Islamic University of Gaza

Deanery of higher Studies

Information Technology Program

Department of Computer Science

A Data Mining Based Fraud Detection Model for Water Consumption Billing System in MOG

Prepared By

Eyad Hashem S.Humaid

120092860

Supervised By

Dr. Tawfiq S. Barhoom

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science In Information Technology

2012-1433H

A DATA MINING BASED FRAUD DETECTION MODEL FOR WATER CONSUMPTION BILLING SYSTEM IN MOG

ABSTRACT

Financial losses due to financial frauds are mounting, recognizing the problem of losses and the area of suspicious behavior is the challenge of fraud detection. Applying data mining techniques on financial statements can help in pointing out the fraudulent usage. It is important to understand the underlying business objectives to apply data mining objectives.

Water consumer dishonesty is a problem faced by all water and power utilities that managed by a financial billing system worldwide. Finding efficient measurements for detecting fraudulent electricity consumption has been an active research area in recent years. This thesis presents a new model towards Non-Technical Loss (NTL) detection in water consumption utility using data mining techniques.

This work applies a suitable data mining technique in this field based on the financial billing system for water consumption in Gaza city. Selected technique used in developing a fraud detection model. The efficiency and accuracy of the model were tested and evaluated by one scientific method and reached one accepted technique.

The intelligent model developed in this research study predicts and select suspicious customers to be inspected on-site by the department of water theft combat (DWTC) teams at the municipality of Gaza (MOG) for detection of fraud activities.

The model increases the detection hit rate of 1-10 % random manual detection to 80% intelligent detection.

This approach provides a method of data mining, which involves feature selection and extraction from historical customer's water consumption data. The Support Vector Classification technique (SVC) applied in this research study uses customer's load profile information in order to expose abnormal customer’s load profile behavior.

نموذج لكشف الاحتيال في استهلاك المياه في مدينة غزة بالاعتماد على تقنية تنقيب البيانات ونظام الفاتورة المحوسبة

الخسارة المالية الناتجة عن عمليات الاحتيال المالي في تزايد مستمر. معرفة المشكلة التي تسبب تلك الخسارة ومعرفة المنطقة أو مجموعة الأشخاص المشكوك في أمرهم يعتبر تحدي في علم كشف الاحتيال. تطبيق تقنية تنقيب البيانات على البيانات الناتجة عن الأنظمة المالية يساعد في عملية الكشف عن عمليات الاحتيال أو المحتالين .إنه من المهم دراسة و فهم آلية و أهداف العمل في الدائرة قيد الدراسة للتوصل إلى أهداف تقنية البيانات في كشف الاحتيال.

ممارسة التضليل في عملية استهلاك المياه والكهرباء هي مشكلة متواجدة لدى جميع المؤسسات المسئولة عن عملية توزيع المياه والطاقة والتي تدار من قبل أنظمة الفواتير المالية المحسوبة على مستوى العالم.الأبحاث العلمية الخاصة بإيجاد أدوات للقياس للكشف عن عمليات الاحتيال في الاستهلاك المرتبط بعداد الكتروني أو ميكانيكي لحساب كمية الاستهلاك وخاصة الكهرباء كانت نشطة في السنوات الأخيرة.

يقدم هذا البحث نهجا جديدا نحو الخسارة الغير فنية (NTLs) للمساعدة في الكشف عن الذي يمارسون الاحتيال في سحب كميات استهلاك المياه نظرا لوجود تشابه كبير بين طريقة استهلاك المياه والكهرباء.

هذا العمل يهدف إلى إيجاد طريقة مناسبة من طرق تقنية تنقيب البيانات المتخصصة في هذا المجال وبالتعامل مع نظام للفوترة المالية الخاص بإدارة حسابات مشتركين المياه لمدينة غزة والمتواجد في بلدية غزة.

ولذا تم تصميم نموذج لكشف الاحتيال باستخدام الطريقة المختارة وتم إخضاعه لإحدى عمليات التقييم حتى تم التوصل إلى مستوى أداء ودقة مقبولين.

هذا النموذج قادرا على تحديد مجموعة المشتركين المشبوهة وهذا بدوره عند تطبيقه كنظام متكامل سيساعد فريق الكشف عن سرقات المياه في بلدية غزة (DWTC) في عمليات التفتيش الميدانية حيث سيزيد معدل الكشف من 5% إلى 80%.

