. Unlike simple statical evaluations, system is capable of identifying stock market anomalies in a better rate due to supervised learning techniques.

System was tested based on real transaction data collected from Saudi Stock Exchange. First stage is supervised learning for price, volume anomaly detection and second stage is optimize results using customer behavior. More than 30,000 real transactions are used for testing in various models. Degree of anomaly of a transactions is marked based on conclusions of three domain experts and system output is evaluated based on them. Best System output was 96% of Precision, 100% of recall, 75% of Accuracy and 88% of F1 score. All the percentages are calculated with respect to conclusions of domain experts.

Contents

List of Figures

List of Tables

Chapter 1: Introduction

Anomaly detection systems have become more and more interesting with time but most of the anomaly detection systems such as anti-virus systems are programmed in order to detect known signatures of anomalies. Researchers are approaching on new techniques to identify new attacks along with known attacks, which cannot be detected by programmed detection systems. This scenario has opened a new research are to discover optimized techniques to detect anomalies which are evolving.

A stock market is the place to trade company stocks among market participants at an agreed price. This is a place where various kinds of anomalies take place which leads to unexpected results such as new comers lose their money, powerful people get more profits, market indices are not denoting real picture of market and stakeholders and even companies fell down as a whole and etc. In order to maintain stock market as a fair and safe place to all parties, evolving anomalies should be immediately identified and reported.

When it comes to anomaly detection methods, there are detection techniques such as data mining methods, data profiling methods, pattern recognition methods and statically implemented methods. New researches have come up to analyze the possibility of applying artificial immune system techniques and model to resolve problems of anomaly detection. Natural immune system is the best known natural anomaly detection system which consists of various techniques to protect the body from foreign invaders like viruses and bacteria, and having capabilities like adaptability, autonomous, accuracy and identifying attacks and act upon them. Above qualities have lead researches to apply this model to solve this problem.

Even though there are stock market anomaly detection systems in previous work, there was a gap of using machine learning techniques along with Artificial Immune System theories. This was important to identify advanced fraud transactions which cannot be detected using simple statical methods. Various models have been evaluated with real transaction data and came up with a better solution in this research to fulfill that gap.

1.1 The Problem

A stock market is the place to trade company stocks for market stakeholders upon agreement. Investors have to have a good knowledge about fluctuations of parameters of market and there is a possibility of novel investors get in trouble due to lack of awareness of fluctuations in Market.

People normally tend to react on sudden fluctuations expecting huge profits. But these movements are created artificially sometimes to cheat people and get their money. This stock manipulation is one of the major problems in stock markets which cause the stock market to lose its credibility. Some countries are having laws against illegal transactions of stock market.

Rule based patterns are widely used in practice in the existing manipulation detection methods. However manipulators constantly change their strategies and they find new ways to manipulate markets. Therefore rule based or static detection methods fail to detect these new evolving manipulation attempts. So the problem of detecting anomalies in stock markets remains open.

1.2 Motivation

Anomaly detection in any particular domain is a very challenging task since anomalies are evolving and the detection systems also should compete to identify and defeat them. Identifying characteristics of anomalies and differentiating it from normal behavior is not a straight-forward

task. Applying natural immune system model is also interesting because they address most difficult problems using simple techniques. Stock Market is getting more and more popular in business world and countries economy also depends on stock market behavior and vice versa. Therefore it is very important to have an advance security system to make it a safe and fair system.

1.3 Objective of the Project

Since rule based solutions are not accurate with evolving stock manipulation strategies, these manipulations have to be detected and prevented in an advanced way. The main project objective is to research and implement a method to detect these evolving stock abusing patterns.

Main Objective is divided into two sub objectives.

- Implementation of price, volume fluctuation detector which is capable of finding abnormalities of a data stream and calculating degree of abnormality of a suspected fluctuation.
- Optimization for above system to increase precision and recall values when detecting suspected scenarios.

Even though price, volume fluctuation detector points out several danger signals, some of them are not due to market anomalies. Objective is, In order to increase precision and recall values when detecting, secondary technique will be implemented based on characteristics of involving stakeholders. This methodology will be capable of identifying some of the stock market anomalies to a greater extend. Ex: Insider Trading, Front Running

Technologies and techniques: Danger Theory, Artificial Immune Systems, Negative Selection, Machine Learning, Natural Algorithms, Clonal Selection.

1.4 Research Contributions

Research contributions have analyzed in artificial immune system domain, stock market anomalies domain and target for applications of artificial immune techniques detecting stock market anomalies. Few of the researches conducted in above domains are listed down below. Research contributions and related research work are explained in detail in next chapter.

1.4.1 The Danger Theory and Its Application to Artificial Immune Systems

New idea challenging the classical self-non-self-viewpoint has become popular amongst immunologists. It is called the Danger Theory. They approach this theory from the perspective of Artificial Immune System techniques. A summary of the Danger Theory is analyzed with particular emphasis on analogies in the Artificial Immune Systems world. A number of potential application areas are then used to provide a framing for a critical assessment of the concept, and its relevance for Artificial Immune Systems.[1]

1.4.2 Abnormal Pattern Detection in Time Series Data via Artificial Immune System Model

This project has been conducted with the target of detecting abnormal transactions performed in stock exchange. The project has executed its methodology using price/volume data stream and then classifying using statical calculations. Implemented solution tested by techniques used in natural immune system such as danger theory, negative selection, clonal selection and immune network theory which ultimately guide to identify unknown signatures of anomalies in daily transactions. [8].

1.4.3 Stock Option Returns and Stock Anomalies: Cross Market Efficiency and the Cost of Hedging Value vs Growth Firms Stock Returns

The empirical literature on stock returns shows overwhelming evidence of stock anomalies related to value investing. This paper studies the relative performance of stock options of value and growth stocks. [10]

1.4.4 Stock market anomalies: A re-assessment based on the UK evidence

This paper reports evidence documenting the presence of a number of irregularities in stock price behavior of firms on the London Stock Exchange. The size effect is not only not the sole anomaly but is not even the most dominant one. Specifically, investment strategies based on dividend yield, PE ratios and share prices appear as profitable, if not more, as a strategy concentrating on firm size. Although there is a large degree of interdependency between all four effects, it is still apparent that the dividend yield and PE ratios subsume the size and share price effects. [6]

1.5 Scope of Project

The scope of the project is limited to detect anomalies of stock market transactions which are reflected through abnormal fluctuations of price and volume. Results will be improved using behavior of customers related to those abnormal transactions detected. This anomaly detection system consists of price and volume fluctuation detector and behavior detector to measure degree of abnormality based on customer behavior. Following anomalies will be mainly targeted by mentioned system.

- Insider Trading – Even though all parties must have same information about company's

financial profiles, company managers/directors (insiders) get to know about company financial profile in advance. They act based on these and get huge profits with minimum risk.

- Front running – When some individuals going to buy huge amount of stocks, People who know about that in advance will buy stocks at current price and create artificial price increment and sell those stocks later.

- Painting the tape/ Wash Sales – Stock transactions are performed within known parties/ individual and build a trust of that company stocks among community and sell those initial stocks at a higher price.

- Pump and Dump/ Poop and Scoop – Group of people attempt to push up/down the price by spreading rumors.

Target market is TDWL (Saudi Stock Exchange). Exchange feed will be modeled to real customer behavior in order to analyze individual behavior since we have source of information of real transactions flown through DirectFN (DirectFN Technologies Pvt Ltd.) Order Management Systems. Feed data contains price and volume data of companies for transactions taken place which requires for evaluation.

Chapter 2: Background Study

Background concepts and related research work are explained in this chapter. There are researches related to artificial immune systems and its applications which we use as fraud detection system, and researches related to fraud detection of stock market anomalies. Therefore this chapter consists of three sections such as,

1. Natural Immune System and techniques.
2. Artificial Immune systems and its applications
3. Anomaly detection of stock market.

2.1 Natural Immune System and techniques

2.1.1 Natural Immune System

There are many systems which are capable of executing amazing functionalities in order to keep human body in normal state. They are complex in structure but using simple techniques to fulfill tasks. Immune system is also very important to humans since it is responsible for protecting human body from virus and bacteria and etc.

Immune system use negative selection technique to detect abnormal patterns. After detecting, immune response is mounted on foreign invaders. This process is called Clonal selection. Im-

mune system maintains valuable detector repository by eliminating redundant detectors. Apart from this, dangerous alerts are identified using Danger theory.

2.1.2 Negative selection

The main objective of this process is to produce detectors which are capable of identifying foreign invaders.

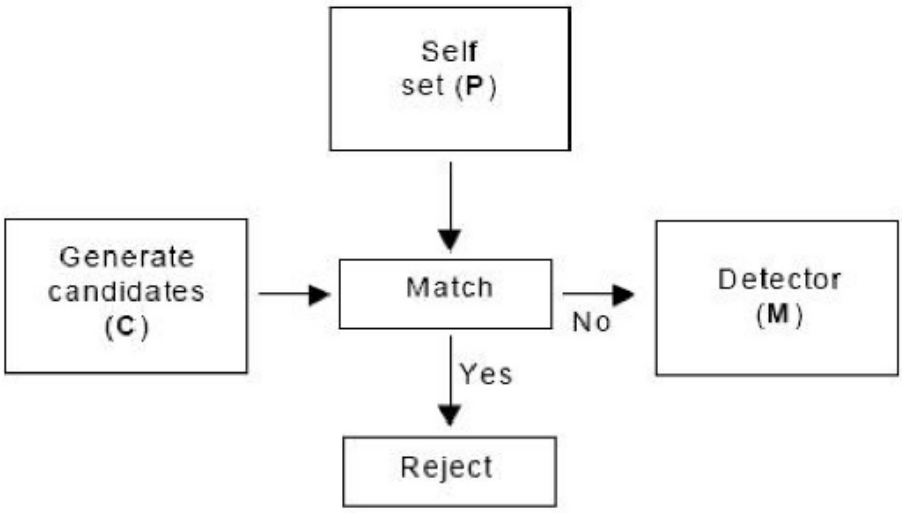

Figure 2.1: Negative Selection (Castro & Timmis 2002)

As described in above figure, immune system generates possible detectors randomly and then those detectors are sent through a mutation process. Generated candidates (detectors) are matched with sample self-cells and destroyed if matched. Likewise detectors are generated which might be matched with non-self-cells [2]

2.1.3 Clonal Selection

In Clonal selection process, when a detector identifies an antigen, it is subjected to proliferate process which is diversified detectors generated which are more capable of capturing same

antigen next time. Detectors which are closely related to antigen will eliminate it and finally it will be stored [2].