هذا البحث يتضمن طريقة من طرق تنقيب البيانات تشمل اختيار واستخراج الملامح أو العناصر ذات الصلة من البيانات التاريخية الخاصة باستهلاك المياه للمشتركين مع تطبيق المصنف SVM على الملفات الخاصة بأحمال أو استهلاكيات المشتركين.

Acknowledgements

First and foremost, I wish to thank Allah for giving me strength and courage to complete this thesis and research, and also to those who have assisted and inspired me throughout this research.

There are so many people to whom I am indebted for their assistance during my endeavors to complete my Masters candidate in information technology at the Islamic University. First and foremost, I would like to express my gratitude to my supervisor Dr. Tawfig Barhoom whose invaluable guidance and support was very helpful throughout my research. Gratitude to Information technology teachers staff, Dr. Alaa'a AL-Halees, Dr. Rebhi Baraka, Dr. Nabeel Hweehy and Dr. Ashraf AL-Attar. It is unlikely that I would have reached completion without their encouragement and support. Besides that, I would like to thank the Non-Technical Loss (NTL) team at the department of water theft combat(DWTC) at the municipality of Gaza (MOG) for providing me a valuable data.

My appreciation also goes to the MOG for providing me the customer data, and other helpful information about business logic and functionality. I express my appreciation to everyone who has involved directly and indirectly to the success of this research.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION

ANN	Artificial Neural Network.
CRISP-DM	Cross-Industry Standard Process for Data Mining.
CV	Cross Validation
CWBD	Customers Water Breaches Data.
DWTC	Department of Water Theft Combat.
ECIBS	Electronic Customers Information Billing System.
FDD	Financial Fraud Detection.
IEEE	Institute of Electrical and Electronics Engineers .
KKT	Karush-Kuhn Tucker.
KNN	k-Nearest Neighbor Algorithm.
LIBSVM	Library of Support Vector Machine .
M3	Cubic Meter.
MMH	Maximum Marginal Hyperplane.
MOG	Municipality Of Gaza.
Monthly_DS	Monthly Data Set.
NTLs	Non-Technical Losses.
RAM	Random Access Memory.
RDBMS	Relational Database management System
ROC	Receiver operating characteristic.
Seasonaly_DS	Seasonally Data Set.
SNN	Simulated Neural Network
SOM	Self-Organizing Map.
SQL	Structured Query Language.
SVC	Support Vector Classification.
SVMs	Support Vector Machines.
TOW	Theft of Water.
U.K	United Kingdom
U.S.A	United States of America.

USD	United States Dollar.
Yearly_DS	Yearly Data Set.

INTRODUCTION

Water theft is a big problem in distributing water fairly to Gaza citizens and decrease the revenue of the organization. The MOG that responsible for water delivering to all Gaza citizens is a non-profit foundation, but it is important to balance the expenses with the income to allow delivering water service to all citizens fairly. "In order to achieve revenue improvement, it is essential to measure the energy consumed accurately" [46], so fraudulent water consumption resulting in an incorrect consumption amount. According to the MOG, Gaza city has about 8000 buildings getting water without the water meter that used by the MOG to calculate the monthly consumption for each customer. During the past 5 years more than 3000 water breach cases have been recorded due to meter tampering, meter malfunction, illegal connections, billing irregularities. These irregularities are internationally known as non-technical losses [16][49].

Manual investigation to detect customer's irregularities and metered water consumption theft that done by DWTC at MOG is hard, slow, costly and performed randomly, so increasing the speed and accuracy of investigation is important to detect the suspicious customer consumption and decrease the cost.

These are several groups of researchers devote a significant amount of effort in studying fraud detection by many ways in various domains. A lot of them applied data mining techniques successfully to solve the problem, but each work has some cons against pros as explained in chapter 3. Therefore, by focusing on the related works' cons and by avoiding their pros, the research proposes a fraud detection model for customer's water consumption profile using classification techniques. The research applies the SVM method and compares it with other methods such as ANN and KNN. The method SVM reached the highest performance and accuracy score in the underlying business fraud detection. The model improves the manual hit rate detection from 1-10% random detection to 80% intelligent detection based on historical consumption data managed by a computerized billing system.

1.1 Non-technical Losses

The mentioned Irregularities known as non-technical losses (NTLs). NTLs originating from electricity theft and other customer malfeasances are a problem in the electricity supply industry. [11][18] NTL is a problem in water supply industry too because of the similarity between water and electricity distribution systems in depending on meter technology and load profiling concept.

NTLs include the following activities: [16][49].