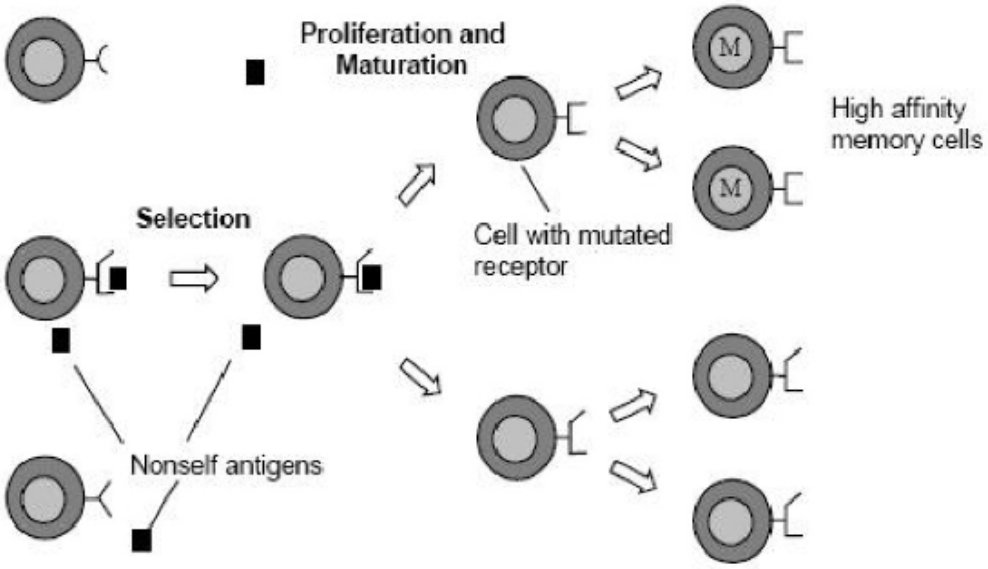

Figure 2.2: Clonal Selection (Castro & Timmis 2002)

2.1.4 Danger Theory

This is introduced to overcome some draw backs of negative selection process. Danger theory is well ahead of negative selection in scalability, fault rate and evolution of detectors.

In this process, danger signal is sent as a confirmation of detecting dangerous cell, therefore all the non-self-cells will not be considered as foreign invaders. [1]

2.2 Artificial Immune System Applications

Artificial immune system is a model of natural immune system and it uses natural immune system techniques to provide solutions in computational world. Some of the applications of artificial immune systems and how it solves given problem will be described below.

2.2.1 Anomaly detection of Time Series Data

Negative selection algorithm is applied to generate detectors from known data set and use them in detecting novelties of time series data [4].

Proposed solution is a pre-defined process of analyzing incoming data set and identifying anomalies. First, data set is divided into separate chunks using sliding windows. Then encode each chunk and store values as self-set. Random strings of values are generated and match them with stored self-values. Finally store values which did not match with self-values.

Fault detection of milling machines is an application of above process. It is very important to have real time monitoring of tool conditions in automated machine operation condition in milling industry. A reliable and effective tool breakage technique is required to provide instant response to unexpected tool failure, in order to prevent damages to work piece and machine tool. Behavior of cutting force is monitored and report possible failures in proposed solution .[4] This solution can be applied to detect variation of data set which has consistent and periodical behavior.

This is not a novel research area, many researches have introduced various methods to identify anomalies in time series data. Paper [3] has introduced statically executed method which is having three parts called data clustering, rule generation and anomaly detection. Clustering is done by using Gecko algorithm. According to that data is clustered in to maximum number of chunks and merge them into specific number of chucks again. “L” method is used to identify that specific number. After generation of clusters, specific rule set is generated to go through identified data points. This solution expects data stream to be compatible with generated rule set. Therefore this solution is not suitable with a data stream which having highly dynamic behavior.

Below Diagram shows how the system works [3].

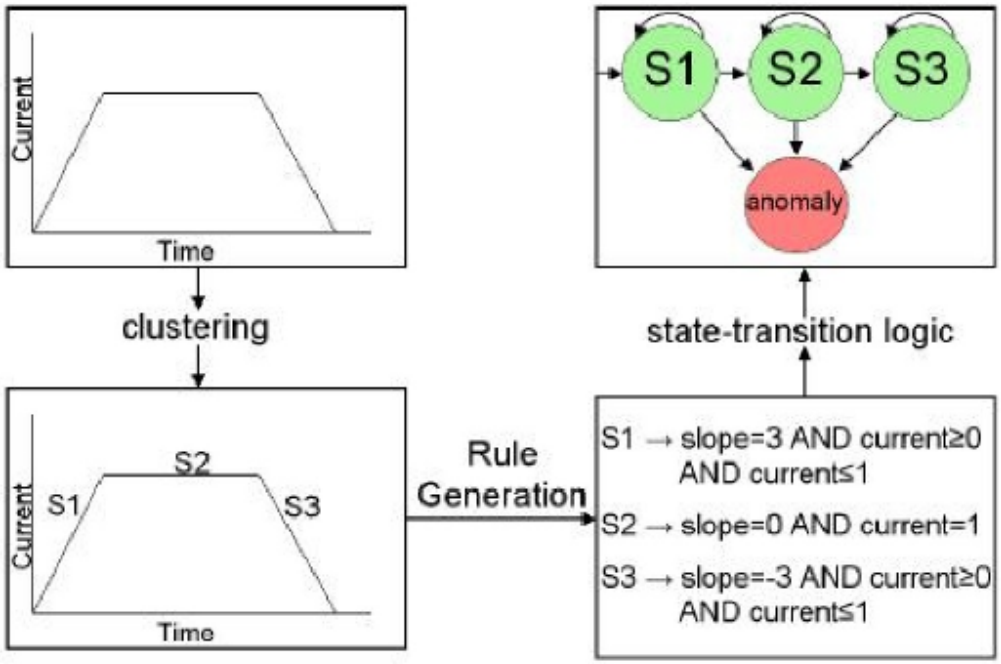

Figure 2.3: Anomaly Detection of Time Series Data (Salvador & Chan 2005)

2.2.2 Anomaly detector for financial fraud in retail sector

Fraud detection of retail sector is stated in reference [5]. With the major involvement of technology for retail sector, ecommerce has been a key component. Electronic money transactions lead to some security problems, even if there are many advantages of it. Some payment methods are completely based on electronic transactions so that, various kind of techniques are used to cheat in those situations since buyers and sellers do not see each other. Apart from that, fake transactions or split transactions are entered to system so that they are highly paid since payments depend on number of transactions too.

In [5], they suggests to detect those anomalies by monitoring transaction patterns and identifying abnormal transaction patterns. A-priori algorithm and stored rules which are in the form of IF THEN are used to perform anomaly detection.

Important part of that paper is using positive selection algorithm. They point out when using positive selection, system can handle large amount of data reducing workload of negative selection.

2.2.3 Anomaly detection of stock markets

Reference [7] categorizes stock market anomalies into two categories.

- Anomalies in Price – Abnormal fluctuations of price data
- Anomalies in individual behavior – Several people is involved in single transactions and some of the individuals behave different than their peers. This is categorized as an anomaly

Anomaly detection solution only focused on detecting anomalies by analyzing abnormal behavior of stock brokers. This is a significant importance because most of the manipulations are executed by stock brokers. Proposed solution is named as Peer Group Analysis (PGA) which initially identifies stock brokers that are having same set of characteristics. Then statically analyzed that identifies group. Finally individuals act differently than others are identified.

Main workflow discussed is illustrated in figure 2.4

This solution is tested in Dhaka Stock Market and shows significant results. But there are number of drawbacks in this solution as listed below.

- Individuals behave normally, but different than peers are identified as abnormal scenario because PGA is not capable of identifying them as normal since it depends only on peer behavior.
- This only considers stock brokers behavior even there are number of individuals involve in single transaction.
- Sudden increment and decrement of price and volume is a significant data when detecting anomalies of stock markets which is not taken into account in this solution.

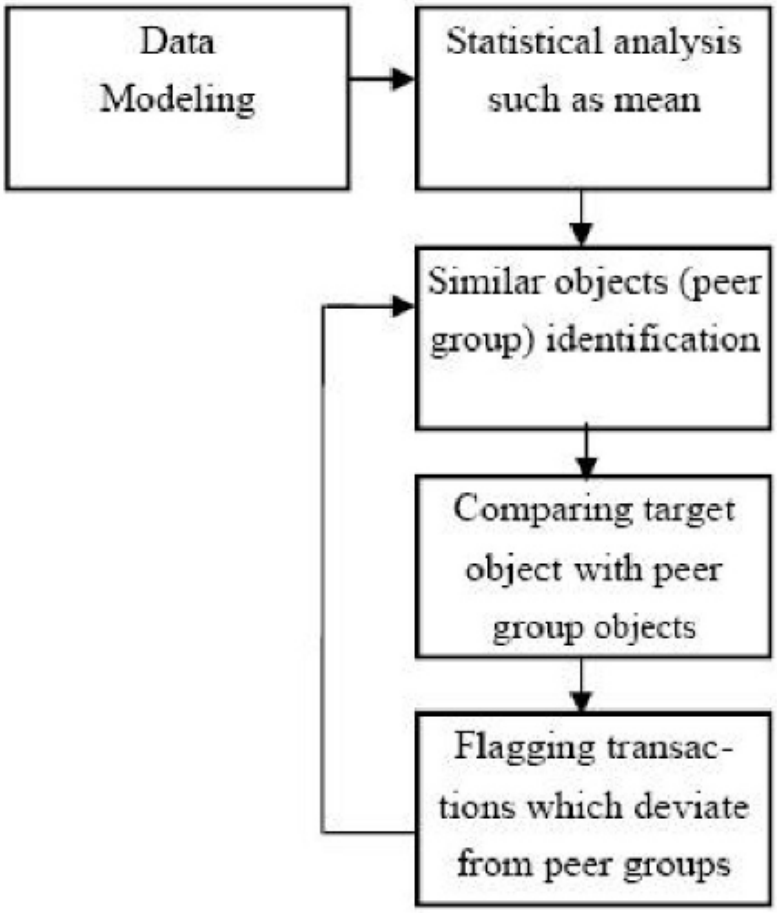

Figure 2.4: PGA Based Anomaly Detection (Ferdousy & Maeda 2006)

Reference [8] has improved detection methodology by referring the daily price and volume of transaction as well as the behavior of individuals. Techniques used in natural immune system such as danger theory, negative selection, clonal selection and immune network theory are used in implementing the solution which has targeted to identify unknown signatures of anomalies in stock market transactions. System has been tested based on data collected from Colombo Stock Exchange and the results were examined by a domain expert. But detection techniques of this proposal is fully dependent on statical calculations.