1) Losses due to faulty meters and equipment.
2) Tampering with meters so that meters record low rates of consumption.
3) Stealing by bypassing the meter or otherwise making illegal connections.
4) Arranging false readings by bribing meter readers.
5) Arranging billing irregularities with the help of internal employees by means of such subterfuges as making out lower bills, adjusting the decimal point position on the bills, or just ignoring unpaid bills.
6) Poor revenue collection techniques.

1.2 Load Profiling

Load profiling is the load shape that to explain the daily and seasonal variations in load as responses to the time of day, the time of a year, by the type of day or season of the year. In this way, it represents habits and usual weather conditions and known or observable customer information [11][50].Thus, the shape of load profiles is influenced by the customer's type of activity, and on the other hand behavior of a customer.[10]. In this thesis three customer load profiling types (monthly, seasonally and yearly) extracted from historical water consuming database and tested by the proposed classification technique. The research will present a methodology to classify customer's load profiles by their form of water consumption. The goal is to determine a customer with typical normal load profile and the customer with the representative fraud load profile. The customer water consumption (load profiling) are recorded in a computerized financial billing system which calculates and issues the water invoices. In the present system, an employee of a utility goes to customer premises to read the meter, records it in a book and then transfers the reading into the computer system and generates a bill. The bill is served to the consumer by courier. The consumer pays the bill at the collection centers opened by the utility for receipt of payments.

1.3 Financial System Frauds Overview

Fraud is a serious problem face information system that implemented in various domains. Credit card transactions as a financial system branch had a total loss of 800 million dollars of fraud in U.S.A. and 750 million dollars in U.K. in the year 2004 [1]. In the area of health care according to transparency international [2], the total expenditure exceeds the amount of 3 trillion euro worldwide. That size in the health care industry induces several actors in the field to make a profit by using illegal means, forbidden financial operation committing health care fraud. fiscal frauds have received considerable attention from the public, press, investors, the financial community and regulators because of several high profile frauds reported at large corporations such as Enron, Lucent, and WorldCom over the last few years. Most editors lack the experience necessary to detect it. Statistics and data mining methods have been applied successfully to detect activities such as money laundering, e-commerce credit card fraud, telecommunications fraud, insurance fraud, electricity fraud, etc. [8]. There has been a large body of research and practice focusing on exploring data mining techniques to solve financial problems. The competitive advantages achieved by data mining include increased revenue, reduced cost, and much improved marketplace responsiveness and awareness [9]. Fraud detection by manual analysis is too slow and hard because of huge data records but by data mining techniques can analyze huge data sets and extract the knowledge that helps in fraud detection.

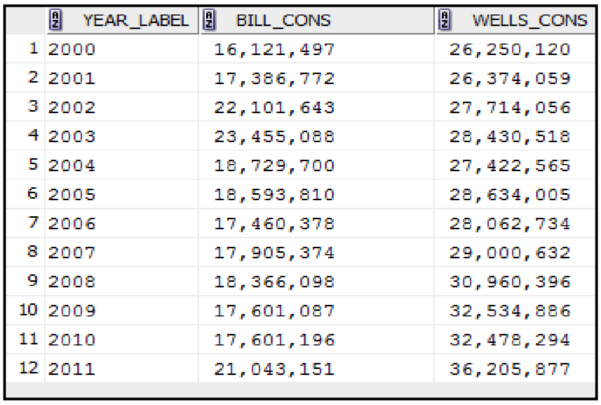
1.4 The Organization Under Study.

The research study focus on MOG financial billing system that manages the cost and consumption quantity of the water delivering to Gaza citizens. Water consumption theft cases that continuously occur, cause cumulative budgetary losses. MOG serves the Gaza city that has about 40,000 customers of water subscriptions among 67 areas divided Gaza city. Manual investigation to detect irregularities and water theft by DWTC that done randomly is costly, hard and slow, so increasing the speed and accuracy of investigation is important to detect the suspicious customer consumption and decrease the cost. All municipalities and organizations that interested in water consumption and responsible for water delivering, concern about that to avoid financial losses and satisfy the other committed customers. There cannot be any business prospects without satisfied customers who remain loyal and develop their relationship with the organization [23][40].

1.5 Research Motivation

The main motivations of this study is to

- Assist the MOG to reduce its NTLs in the water distribution sector.
- Investigate the capability of using data mining classification techniques for the detection and identification of NTL activities [17][22] and to solve the existing drawbacks in [16][17] that deal with 20% of the available customer's load profile data.
- Fraud activities Inspected on-site by MOG DWTC teams manually and randomly, so the intelligent model developed in this research study can predicts suspicious customers to help them in detection of fraud activities.
- According to the new statistics at the department of water distribution at the municipality of Gaza, the financial losses that caused by water consumption emerging from the big difference between water wells productions that located in Gaza city and Gaza water consuming as illustrated in Fig 2. The difference in 2011 reaches 15 million cubic meter of water considered as water loss.