Following are limitations of above solution proposed.

- Fully dependent on statical calculations, and no machine learning techniques involved.
- It doesn't have a separate learning phrase even if it is capable of identifying the abnormalities by examining features of the given data stream rather than globally assigning boundary values to differentiate normal and abnormal ranges
- All abnormal price fluctuations are captured gave priority to them rather than volume changes, which lead significant false positive error rate of results.
- Not capable of identifying abnormal movements of individuals when their figures do not exceed 100000. That means split anomalies can easily bypass the given solution.

Future work of the solution is illustrated as concentrating of representing identified anomalies in more flexible manner which enhances system decisions and matching process. This solution can be further optimized with another parameter to reduce false positive and false negative errors.

Chapter 3: Problem Analysis

Stock market anomaly detection problem will be analyzed in this chapter. What is known as stock market anomalies, how and who can it be performed, how to identify them and the characteristic features of stock market anomalies will be discussed.

A stock exchange is a place where traders, stock brokers and customers can buy and sell some amount of stock, bonds, and other securities depending on market behavior. Other than stock issued by companies which are listed in exchange, unit trusts, investment products, derivatives and bonds are also traded. Stock exchanges act as auctions which trading happens often, but markets with sellers and buyers consummating transactions at a common ground.

And it is a great income source for investors where they can hold some percentage of ownership of several companies and get profits when company earns money (as dividends) and when the stock price goes high. And also company owners can get their company listed and sell a particular percentage of company and invest that money to expand the company. Majority of companies of all countries are listed in a stock exchange and therefore market indices vary with countries economy level. Stock market fluctuations indicate countries economy level.

Investors have to have a good knowledge about fluctuations of parameters of market and there is a possibility of novel investors get in trouble due to lack of awareness of fluctuations in Market. People normally tend to react on sudden fluctuations expecting huge profits. But these movements are created artificially sometimes to cheat people and get their money. This stock manipulation is one of the major problems in stock markets which cause the stock market to lose its credibility. Some countries are having laws against illegal transactions of stock market.

3.1 Stock Manipulations

Stock manipulation is performing transactions different than peers, with intention of cheating others and get more profit. Some of those techniques are not illegal but that will cheat other stake holders.

Figure 3.1 shows price graph of a company which has a suspicious behavior (marked in circles).

Figure 3.1: Stock Value Variation - 4300

Laws for stock manipulations are different from market to market and country to country. Some of the markets are having surveillance systems which identify stock manipulations real time and report them. There are limitations of surveillance systems because they report well known stock manipulations only.

Below is some of the well-known stock manipulations

1. Front running – When some individuals going to buy huge amount of stocks, People who know about that in advance will buy stocks at current price and create artificial price increment and sell those stocks later. Most of the time stock brokers are involved in these type of stock manipulations.

   As an example when a customer is willing to buy huge amount of stocks, broker buys those in advance, therefore stock price increases due to increment of demand and then sell those stocks of broker to customer to higher price.

2. Insider Trading – Even though all parties must have same information about company's financial profiles, company managers/directors (insiders) get to know about company financial profile in advance. That is a violation of a law in stock market which states that all parties should have same information about company's financial profiles. Insiders act based on advance information and get huge profits with minimum risk.

   As an example, if the company is about to get a huge profit, insiders buy a lot of stock early and sell them after increment of company stock price.

3. Pump and Dump – this is about group of people attempt to push up the price by spreading rumors within the community. They have a front seat when considering particular company and their comments have a high validity. They spread those misleading statements in online collaborative places and after increment of the stock price, they sell their stocks.

4. Poop and Scoop – this activity is performed by group of people spreading rumors which is bad for companies reputation and after the stock price is decreased they buy stocks at a low price. This is opposite manipulation action of Pump and Dump.

5. Bid Support – this is a form of market manipulation, which happens due to multiple bids are placed for small amounts of share but the value is just below the highest bid. This causes of attracting sell orders and creating a false multiple transactions for particular stock, while giving the impression that plenty of buyers are waiting to buy.

6. Painting the Tape – Stock transactions are performed within known parties and build a trust of that company stocks among community and sell those initial stocks at a higher price. This manipulation is done by group of buyers. They perform huge amount of transactions among themselves and then external investors starts to trade with that company. Then that group of manipulators can sell their stocks to higher price.

7. Wash Sales – This activity is kind of similar to Painting the Tape manipulation type, the only difference is, not the group of people but and individual performs transactions with two or more accounts to create artificial demand on a particular symbol.

There are many other manipulation techniques which can be identified by common surveillance

systems. But there are few domain components (stock price, volume etc.) involves in those manipulations and in detail component identification is done in next section.

3.2 Identifying domain components

There are domain components which vary abnormally during stock manipulation. Main such domain parameters are price, volume of transaction and individual behavior. But in some of the manipulations, those parameters vary normally as well.

When considering the normal manipulation identifying procedure, there is a responsible party in most Stock Exchanges called Security Exchange Commission. They mainly observe fluctuation of price and volume of transactions and identify cases where those domain parameters vary abnormally enough. For such cases, they search for involved parties which could be broker, trader and customer who was involved in buying or selling. Then they analyze the behavior of above parties comparing with their peer groups.

Investigators of stock manipulations examine past behavior of involved parties who has suspected behavior. Then they confirm for suspects for possible fraud. But the important fact is one individual is hardly performs a manipulations, but multiple parties do. It is challenging to extract the connection of several parties involved in fraud behavior.

We will be discussing how to identify possible fraud initially using Price and Volume charts. As an example Figure 1 illustrates Price and volume charts of an Insider Trading case. As indicated in white arrows, price and volume values are increased exceptionally during the same time period. In sell customer's chart, they have sold large amount of volume at the same time. When analyzing sell customers chart for past transactions, they have purchased large amount of volume some time back as indicated in red arrow. This observation confirms a possible manipulation of insider trading.

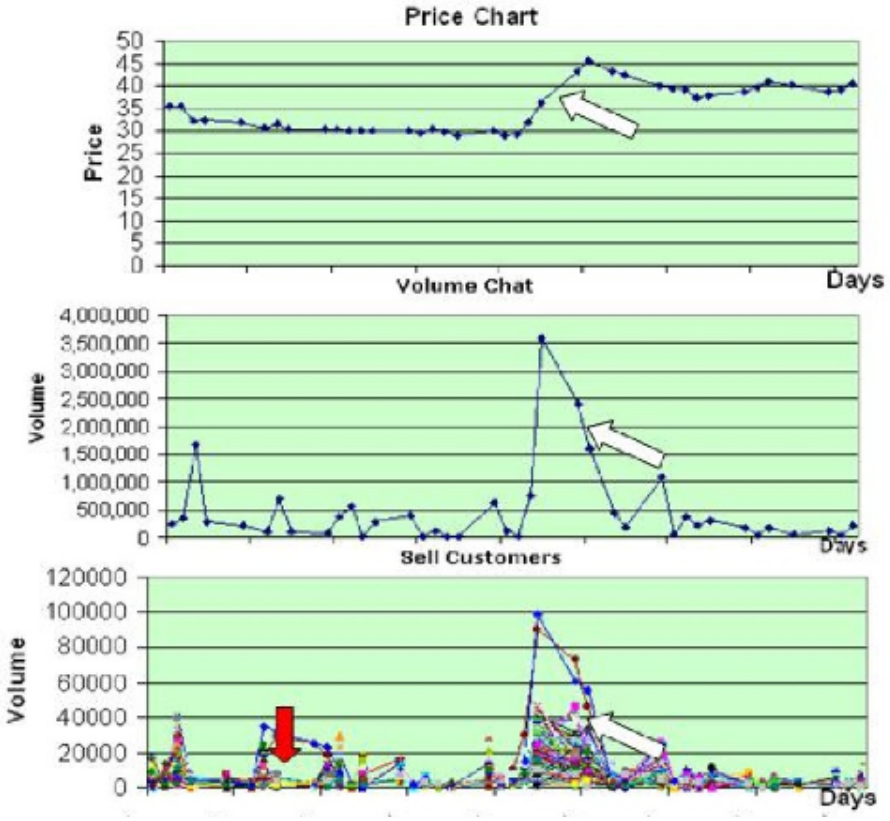

Figure 3.2: Insider Trading (P Perera 2008)

Domain experts identifies such cases as explained above considering their past experience. There are multiple parameters have to be taken into consideration when confirming a fraud behavior. It is hard to implement a high performing surveillance system because it is not a clear-cut task. It is very challenging to analyze real time because transactions occurs at a very high frequency in a market.

Chapter 4: System Design

Design for Artificial Immune System based Anomaly Detection System for stock markets will be discussed in this chapter. Natural Immune System model and its techniques will be used in this design to optimize the output of the system.

Since the anomaly detection system model is not directly supports Immune system model, using those techniques in anomaly detection system model has been done after numerous design considerations which will be stated in the rest of the chapter.

4.1 Design Problems and Solution Analysis

Difficulties faced when designing the solution for the main problem are listed below along with the solutions for each design problem.

1. When we try to detect anomalies with the input data stream, we cannot directly identify a fluctuation as an anomaly. That depends on degree of average fluctuation of whole data stream. Even though we captured a fluctuation in input data stream, that may not abnormal when considering the past behavior of data values.

   Considering the nature of the problem as above, we cannot create anomaly detectors to be used in negative selection algorithm.

   Solution for the above design problem is proposed as to get the benefits of Danger Signal concept. Danger for the system can be defined as abnormal fluctuation of Price along

with abnormal fluctuation of the corresponding Volume. System generates danger signal as above and system will act accordingly. After investigations help to prove whether identified scenario is a stock manipulation, system then store that and will use it to identify future manipulations. In that way we can approach implementing Negative selection algorithm.

2. It is not possible to have a global threshold value for data streams and individual behaviors for all Symbols. Abnormality cannot be identified for price and volume by a universal value. If a company stock price is 500, then price change of a 10 may not an abnormal change. But if another company has a stock price of 10, there is a more probability to be an abnormal change if their price changed by 10.

   Solution for the above design problem is to calculate dynamic threshold values. That threshold values will be valid for particular company and particular time period only. They will not be reusable even for same company because price or volume change is not consistent.