	YEAR_LABEL	BILL_CONS	WELLS_CONS
1	2000	16,121,497	26,250,120
2	2001	17,386,772	26,374,059
3	2002	22,101,643	27,714,056
4	2003	23,455,088	28,430,518
5	2004	18,729,700	27,422,565
6	2005	18,593,810	28,634,005
7	2006	17,460,378	28,062,734
8	2007	17,905,374	29,000,632
9	2008	18,366,098	30,960,396
10	2009	17,601,087	32,534,886
11	2010	17,601,196	32,478,294
12	2011	21,043,151	36,205,877

Fig 1.1. Water wells production vs. water consumptions per year from 2000 to 2011.

1.6. Statement Of Problem.

Manual investigation to detect customer's irregularities, the suspicious customer consumptions and metered water consumption theft by DWTC at MOG to detect is hard, slow, costly and done randomly.

1.7 Research Objectives

There are some objectives for this research as follows.

1.7.1 Main Objective.

Find and apply a rapid intelligent model to detect metered water consuming frauds.

1.7.2 Specific Objectives.

- Review the related works and the mining methods in this field.
- Investigate the most appropriate techniques based on FFD and load profile concept.
- Apply a classification model for water consumption based on one appropriate technique.
- Evaluate the applied model accuracy.
- Selecting the best attributes that improve the accuracy.

1.8. Scope And Limitations.

The idea of this research assumption is limited to any financial billing system that's responsible for managing and calculating water consumption for customers. This research will focus on MOG's historical transactions data and will use SVM classification data mining technique that internationally applied successfully. Water theft can be done by illegal connections without consumption meter or within a metered consumption. The study of this research is limited to the customers with metered water consumption and with buildings have at least one water agreement participation.

Classification methods are the dominant techniques in this field [8]. So in this study three classification techniques were tested and evaluated upon the underlying business dataset structures, which are support vector machines (SVM), artificial neural network (ANN), K-nearest neighbor (KNN).

Thesis research divided into six chapters. Chapter 1 presents general overview, research objectives, problem definition, research motivation and scope and limitation. Chapter 2 presents literature review about thesis literature requirements which include non technical losses, load profiling definitions, data mining methodologies, used classification methods, and the selected performance evaluation method. Chapter 3 introduces and discusses some related work in fraud detection. Chapter 4 presents the model development methodology and underline business data sets. Chapter 5 provides the final evaluated results between the selected classification techniques on all datasets. Chapter 6 presents the conclusion of the thesis research work and what are future demands.

CHAPTER 2

LITERATURE REVIEW

This chapter presents the background and theoretical concepts of the data mining techniques applied in this research study. Start by discussing the importance of data mining techniques in the business domain and explain the concept of supervised and non supervised learning methods. Next clarify the statistics versus data mining, review the methodology of data mining. In the last part of this chapter, the background and theoretical concepts of SVMs, ANN and KNN were presented and explained.

2.1 Data Mining.

The benefits of data mining can be extracted from knowing its definition. Data Mining is the science of extracting useful information from large datasets or database is known as data mining. It is a new discipline, lying at the intersection of statistics, machine learning, data management and databases, pattern recognition, artificial intelligence, and other areas [1]. It converts data into knowledge and actionable information [23]. The modern technologies of computers, networks, and sensors have made data collection and organization an almost effortless task. However, the captured data need to be converted into information and knowledge from recorded data to become useful. Traditionally, analysts have performed the task of extracting useful information from the recorded data, But the increasing volume of data in modern business and science calls for computer-based approaches. As data sets have grown in size and complexity, there has been an inevitable shift away from direct hands-on data analysis toward indirect, automatic data analysis using more complex and sophisticated tools. Data mining is the entire process of applying computer-based methodology, including new techniques for knowledge discovery, from data [2]. Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner [1]. Data mining techniques play a big role in analyzing large databases with millions of records to achieve new unknown patterns. These patterns help with study the customer transactions behavior in many scientific researches as in mine. When we get the suspicious customer patterns using data mining techniques, then we can narrow the circle of investigation to get fraudulent issues very fast according to the manual way that could be impossible.