3. Since the defining boundary values to detect abnormal or normal cases is not easily and straight forward task, possibility of growing false negative and false negative error rate is high. Even the experts cannot define exact threshold values directly. One of the main consideration of the main solution is to minimize false negative error which is considering particular abnormal scenario as normal.

   Solution for this design problem is to facilitate to change sensitivity if the system by allowing external change of sensitivity variables of the system, which leads to reduce error rate for particular data set.

4. There is a problem that even though the system captured certain scenarios as stock manipulation, there are other factors that need to be considered when concluding a scenario as a stock manipulation. Those factors are hard to capture but scenario depends on it.

   Solution is to that particular problem is to get the expert knowledge on these kind of scenarios. Otherwise same kind of errors can be reoccurred by the system. Since experts carry investigations on suspected scenarios, feedback can be fed to the system.

5. After the initial learning phase of the system, new suspected scenarios should be able to be identified. System anyway keep on learning with feedback continuously. System identifies a new suspected scenario in two steps.

   - Initially Suspicious – When new transaction is fed to Price/Volume anomaly detecting component it predicts the possibility of being a fraud and if it passes the threshold, it is considered as initially suspicious transaction.
   - Confirmed Suspicious - After finding so called initially suspicious transaction, it is fed to 'Customer Anomaly Detector' and match customer details and confirm for suspicious scenario. Details of confirmed suspicious scenarios are fed back to the system to optimize next identification.

4.2 System Components

Below diagram illustrates system design, what are the key components and the work flow.

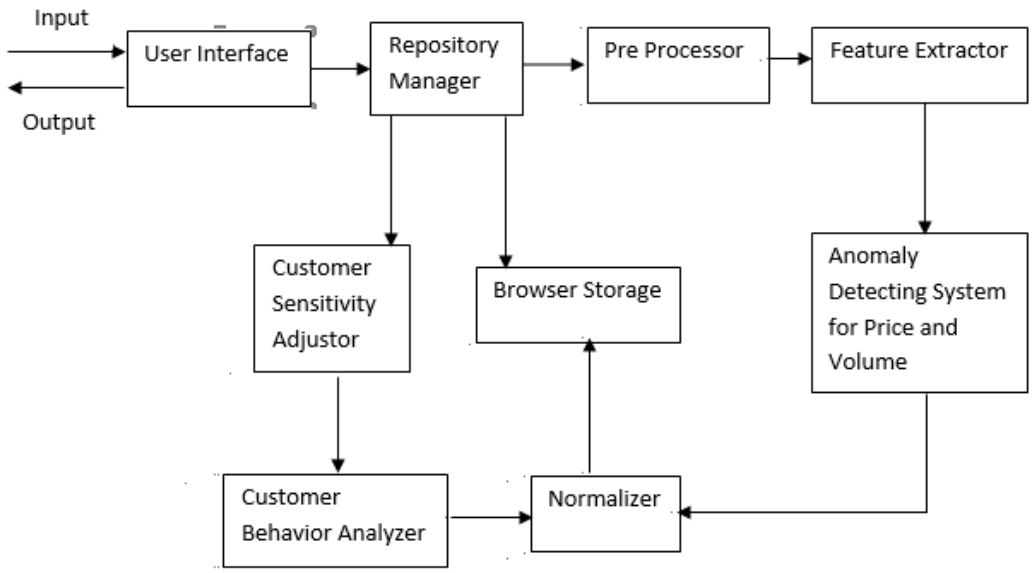

Figure 4.1: System Work flow

The role of each component in the system will be discussed in the rest of the section.

4.2.1 Web Interface

This component represents abstract system to external user. User inputs are taken to system and system outputs are given to user via User Interface. Below are the list of functionalities of Web Interface Component.

- Showing Results of system output
- Set Initial parameters to system
- Adjust sensitivity by user
- Get user feedback

This is a critical design decision to use browser web as system container because there are limitations of web application than a desktop application such as limitations of accessing computer storage etc. But current usages of web applications are growing due to many reasons such as easily accessible, use with many devices which having Internet facility, light weight etc. Therefore is solution will be more closer to end users but very challenging for researcher. Challenges and the way they have overcome will be explained under each topic of rest of the chapters.

4.2.2 Repository Manager

This component is responsible performing three tasks during system run.

1. Format input dataset and make them ready to pre-process.
2. Improving training data set which gives better results
3. Executing normalizing process to improve suspicious scenarios

This functionality is based on Immune network theory concepts. This is executed depending on expert feedback given for the system output. This will ensure the quality of the detectors in a way that the feedback from environment is received and controls the memory accordingly. But immune network theory component to demoting and killing detectors which are not used recently is not used here because anomalies which can be popped up very rarely should be identified as well.

4.2.3 Browser Storage

Browser Storage is responsible for holding the detector set for the system. Confirmed cases with price/volume fluctuations, normal transactions and suspected customer behavior are stored in browser storage. Stored detectors are used in identifying newly suspected cases.

This component gives stored cases with degree of anomaly as the output. This aligns with the detector set concept of the negative selection algorithm whereas repository manager optimizes the quality of the dataset.

This also is the main repository to store all transaction details fed to system. Information included symbol information, price and volume information for certain time period, transaction no, exchange commission, customer information and etc.

4.2.4 Data Processor

This component is responsible for process input data and prepare them for feature extraction phase. Corresponding steps of data processing are stated below.

1. Eliminate invalid transactions.

2. Eliminate unwanted columns.

3. Data Categorization

Implementation details of this component is described in the next chapter.

4.2.5 Feature Extractor

This component is responsible for calculating necessary feature values required for learning and testing phases.

Price and volume anomaly detector is based on machine learning techniques and required features are calculated and data models are prepared according to input models of classifier.

4.2.6 Customer Sensitivity Adjuster

Functionality of this component is to adjust the impact of the customer behavior to make a predicted transaction to a abnormal one. This component is managed externally and calibrated adjuster value to a optimized one with training data provided. if there is no training data for current scenario, a default value will be used.

4.2.7 Customer Behavior Analyzer

Functionality of this component is to analyze customer behavior in transactions whether suspicious or not.

Analyzer will provide results such that system will store analyzed behavior in order to normalize abnormal outputs received from price, volume anomaly detector and will be used in identifying future anomalies in customer behavior.

4.2.8 Anomaly Detector for Price and Volume

This component's functionality is to detect abnormal fluctuations of price and volume for particular time period. Machine Learning techniques are used in identifying abnormality. Since there are no global threshold values and those vary even for one company, probability of being a anomaly will be received. Methodology is based on supervised learning techniques.

Price and volume data streams are fed as the input and degree of abnormality of the input values are provided as the output. Functionality of this component is aligned with the Danger Signal theory concepts.

4.2.9 Normalizer

This component takes of price, volume anomaly probability and analyzed results of customer behavior as input and provides possible suspicious scenarios as the output.

Price and volume anomaly detector initially identifies suspected fluctuations in price and volume and this component then normalize the probability depends on suspected individual behavior by comparing their behavior with its peer groups. Comparison is performed assuming that all peers should behave in same way since all the stakeholders are having same information about market behavior.

Chapter 5: System Implementation

This chapter will explain detail implementation of each component designed in previous chapter. Related techniques will be pointed out which will be solutions for particular design problems explained in System Design chapter.

5.1 Web Interface

This system consists of a user interface because there are some external user interactions need to handle the system. Web implementation has been done using JavaScript framework called Ember. There was a doubt regarding implementation architecture can be limited using web interface, but with latest web technologies like HTML5 and Ember has quite useful features to implement whole system in web based manner.

Below are some requirements of User Interface.

1. Sensitivity adjustment - sensitivity is different from company to company, so they should be adjusted via UI
2. Show tables indicating suspected scenarios
3. Show system output
4. Get user feedback
5. Show real time price changes of companies

In this web component, there are separate pages for each task and models that has been implemented. Below is the categorization of pages.

1. Data page - Training and Testing data are provided to the system
2. Training and Testing pages - pages contains different models that are used for system implementation.

Detail evaluations of different models and results will be discussed in the next chapter.

5.2 Repository Manager

This component manages input data and processed data of some of the system components in order to provide quality data for components to get an accurate output.

First task is to format dataset in order to provide them to pre-processor. Data is read and identifies details of a single transaction and group them transaction basis.

Then the header component is filtered in order to make transaction objects. Transaction objects are created in Json format as key value pairs containing transaction details. Transaction objects are put in to arrays and send to pre processing.

And also Repository Manager manages storing customer details sent by Customer Behavior Analyzer and probability values taken as the output of anomaly detector for price and volume.

5.3 Browser Storage

Browser storage is considered as the main repository of the system as latest browsers have features like Local Storage, Session Storage and Application Cache which can be used to contain substantial amount of meta data. Storages are in mega byte scale and all the information stored in this system is stored in Local Storage which was well enough. Browser Storage is responsible for holding the detector set for the system. Below are the categories of data stored in browser storage.

1. Train data - Training data sets are stored which are used in training various models of price, volume anomaly detectors
2. Test data - Testing data sets are stored which are used in testing various models of price, volume anomaly detectors
3. Customer Data - Customer behavior map with weights are stored learned from previous transactions.

5.4 Data Processor

Implementation of this component is performed with a purpose of process input data and prepare them for feature extraction phase. Corresponding steps of data processing are described below.

1. Eliminate invalid transactions.

   In the case of processed transaction (mainly in training data), validity of the transaction is checked here. Invalid transactions are the transactions which are not successfully processed by exchange. Price mismatch, quantity mismatch, invalid symbol, buying power is not enough are some of the reasons to transactions become invalidated.

2. Eliminate unwanted columns.

   After filtering successful transactions, system is working on extracting necessary details of corresponding transactions. From set of the details of transaction, only price, volume, customer details and company will be extracted.

3. Data Categorization

   Data Categorization has been done to be used in different kinds of models proposed for solution. Ex: Categorizing company wise data

5.5 Feature Extractor

This component is implemented in order to calculate necessary feature values required for learning and testing phases.

From transaction details, only price and volume related features are extracted from the feature extractor. System evaluates customer behavior in separate component. Other details are dependent fields of price and volume fluctuations. Following are the functional steps of feature extractor.

1. Calculate 'Average Price' of normal transactions for a particular model - MP
2. Calculate 'Average Volume' of normal transactions for a particular model - MV
3. Calculate price difference

$$PD = P - MP$$

   PD - Price Difference

   P - Price of transaction

   MP - Mean Price

4. Calculate volume difference

$$VD = V - MV$$

VD - Volume Difference

V - Volume of transaction

VP - Mean Volume

Particular array consists of price differences and volume differences will be prepared which requires for price, volume anomaly detector

5.6 Customer Sensitivity Adjuster

As mentioned in Design chapter, functionality of this component is to adjust the impact of the customer behavior to make a predicted transaction to a abnormal one.

This adjustment value is provided externally by user in order to optimize customer weight values defined through Customer Behavior Analyzer. Optimum adjusted value vary for each model tested, obtained via testing system to achieve better anomaly detection rate.

5.7 Customer Behavior Analyzer

Main objective of this component is to detect anomalies of customer behavior which is identified by acting quite different from their peers.

Since price variations are there due to amount of volume in transactions, main technique is to check abnormal behavior of parties, is to check abnormal level of volume is there in their transactions. As an example, if a buyer buys 50,000 of volume but the average volume in

transactions of its peers is around 100, that buyer is noted as a suspect, initially. And also due to sudden high demand, price of the particular stock goes up.

Using training and later with testing data, suspicious behavior of customer are analyzed and create a map of customers with corresponding anomaly weight. This customer behavior summary is used in normalizer to normalize suspicious transactions received from price and volume anomaly detector.

Another technique used to detect abnormal customer behavior is analyze the time difference between same customer's successive transactions. Weight of the time difference is higher when customer does transactions more frequently. This feature is fed to system to identify some of the advanced anomalies such as 'splitting'.

5.8 Anomaly Detector for Price and Volume

Main objective of this component is to detect anomalies of price and volume values of stock market transactions. Probability value which represents degree of anomaly will be calculate for price and volume differences.

There were several tests have been carried out and following technique is found which gives most promising results for Price and Volume Anomaly Detector.

In order to get an accurate output of the component, supervised machine learning techniques are used to detect anomalies of price and volume. System is tested with various machine learning algorithms (results comparison will be stated in the next chapter) and optimum algorithm among them was Naive Bayes Algorithm.

Naive Bayes Classifier implemented using NodeJS is used to train and test the price and volume anomalies. Features produced from Feature Extractor component is used as the input, and

output is the component is probability of being an abnormal transaction.

Below are the functional steps of the component.

1. Initialize Classifier with required parameters
2. Input Training data array consists of features extracted for price and volume - Training data headers are Price Difference, Volume Difference and Anomaly Index
3. Train the classifier with provided data
4. Testing data are fed to classifier - Testing data headers are Price Difference and Volume Difference
5. Obtain results for test data

These abnormal probabilities are fed to Normalizer component to add the affect of customer behavior to transaction abnormality.

5.9 Normalizer

As mentioned in design chapter, this component takes of price, volume anomaly probability and analyzed results of customer behavior as input and provides possible suspicious scenarios as the output.

Methodology of the Normalizer is, getting inputs of price, volume abnormal probabilities and customer map consists of customer anomaly weights and customer sensitivity adjuster value.

Lets take a particular transaction with customer sensitivity adjuster value SAV, customer anomaly weight WC and price, volume probability PPV. Below equation gives the normalized abnormal probability PN of the transaction.

$$PN = PPV + WC * SAV$$

If this normalized abnormal probability of a transaction is greater than that of its normal probability, this transaction is considered to be abnormal.

Chapter 6: Evaluation

Evaluation criteria consists of following steps to find anomalies in stock market transactions.

1. Anomaly detection of Price and Volume data
2. Anomaly behavior of customer
3. Anomaly of overall transaction

Below are the various models and combinations tested in this phase. Detailed evaluation results and related comparisons are done in the rest of the chapter.

1. Evaluation models of different classifiers
2. Evaluation of the models with different feature sets
3. Evaluation of system with single company data
4. Evaluation of system with multiple company data with parallel classifiers
5. Evaluation of system for real time transactions of all companies traded at particular time period - single classifier
6. Comparison this system performance with DirectFN rule based anomaly detector

6.1 Transaction Data

Testing has been carried out with real market data collected for several companies listed in Saudi Stock Exchange. System training and evaluation phase is performed effectively with real transaction data along with real customer data.

A set of data has been analyzed by three different domain experts and marked real manipulation scenarios, which is effectively used to train the system. sensitivity parameters has been adjusted such that to identify known cases with low error rate with the use of above data set.

Other transaction data is used to test the system. testing has been carried out with data sets of different companies in various models stated earlier.

6.2 Training and Testing Models

Training and testing has been carried out in following models and accuracy, precision, false negative rate are calculated which is required for system evaluation.

6.2.1 System Models with Different Classifiers

In this case, system is trained and tested for single company data for different classifier models. Below are the classifiers used for evaluation.

1. Naive Bayes Classifier
2. Support Vector Model Classifier

3. Logistic Regression Classifier

Naive Bayes Classifier

Extracted features are arranged suitable for Naive Bayes classifier inputs and initialize the classifier. After that arranged input data of company 'Saudi Industrial Investment Group' is fed to system.

```
73              var classifier = new bayes['default'].NaiveBayes({  classifier = NaiveBayes
74                  columns: trainingColumns,  trainingColumns = (3) ["priceDiff", "volumeDi
75                  data: arrayTraining,  arrayTraining = Array(610)
76                  verbose: true
77              });
78
79              classifier.train();  classifier = NaiveBayes {stripWhitespace: true, columns
80
81              this.testDataBayes(classifier);
82          },
```

Figure 6.1: Bayes Classifier Implementation

Below table denotes the classifier output results of Naive Bayes Classifier. Please note that results are not normalized with customer data. Degree of anomaly of a transactions is marked based on conclusions of domain experts and system output is evaluated based on them. All the percentages are calculated for system output with respect to conclusions of domain experts.

Measurement	Value
Accuracy	88.81%
Recall	100%
Precision	38.18%
F1 Score	55.26

Table 6.1: Naive Bayes Classifier results

Support Vector Model Classifier

Support Vector Machines (so called SVMs) are also generally used as supervised learning technique and since this solution also using supervised leaning techniques, tested the system and obtained results with SVM classifier.

Support vector machine builds single or multiple hyperplanes in a high or infinite dimensional space, which can be used for classification. Extracted features are arranged suitable for SVM classifier inputs and initialize the classifier. After that arranged input data of company 'Saudi Industrial Investment Group' is fed to system and evaluate. One significant difference of input data of SVM compared to Naive Bayes is, abnormal indexes of training data provided separately as an array.

```
302
303             var svm = new ml['default'].SVM({
304                 x: arrayTrainingIn,
305                 y: arrayTrainingOut
306             });
307
308             svm.train({});
309
310             this.testDataSVM(svm);
```

Figure 6.2: SVM Implementation

Below table denotes the classifier output results of SVM Classifier. Please note that results are not normalized with customer data. Degree of anomaly of a transactions is marked based on conclusions of domain experts and system output is evaluated based on them. All the percentages are calculated for system output with respect to conclusions of domain experts.

Measurement	Value
Accuracy	93.03%
Recall	0%
Precision	0%
F1 Score	0

Table 6.2: SVM Classifier results

Above table shows that SVM model is not capable of identifying any of the positive anomaly. Therefor Recall, Precision and F1 score is zero. And also this classifier took longer time to classify compared to other two classifiers.

High accuracy does not make this classifier interested because anyway suspicious cases are not very common as normal cases in stock market transactions, but need to identify those suspicious cases.

Logistic Regression Classifier

Logistic Regression Classifier is also generally used as supervised learning technique, same as previously discussed classifiers, and since this solution also using supervised leaning techniques, tested the system and obtained results with Logistic Regression classifier.

Logistic regression is a technique which is borrowed by machine learning from the field of statistics. Unlike linear regression which generates outputs with the format of continuous number values, with the logistic regression technique its output is transformed using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes. For this solution we need only binary classification.

Extracted features are arranged suitable for Logistic Regression classifier inputs and initialize the classifier. After that arranged input data of company 'Saudi Industrial Investment Group' is fed to system and evaluate. No of training epochs used is 800 and learning rate is 0.05.

One significant difference of input data of Logistic regression compared to Naive Bayes is, abnormal indexes of training data provided separately as an array and it is same as in SVM.

```
214             var classifier = new ml['default'].LogisticRegression({  class
215                 'input': arrayTrainingIn,  arrayTrainingIn = Array(610)
216                 'label': arrayTrainingOut,  arrayTrainingOut = Array(610)
217                 'n_in': 2,
218                 'n_out': 2
219             });
220
221             var training_epochs = 800,  training_epochs = 800
222                 lr = 0.05;  lr = 0.05
223
224             classifier.train({  classifier = module.exports {x: Array(610)
225                 'lr': lr,  lr = 0.05
226                 'epochs': training_epochs  training_epochs = 800
227             });
228
229             this.testDataReg(classifier);
230         },
231
```

Figure 6.3: Logistic Regression Implementation

Below table denotes the classifier output results of logistic regression Classifier. Please note that

results are not normalized with customer data. Degree of anomaly of a transactions is marked based on conclusions of domain experts and system output is evaluated based on them. All the percentages are calculated for system output with respect to conclusions of domain experts.

Measurement	Value
Accuracy	93.03%
Recall	0%
Precision	0%
F1 Score	0

Table 6.3: Logistic Regression Classifier results

Above table shows that logistic regression gives same results as SVM model and it too is not capable of identifying any of the positive anomaly. therefor Recall, Precision and F1 score is zero. High accuracy does not make this classifier interested because anyway suspicious cases are not very common as normal cases in stock market transactions, but need to identify those suspicious cases.

After evaluating above classifier performances, Naive Bayes Classifier is selected as the promising classifier and it will be used in rest of the models evaluated in this chapter.

6.2.2 Models with different feature sets

In this model, system is trained and tested for single company data only. But different feature sets are used to test the system and results will be compared to chose the best feature set. Examples of feature sets are described below with its output.

Price, Volume, Commission and Transaction Time

Results of the classifier prediction is stated below. Capability of identifying anomaly scenarios are not up to the standard. Degree of anomaly of a transactions is marked based on conclu-

sions of domain experts and system output is evaluated based on them. All the percentages are calculated for system output with respect to conclusions of domain experts.

Measurement	Value
Accuracy	70.34%
Recall	89.35%
Precision	22.38%
F1 Score	34.58

Table 6.4: Results with Price, Volume, Commission and Transaction Time as features

Below table denotes other feature combinations tested and their results.