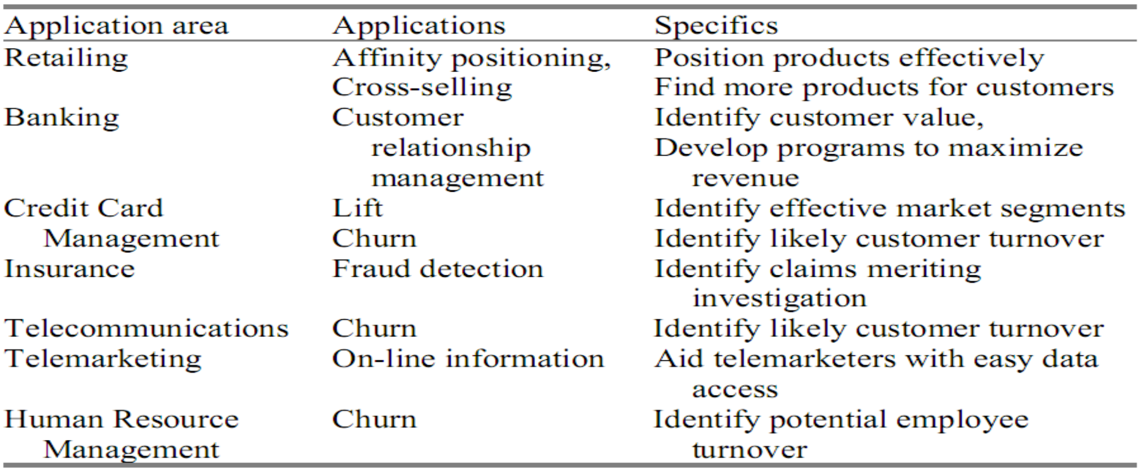
2.1.1 Types of Data Mining.

Data mining techniques divided into two kinds as supervised and non supervised techniques. “In supervised modeling, whether for the prediction of an event or for a continuous numeric outcome, the availability of a training dataset with historical data is required. Models learn from past cases” [24]. Supervised techniques such as classification models depend on a label class must exist in the database under study. This type of mining used in several fields like direct marketing, Credit/loan approval, Medical diagnosis if a tumor is cancerous or benign**,** fraud detection if a transaction is fraudulent, ..Etc. The non-supervised learning such as clustering models, there is no class label. Clustering is a method of grouping data that share similar trend and patterns, applied in Insurance, city-planning to identify groups of houses according to their house type, value, and geographical location,..Etc. Supervised learning techniques is the dominated methods for detecting financial frauds such as classification data mining techniques, which are the most recent used in public organization [27][13]. So that the thesis concentrate on supervised classification learning techniques.

2.1.2 Business Data Mining

Data mining has been very effective in many business venues. The key is to find actionable information, or information that can be utilized in a concrete way to improve profitability. Some of the earliest applications were in retailing, especially in the form of market basket analysis. Table [1] shows the general application areas. Note that they are meant to be representative rather than comprehensive [7].

Table 2.1.Data mining application areas



Application area	Applications	Specifics
Retailing	Affinity positioning, Cross-selling	Position products effectively Find more products for customers
Banking	Customer relationship management	Identify customer value, Develop programs to maximize revenue
Credit Card Management	Lift Churn	Identify effective market segments Identify likely customer turnover
Insurance	Fraud detection	Identify claims meriting investigation
Telecommunications	Churn	Identify likely customer turnover
Telemarketing	On-line information	Aid telemarketers with easy data access
Human Resource Management	Churn	Identify potential employee turnover

2.2 Data Mining Methodology

There is more perspective represented the process steps for data mining (knowledge discovery) methodology. Knowledge discovery as a process is depicted in Figure [1] and consists of an iterative sequence of the following steps:

1. Data cleaning (to remove noise and inconsistent data).
2. Data integration (where multiple data sources may be combined).
3. Data selection (where data relevant to the analysis task are retrieved from the database).
4. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance). Data mining (an essential process where intelligent methods are applied in order to extract data patterns).
5. Pattern evaluation (to identify the truly interesting patterns representing knowledge Based on some interesting measures).
6. Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user) [20].