Price	Volume	Commission	Time	side	Accuracy	Recall	Precision
X	-	-	-	-	55.98%	35.36%	15.33%
X	X	-	-	-	65.45%	38.38%	23.55%
X	X	X	-	-	64.33%	37.25%	25.85%
X	X	X	X	-	70.34%	89.35%	22.38%
X	X	X	X	X	66.55%	88.22%	26.90%
X	-	-	X	X	41.23%	33.24%	12.35%
-	X	-	X	-	40.25%	34.55%	11.53%
-	X	-	X	X	42.44%	38.22%	12.29%
X	-	X	X	-	63.76%	37.55%	22.43%
X	-	X	-	X	64.55%	38.63%	24.36%
-	X	X	X	-	48.98%	66.70%	33.54%
-	X	X	X	X	66.45%	67.26%	36.76%

Table 6.5: Results with Different Combination of features

Some of the possible reasons for getting results of less accuracy have been listed for tested features.

1. Price and Volume values itself does not are cannot be considered as features because they fluctuate with time even in normal cases

2. Commission linearly depends on Price and Volume and useless feature to be used along with Price and Volume

3. Abnormality of a transaction hardly depends on Transaction Time and also dateTime string itself is not a feature.

Price Difference, Volume Difference, Time Difference

Price difference is calculated as the difference between price and and mean price. And Volume difference is calculated as the difference between volume and and mean volume. Time difference is calculated as time taken to perform this transaction from previous transaction, in seconds. Results of the classifier prediction is stated below. Capability of identifying anomaly scenarios are optimum in this case. Degree of anomaly of a transactions is marked based on conclusions of domain experts and system output is evaluated based on them. All the percentages are calculated for system output with respect to conclusions of domain experts.

Measurement	Value
Accuracy	96.38%
Recall	100%
Precision	65.63%
F1 Score	79.25

Table 6.6: Results with Price Difference, Volume Difference and Time Difference as features

After analyzing these results of different feature sets, price difference and volume difference and Time Difference are considered as features for other models discussed in this chapter.

6.2.3 Single Company Data Model

In this model, system is trained and tested for single company data only. Sensitivity is also set depending only on this company data.

Company symbol '2250' (company Name is 'Saudi Industrial Investment Group') which is listed in Saudi Stock Exchange has been selected for testing.

After cleaning and pre-processing data, 914 valid transaction records were there in data model, which is divided nearly 2 : 1 ratio for training and testing data.

Company	Price	Quantity	Tr. Time	Status	Commission
2250	24.75	1,200	20141130-08:37:23	2	5.35
2250	24.60	2,000	20141130-08:32:37	2	8.86
2250	24.50	2,000	20141130-08:32:21	2	8.82
2250	24.60	2,000	20141130-08:26:49	2	8.86
2250	24.50	2,000	20141130-08:23:25	2	8.82
2250	24.40	1,000	20141130-08:19:33	2	4.39
2250	24.35	1,500	20141130-08:18:59	2	6.57
2250	24.30	2,000	20141130-08:14:35	2	8.75
2250	24.25	1,000	20141130-08:14:00	2	4.36
2250	24.30	2,000	20141130-08:10:16	2	8.75
2250	24.30	1,000	20141130-08:06:58	2	4.37
2250	24.15	500	20141130-08:05:13	2	2.17
2250	24.10	1,000	20141130-08:05:11	2	4.34
2250	24.15	500	20141130-08:04:05	2	2.17
2250	24.05	3,000	20141130-08:00:01	2	12.99
2250	27.00	2,000	20141127-10:45:42	2	9.72
2250	27.00	4,000	20141127-10:45:20	2	19.44

Figure 6.4: Transactions of Company Symbol 2250

```
48
49          this.set('trainContent', train);  train = Array(610)
50          this.set('testContent', test);  test = Array(304)
51
```

Figure 6.5: Training and Testing Data

After that features are extracted and training the system. Classifier is selected as Naive Bayes classifier for this model. Performance comparison of different classifier models has been discussed in previous sections.

```
73          var classifier = new bayes['default'].NaiveBayes({  classifier = NaiveBayes
74              columns: trainingColumns,  trainingColumns = (3) ["priceDiff", "volumeDi
75              data: arrayTraining,  arrayTraining = Array(610)
76              verbose: true
77          });
78
79          classifier.train();  classifier = NaiveBayes {stripWhitespace: true, columns
80
81          this.testDataBayes(classifier);
82      },
```

Figure 6.6: Bayes Classifier Implementation

After classification by price, volume detector accuracy, precision, recall and F1 score has been calculated. Then those abnormal probability values are normalized using customer weights. Degree of anomaly of a transactions is marked based on conclusions of domain experts and system output is evaluated based on them. All the percentages are calculated for system output with respect to conclusions of domain experts. Results of the system before normalization and after normalization are stated below.

Measurement	Value
Accuracy	88.81%
Recall	100%
Precision	38.18%
F1 Score	55.26

Table 6.7: Single Company Results without Normalize

Below is company results after normalization.

Measurement	Value
Accuracy	96.38%
Recall	100%
Precision	65.63%
F1 Score	79.25

Table 6.8: Single Company Results with Normalize

6.2.4 Parallel Classifier Model

In this model, system is trained and tested for multiple company data. Each classifier is trained with corresponding company data and multiple classifiers are activated and test company data parallelly.

One classifier testing and evaluation steps are same as that described in previous sub section. Apart from company symbol '2250' (company Name is 'Saudi Industrial Investment Group'), company symbol '1010' (company Name is 'Riyad Bank') has been selected for testing.

Company	Price	Quantity	Tr. Time	Status	Commission
1010	17.85	1,900	20141130-09:19:26	2	6.10
1010	18.70	3,352	20141120-11:42:26	2	11.28
1010	18.70	2,000	20141120-11:21:08	2	6.73
1010	18.70	1,000	20141120-11:19:46	2	3.37
1010	18.65	350	20141120-11:01:20	2	1.17
1010	18.65	1,000	20141120-10:58:19	2	3.36
1010	18.65	2,000	20141120-10:52:33	2	6.71
1010	18.65	600	20141120-10:27:38	2	2.01
1010	18.65	1,000	20141120-10:15:50	2	3.36
1010	18.70	2,000	20141120-09:57:05	2	6.73
1010	18.65	2,000	20141120-09:39:26	2	6.71
1010	18.65	500	20141120-09:19:04	2	1.68
1010	18.70	500	20141120-09:07:14	2	1.68
1010	18.70	500	20141120-09:05:48	2	1.68
1010	18.70	1,000	20141120-09:05:37	2	3.37
1010	18.70	500	20141120-08:50:55	2	1.68
1010	18.70	250	20141120-08:47:29	2	0.84

Figure 6.7: Transactions of Company Symbol 1010

After classification by price, volume detector, those abnormal probability values are normalized using customer weights and accuracy, precision, recall and F1 score have been calculated. All the percentages are calculated for system output with respect to conclusions of domain experts.

Results of the parallel classifier models are stated below.

Measurement	Value
Accuracy	95.07%
Recall	65%
Precision	100%
F1 Score	78.78

Table 6.9: Parallel Classifier Model results for 1010

We can evaluate that abnormal detection of the system with parallel classifiers performs same as single company data model by comparing results of Company '2250'.

Measurement	Value
Accuracy	96.38%
Recall	100%
Precision	65.63%
F1 Score	79.25

Table 6.10: Parallel Classifier Model results for 2250

6.2.5 Real Time All Company Data Model

In this model, system is trained and tested for all companies with single classifier. This scenario is more towards practical situations as solution is expected to detect transaction anomalies in real time if possible. Training and testing data obtained for all companies traded for particular time period are sorted with transaction time and provide as inputs to the system.

After classification by price, volume detector accuracy, precision, recall and F1 score have been calculated. Then those abnormal probability values are normalized using customer weights. Degree of anomaly of a transactions is marked based on conclusions of domain experts and system output is evaluated based on them. All the percentages are calculated for system output with respect to conclusions of domain experts. Results of the system after normalization are stated below.

Company	Price	Quantity	Tr. Time	Status	Commission
4007	65.00	288,461	20140803-08:00:03	2	3,374.99
6002	104.00	111	20140803-08:00:03	2	2.08
2010	129.00	90	20140803-08:00:01	2	2.09
1810	131.75	88	20140803-08:00:01	2	2.09
4240	107.25	108	20140803-08:00:01	2	2.08
4002	107.50	108	20140803-08:00:01	2	2.09
4001	112.00	104	20140803-08:00:01	2	2.10
2290	72.50	160	20140803-08:00:01	2	2.09
2040	139.00	3,000	20140724-12:32:05	--	19.49
1040	49.50	1,000	20140724-12:20:21	2	8.91
2060	36.70	10,000	20140724-12:13:33	2	66.06
6004	189.75	450	20140724-12:11:41	2	15.37
1090	43.90	1,500	20140724-12:04:19	2	11.85
1040	50.00	7,000	20140724-11:58:14	2	63.00
1040	49.90	7,000	20140724-11:56:25	2	62.87
1040	49.90	5,000	20140724-11:55:59	2	44.91
4003	122.00	2,920	20140724-11:53:44	--	12.30

Figure 6.8: Transactions of Multiple Companies Traded

Measurement	Value
Accuracy	92.42%
Recall	13.88%
Precision	55.55%
F1 Score	22.22

Table 6.11: Results of Real Time All Company

Some of the possible reasons for getting results of less accuracy have been listed for tested features.

1. Price and Volume value averages are not much meaningful when many companies contribute for calculation
2. Training data abnormalities are defined initially, are not related with the other companies traded in same time period
3. Abnormality of a transaction changes from company to company, as one transaction is abnormal for particular company but not for another company

6.2.6 Comparison with DirectFN rule based anomaly detector

DirectFN is one of the largest ICT companies in Sri Lanka and Middle East, who offers stock market related solutions to trade and view price data, and also provides a rule based anomaly detection system called AML

In AML, we can filter suspected transactions for particular time period. For the same period of time and for particular company (we have chosen company '2250'), we have collected predictions from DirectFN AML, predictions from our solution and predictions of domain expert who analyzed transactions of that particular period for company '2250'. Evaluation is executed with the assumption of domain expert's predictions are 100% correct and results of other two predictions are analyzed compared to it.

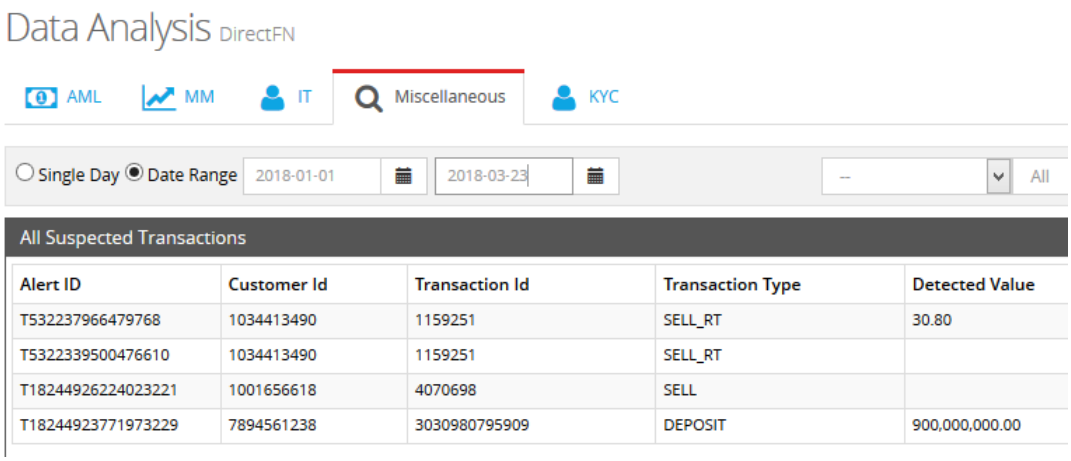

Figure 6.9: DirectFN AML

Initial Detail	Value
Total No of Records	304
Expert's Positive Count	21

Table 6.12: Initial Data Parameters

Below tables are showing predictions of this system and DirectFN AML, and accuracy, precision, recall and F1 score have been calculated compared to domain expert predictions.

Below table shows obtained results from DirectFN AML.

Measurement	Value
Predicted True Positive Count	5
Accuracy	91.12%
Recall	23.81%
Precision	31.25%
F1 Score	27.02

Table 6.13: Results of DirectFN AML

Below table shows obtained results from our solution for same input data.

Measurement	Value
Predicted True Positive Count	21
Accuracy	96.38%
Recall	100%
Precision	65.63%
F1 Score	79.25

Table 6.14: Results of our solution

Some of the possible reasons for getting erroneous results for DirectFN AML have been listed

below.

1. Accuracy of AML is fair but relatively low, but that measurement is not that valid for systems that have low no of positive scenarios.

2. Since rules are common for all companies, higher margins for parameters have applied for AML. Therefore true positive recognition ability is low and it is indicated with low precision.

3. Some of the highly traded or rarely traded company transactions are captured as anomalies even if they are normal for the company. That leads to decrease recall value and also F1 score.

Chapter 7: Conclusion and Future Work

In this conclusion chapter, to what extent initial objectives are satisfied, will be discussed.

7.1 Conclusion

The main project objective was to research and implement a method to detect these evolving stock abusing patterns.

Main objective was divided into two sub objectives.

- Implementation of price, volume fluctuation detector which is capable of finding abnormalities of a data stream and calculating degree of abnormality of a suspected fluctuation.
- Heuristic function for above system to increase precision and recall values when detecting suspected scenarios.

First sub objective is achieved with successful rates detecting stock market anomalies. Statical calculations are used initially but detection capability of supervised learning system was much higher. Therefore, after testing and evaluating various classifiers, best classifier system for this task is used in testing other models.

Then feature set is also optimized testing various sets of features and obtaining results and analyzing what are the reasons for those results. Optimum and minimum feature set was price

difference compared to average price, volume difference compared to average volume and time difference compared to previous transaction.

Second sub objective is achieved with analyzing customer behavior and add its biased value to results of price, volume detector and taking the final output of the system. Customer behavior analyzer go through customer details of the transactions and learn its detector map. High fraud customers are having higher weights and having good capable of minimizing error rate. This functionality of customer behavior analyzer and Normalizer is aligned with Clonal Selection of Artificial Immune System theory.

Below is the results obtained for Single Company, Normalized Model

Measurement	Value
Accuracy	96.38%
Recall	100%
Precision	65.63%
F1 Score	79.25

Table 7.1: Results of Best Model Addressed

Below are some reasons for output as above.

1. Good Score for all measures due to optimization of classifier, feature set and methodology depend on Artificial Immune theories.

2. Can be further optimized solution by increasing no of training transactions, adding domain components (as splitting the stocks) and etc.

7.2 Future Work

In this solution, initial identification of frauds with Price and Volume is moved from statical techniques to supervised learning techniques. And further normalized results with customer behavior analyzer.

This solution can be further optimized by adding domain specific scenario effect to transactions more and more. One such example is Stock Splitting. When the stock value become to a saturated level, price is set to half of the initial price and make no of stocks twice. We have to neglect those price, volume fluctuations as they are not anomaly behavior.

Only suspected behavior identification is performed in this solution and have to proceed with manual check for each suspected scenario produced. That manual check should be done without letting customer know because if it is not a fraud, customer will be disappointed for suspecting him.

This solution is not tested with a time series model with time series theories. Taking average price and volume fluctuation value of particular scenario can be further fine tune to small period of time may be lead to better results. And also new feature sets also can be tested obtained from time series data.

Domain expert knowledge is used to pre identification of suspicious scenarios and it will be better if the user can train the system with real confirmed scenarios, then the system will be more towards identifying real transaction anomaly. Real frauds are rare compared to normal transaction and rare anomalies may have to subjected for better classification technique apart from techniques used in this solution.

References

[1] Aickelin. The danger theory and its application to artificial immune system. *Proceedings of the 1st International Conference on Artificial Immune Systems, Canterbury, UK*, ICARIS-2002.

[2] D. & Timmis Castro. Artificial neural networks in pattern recognition. Master's thesis, University of Paislay, UK, 2002.

[3] Salvador & Chan. Learning status and rules for detecting anomalies in time series. Master's thesis, Applied Intelligent 23, 2005.

[4] Dasgupta & Forrest. Novelty detection of time series data using ideas of immunology. in 5th International Conference on Intelligent Systems, 1996.

[5] Ong & Overill (n.g) Kim. Design of an artificial immune system as a novel anomaly detector for combating financial fraud in retail sector. Master's thesis, Department of Computer Science, Kings College London, 2008.

[6] Mario Levis. Stock market anomalies: A re-assessment based on the uk evidence. *M Levis – Journal of Banking & Finance*, 1989 – Elsevier.

[7] Ferdousy & Maeda. Anomaly detection using unsupervised profiling method in time series data. in Proceedings of ADBIS Research Communications, 2006.

[8] Perera P.D.S.U. Abnormal pattern detection in time series data via artificial immune system model. Master's thesis, University of Colombo, School of Computing, 2008.

[9] Sharma D. Sharma A. Clonal selection algorithm for classification, in: Liò p., nicosia g., stibor t. (eds) artificial immune systems. ICARIS, 2011.

[10] Mick Swartz. Stock option returns and stock anomalies: Cross market efficiency and the cost of hedging value vs growth firms stock returns. *M Swartz - Journal of Business Economics and Finance*, 2013.

Appendix A: Immune System

Immune system techniques used in this project are described in detail in this section.

A.1 Natural Immune System

There are many systems which are capable of executing amazing functionalities in order to keep human body in normal state. They are complex in structure but using simple techniques to fulfill tasks. Immune system is also very important to humans since it is responsible for protecting human body from virus and bacteria and etc.

Immune system use negative selection technique to detect abnormal patterns. After detecting, immune response is mounted on foreign invaders. This process is called Clonal selection. Immune system maintains valuable detector repository by eliminating redundant detectors. Apart from this, dangerous alerts are identified using Danger theory.

A.2 Negative selection

The main objective of this process is to produce detectors which are capable of identifying foreign invaders. It is based on the main technique in the thymus that produces a set of mature T-cells which can bind only non-self antigens.

First step of this algorithm is to produce a set of self strings(P), that is considered as normal to the system. The main objective is to generate a set of detectors(M) that only recognize the completely opposite strings of S. These detectors can then be applied to new data in order to classify them as being self or non-self.

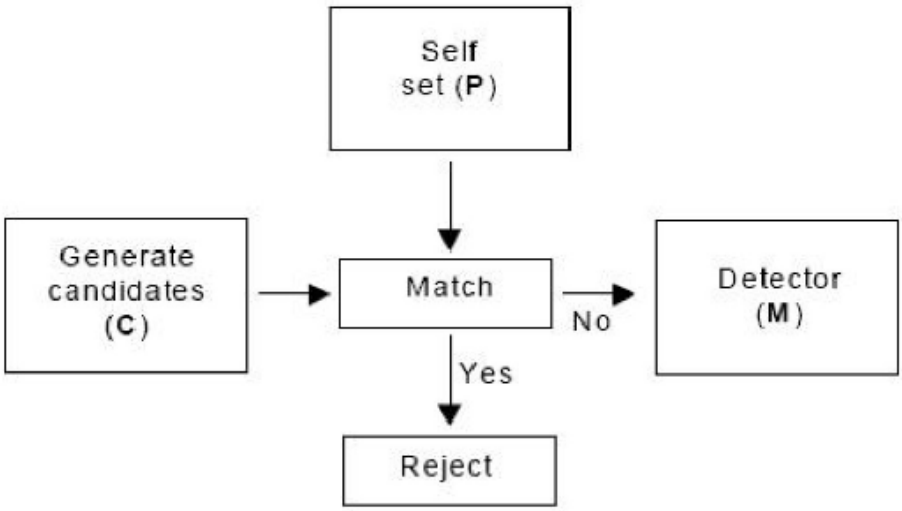

Figure A.1: Negative Selection

As described in above figure, immune system generates possible detectors randomly and then those detectors are sent through a mutation process. Generated candidates (detectors) are matched with sample self-cells and destroyed if matched. Likewise detectors are generated which might be matched with non-self-cells.

A.3 Clonal Selection

In Clonal selection process, when a detector identifies an antigen, it is subjected to proliferate process which is diversified detectors generated which are more capable of capturing same antigen next time. Detectors which are closely related to antigen will eliminate it and finally it will be stored.