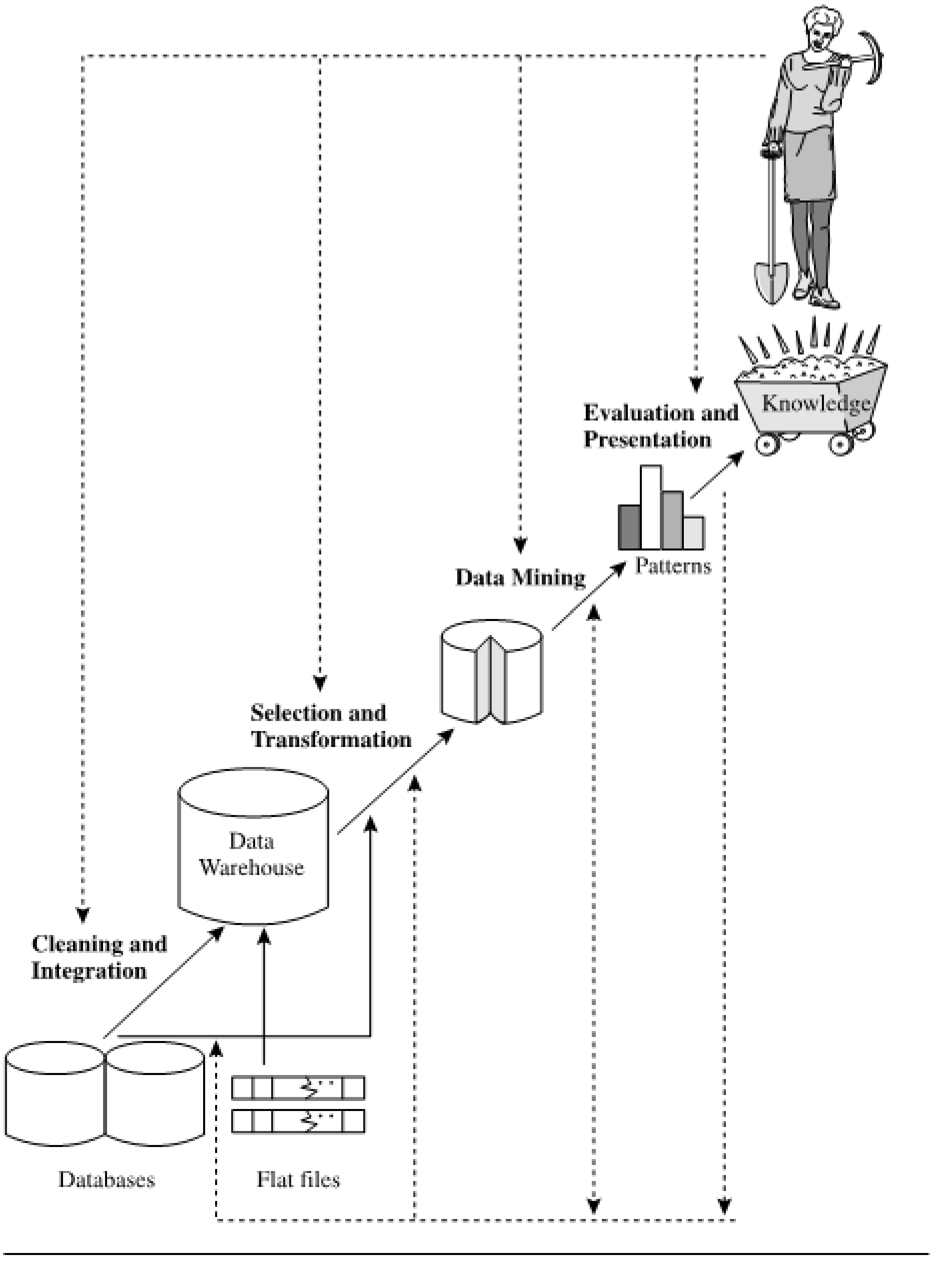


Fig2.1. Data mining process steps. [20]

There is a Cross-Industry Standard Process for Data Mining (CRISP-DM) widely used by industry members. This model consists of six phases in Tended as a cyclical process (see Figure 2): [7][34]