Simply, when antigens get into the body and then attack the body, immune cells (B cells) are

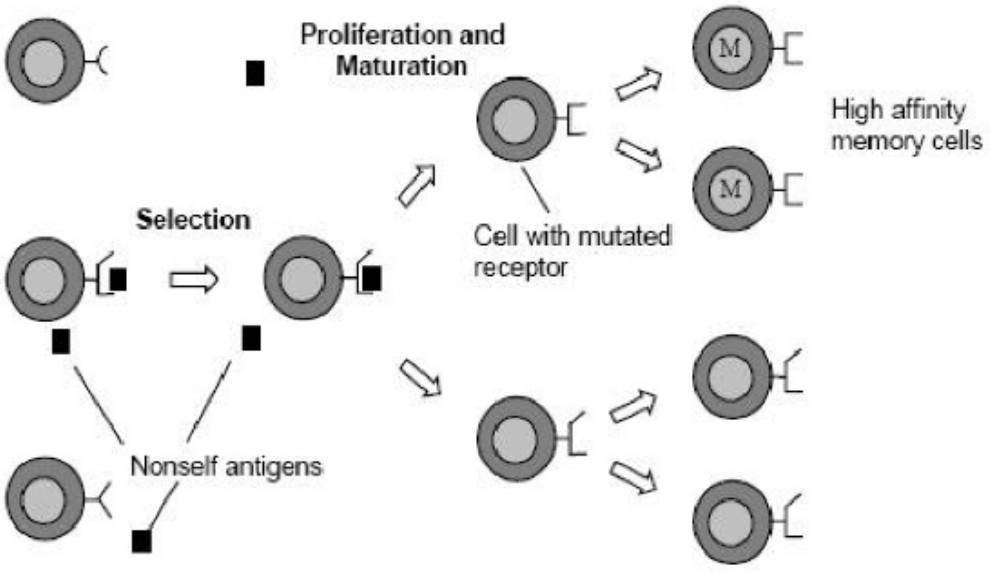

Figure A.2: Clonal Selection

activated and then respond by producing a specific antibodies in order to attack antigens. Antibodies are known to be molecules attached with B cells which are having a target of recognizing and catching antigens. A process called proliferation process will be executed to the cells that have successfully recognized the attacking antigens and produce two new types of cells within the process.

1. Attacking Cells
2. Memory Cells

The attacking cells act as a effective antibodies to defeat the attacking antigens immediately. The memory cells have a long-life span and their task is to attack future exposures of the same or similar antigens.[9]

A.4 Danger Theory

This is introduced to overcome some draw backs of negative selection process. Danger theory is well ahead of negative selection in scalability, fault rate and evolution of detectors.

In this process, danger signal is sent as a confirmation of detecting dangerous cell, therefore all the non-self-cells will not be considered as foreign invaders.

Appendix B: System Implementation Details

System Implementation Details are described in this section.

B.1 Transaction Data

Testing has been carried out with real market data collected for several companies listed in Saudi Stock Exchange. System training and evaluation phase is performed effectively with real transaction data along with real customer data.

Real Customer data has been obtained by DirectFN history databases and real time exchange feed. Since DirectFN is hosting Order Manage Systems for several brokerages, real data is saved in DirectFN data repository.

Data in database have been exported in the format of .csv and uploaded to ADS web site (project web interface). Whole transaction data contained more than 300,000 transactions for requested period. Certain chunks of data of 300k transactions are used to test various models in this project.

A set of data has been analyzed by domain expert and marked real manipulation scenarios, which is effectively used to train the system. sensitivity parameters has been adjusted such that to identify known cases with low error rate with the use of above data set.

	A	B	C	D	E	F	G	H	I	J	K	L
1	T01_SYMBOL	T01_SIDE	T01_ORDE	T01_ORDE	T01_ORDS	T01_TRAN	T01_AVGF	T01_ORDV	T01_EXG_	T01_MUBA	T01_EXTERNAL_	
2	1214	2	2	4000	2	20140713-	82.5	330000	59.4	11700539	200219	
3	1214	2	2	4000	2	20140713-	82	328000	59.04	11700539	200219	
4	7010	2	2	3700	4	20140713-	0	270100	48.62	11700539	200219	
5	2330	2	2	6850	2	20140713-	49.21136	330170	60.68	11180976	108251	
6	1810	2	2	3000	2	20140713-	125.75	377250	67.9	11659493	1.01E+09	
7	4220	1	2	113445	2	20140713-	16	1815120	326.72	11180976	108251	
8	1810	1	2	1030	2	20140713-	126	130295	23.36	11180976	108251	
9	1810	2	2	3000	2	20140713-	126	378000	68.04	11659493	1.01E+09	
10	4260	1	2	1746	2	20140713-	72.75	127458	22.86	11180976	108251	
11	7010	1	2	17950	2	20140713-	71	1274450	229.4	11180976	108251	
12	4002	1	2	490	f	20140713-	0	52920	9.53	11180976	108251	
13	1010	1	2	1	8		0	17	0	11659493	1.01E+09	
14	4260	1	2	3000	C	20140713-	0	214500	38.61	11796181	1001897	
15	1810	1	2	2000	C	20140713-	0	249000	44.82	11796181	1001897	
16	1810	1	2	3500	C	20140713-	0	434875	78.28	11796181	1001897	
17	1810	1	2	7000	C	20140713-	0	868000	156.24	11796181	1001897	
18	4260	1	2	7500	C	20140713-	0	534375	96.19	11796181	1001897	
19	4260	1	2	12500	C	20140713-	0	887500	159.75	11796181	1001897	
20	7010	2	2	3700	m	20140713-	0	270100	48.62	11700539	200219	

Figure B.1: Transaction Data

Other transaction data is used to test the system. testing has been carried out with data sets of different companies in various models stated earlier.

B.2 Naive Bayes Classifier

Naive Bayes Classifier has been proved that this is the best classifier suitable for this solution.Extracted features are arranged suitable for Naive Bayes classifier inputs and initialize the classifier. After that arranged input data of company 'Saudi Industrial Investment Group' is fed to system.

```
var classifier = new bayes.NaiveBayes({
            columns: trainingColumns,
            data: arrayTraining,
            verbose: true
        });

        classifier.train();
```

```
        this.testDataBayes(classifier);
```

Below table denotes the classifier output results of Naive Bayes Classifier. Please note that results are not normalized with customer data.

Measurement	Value
Accuracy	88.81%
Recall	100%
Precision	38.18%
F1 Score	55.26

Table B.1: Naive Bayes Classifier results

B.3 System Functionality

System is trained and tested for company data allocating one classifier for each company data has given the best results. Sensitivity is also set depending only on this company data.

Company symbol '2250' (company Name is 'Saudi Industrial Investment Group') which is listed in Saudi Stock Exchange has been selected for testing.

After cleaning and pre-processing data, 914 valid transaction records were there in data model, which is divided nearly 2 : 1 ratio for training and testing data.

After that features are extracted and training the system. Classifier is selected as Naive Bayes classifier for this model.

Feature extracting code sample is shown below.

```
    integrateVectorValues: function (contentData) {
        var dataContent = contentData;
        var priceKey = 'T01_AVGPX';
```

Company	Price	Quantity	Tr. Time	Status	Commission
2250	24.75	1,200	20141130-08:37:23	2	5.35
2250	24.60	2,000	20141130-08:32:37	2	8.86
2250	24.50	2,000	20141130-08:32:21	2	8.82
2250	24.60	2,000	20141130-08:26:49	2	8.86
2250	24.50	2,000	20141130-08:23:25	2	8.82
2250	24.40	1,000	20141130-08:19:33	2	4.39
2250	24.35	1,500	20141130-08:18:59	2	6.57
2250	24.30	2,000	20141130-08:14:35	2	8.75
2250	24.25	1,000	20141130-08:14:00	2	4.36
2250	24.30	2,000	20141130-08:10:16	2	8.75
2250	24.30	1,000	20141130-08:06:58	2	4.37
2250	24.15	500	20141130-08:05:13	2	2.17
2250	24.10	1,000	20141130-08:05:11	2	4.34
2250	24.15	500	20141130-08:04:05	2	2.17
2250	24.05	3,000	20141130-08:00:01	2	12.99
2250	27.00	2,000	20141127-10:45:42	2	9.72
2250	27.00	4,000	20141127-10:45:20	2	19.44

Figure B.2: Transactions of Company Symbol 2250

```
        var volumeKey = 'T01_ORDERQTY';

        var priceAverage = this.getAverageValue(dataContent, priceKey,
        var volumeAverage = this.getAverageValue(dataContent, volumeKey

        Ember.$.each(dataContent, function (key, trans) {
            var price = trans[priceKey];
            var volume = trans[volumeKey];

            if (price > 1 && volume > 0) {  // Check for valid price/vo
                trans.priceChg = Math.abs(price - priceAverage);
                trans.volumeChg = Math.abs(volume - volumeAverage);
            }
        });

        return dataContent;
    },
```

Customer behavior analyzer code sample is shown below.

```
     // This method creates customer map with fraud weight
    createCustomerMap: function () {
        var dataContent = this.get('trainContent');
        var customerMap = this.get('customerMap');
        var customerKey = 'T01_EXTERNAL_REFNO\r';

        Ember.$.each(dataContent, function (key, trans) {
            var customer = trans[customerKey];

            if (customer) {
                if (!customerMap[customer]) {
                    customerMap[customer] = 1;  // 1 based numbering
                } else {
```

```
                customerMap[customer] = customerMap[customer] + tra
            }
        }
    });

    this.set('customerMap', customerMap);
},
```

Normalizer code sample is shown below.

```
 // This method adds a bias depend on customer behavior to predicti
normalizePrediction: function (transaction, predict) {
    var anomalyIndex = predict['1'] / predict['0'];
    var biasedAnomalyIndex = 0;
    var customerSensitivity = this.get('customerSensitivity');
    var customerMap = this.get('customerMap');
    var customerKey = 'T01_EXTERNAL_REFNO\r';
    var customerAnomalyIndex = customerMap[transaction[customerKey]

    if (customerAnomalyIndex && anomalyIndex < 1) {     // If it is
        biasedAnomalyIndex = customerSensitivity * customerAnomalyI

        if (biasedAnomalyIndex > 1) {
            predict.answer = '1';
        }
    }

    return predict;
},
```

After classification by price, volume detector accuracy, precision, recall and F1 score has been calculated. Then those abnormal probability values are normalized using customer weights.

Results of the system before normalization and after normalization are stated below.

Measurement	Value
Accuracy	88.81%
Recall	100%
Precision	38.18%
F1 Score	55.26

Table B.2: Single Company Results without Normalize

Below is company results after normalization.

Measurement	Value
Accuracy	96.38%
Recall	100%
Precision	65.63%
F1 Score	79.25

Table B.3: Single Company Results with Normalize