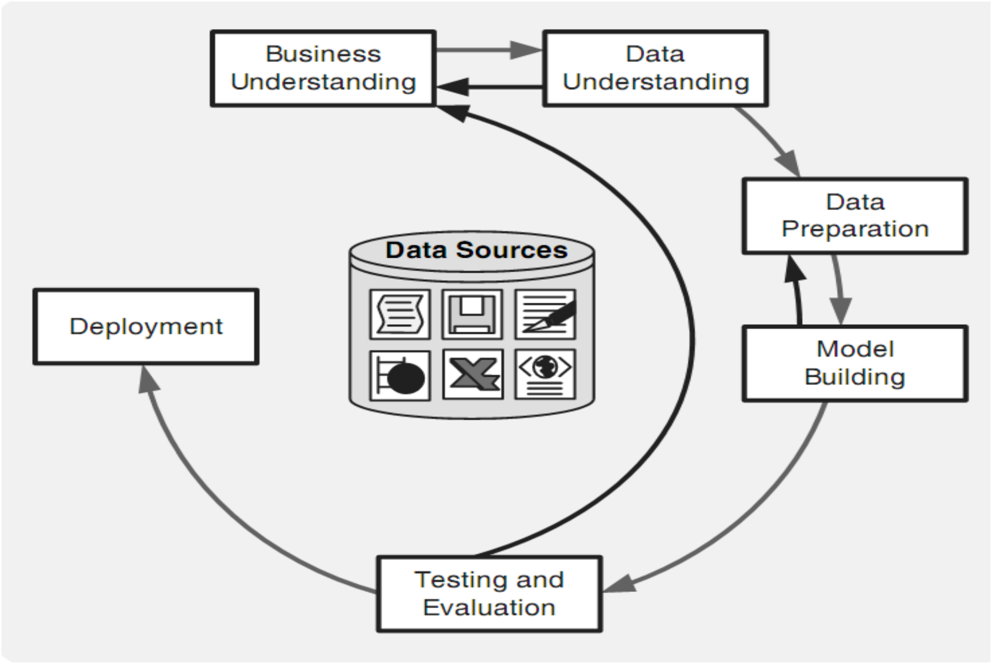


Figure2.2 Phases of CRISP-DM reference mode [34]

In this work, a business intelligent model has been developed, to detect water consumption fraud, based on a specific business structure deal with water consumers using a suitable data mining technique. The model was evaluated by a scientific approach to measure accuracy.

2.3 Classification

Classification models are linear and non-linear, "Linear models are analytical models that assume linear relationships among the coefficients of the variables being studied." [40]. Furthermore, defined as "Approximation of a discriminant function or regression function using a hyperplane. It can be globally optimized using simple techniques, but does not adequately model many real-world problems." [41]. The other type of classification prediction models is the non-linear models in which defined as "Non-Linear Predictive Models are analytical model that does not assume linear relationships among the coefficients of the variables being studied." [40][41]. Linear models in spite of their computational simplicity, stability they have an obvious potential weakness. “The actual process may not be linear, and such an assumption introduces uncorrectable bias into the predictions.” [28]. Data mining applications inherently involve large data sets, and so “the general trend is almost always to use non-linear methods” [28], implying that most data miners feel that their data is closer to the situation in Figure 2.3(b). Although the same SVM technique generated both data sets, the increase in data in Figure 2.3(b) makes the linear model less appealing [28].This section present three selected popular non-linear prediction method to test and evaluate. These methods allow f (x) to take on a more flexible form.

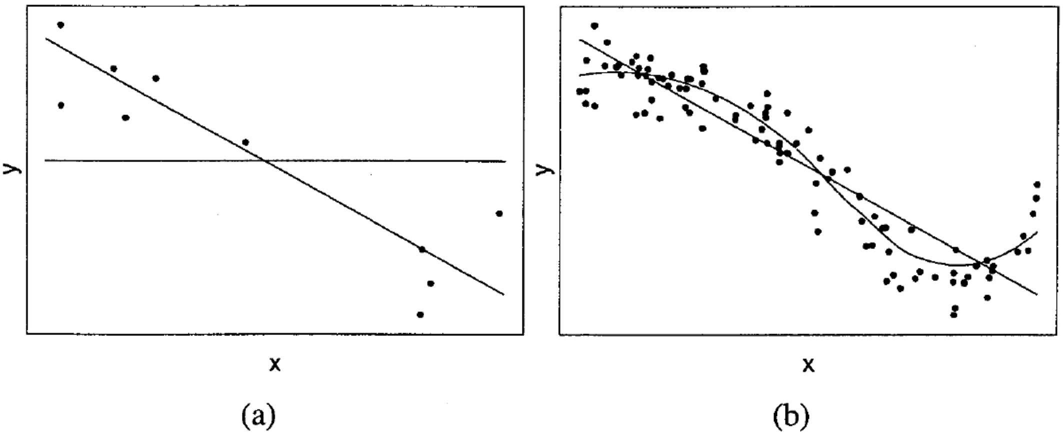


Fig 2.3 utility of linear and nonlinear model [28]

2.3.1 support Vector Machines

The first selected mining technique is SVM that can be used as linear and non linear. "In machine learning, SVMs (SVMs, also support vector networks) is supervised learning model with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. In addition to performing linear classification, SVMs can efficiently perform non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces." [39].

2.3.1.1 The Case When The Data Are Linearly Separable

SVMs are supervised learning methods that generate input-output mapping functions from a set of labeled training data. “The mapping function can be either a classification function (used to categorize the input data) or a regression function (used to estimation of the desired output)” [7]. For classification, “nonlinear kernel functions are often used to transform the input data (inherently representing highly complex nonlinear relationships) to a high dimensional feature space in which the input data becomes more separable (i.e., linearly separable) compared to the original input space. Then, the maximum-margin hyperplanes are constructed to optimally separate the classes in the training data” [7]. Two parallel hyperplanes are constructed on each side of the hyperplanes that separates the data by maximizing the distance between the two parallel hyperplanes. An assumption is made that the larger the margin or distance between these parallel hyperplanes the better the generalization error of the classifier will be [7]. “SVMs can be used for prediction as well as classification. They have been applied to a number of areas, including handwritten digit recognition, object recognition, and speaker identification, as well as benchmark time-series prediction tests” [20]. SVMs have demonstrated highly competitive performance in numerous real-world applications, such as medical diagnosis, Bioinformatics, face recognition, image processing and text mining, which has established SVMs as one of the most popular, state-of-the-art tools for knowledge discovery and data mining. Similar to artificial neural networks, SVMs possess the well-known ability of being universal approximators of any multivariate function to any desired degree of accuracy. Therefore, they are of particular interest to modeling highly nonlinear, complex systems and processes. [7].

To understand the case when the data are linearly separable to explain the mystery of SVMs, let’s first look at the simplest case, a two-class problem where the classes are linearly separable. Let the data set D be given as (X1, y1),(X2, y2), …., (X|D|, y|D|), where Xi is the set of training tuples with associated class labels, yi. Each yi can take one of two values, either +1 or -1 (i.e., yi $\in \{+1, -1\}$), to aid in visualization, let’s consider an example based on two input attributes, A1 and A2, as shown in Figure 2.4. From the graph, we see that the 2-D data are linearly separable because a straight line can be drawn to separate all of the tuples of class +1 from all of the tuples of class -1. There is an infinite number of separating lines that could be drawn. To find the best one, that is, one that will have the minimum classification error on previously unseen tuples. To find this best line, note that if the data were 3-D (i.e., with three attributes), the goal is to find the best separating plane. Generalizing to n dimensions, the goal is to find the best hyperplane. The term “hyperplane” will be used to refer to the decision boundary that has to be reached, regardless of the number of input attributes. So, in other words, how to find the best hyperplane? An SVM approaches this problem by searching for the maximum marginal hyperplane [20]. Consider Figure 2.4, which shows two possible separating hyperplanes and

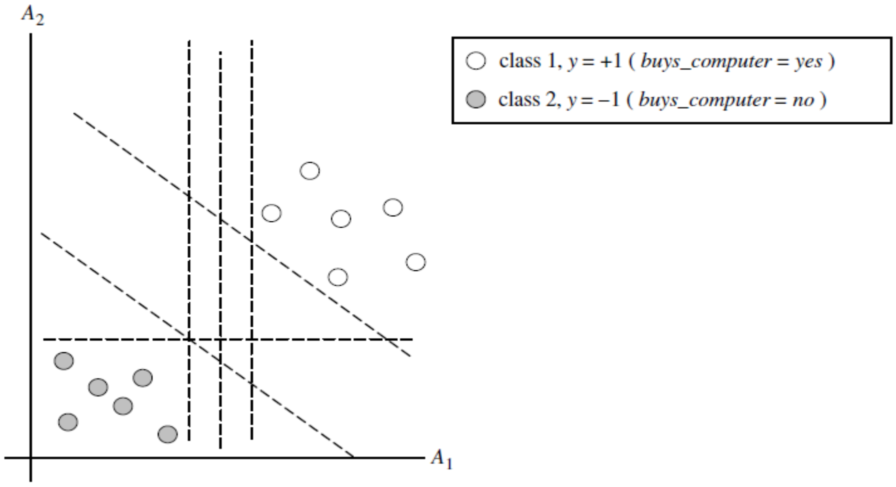

A_2
A_1
○ class 1, $y = +1$ (*buys_computer* = *yes*)
● class 2, $y = -1$ (*buys_computer* = *no*)

Fig 2.4. The 2-D training data are linearly separable. There are an infinite number of (possible) Separating hyperplanes or “decision boundaries.” [20]

their associated margins. "Before getting into the definition of margins, let’s take an intuitive look at figure 2.5. Both hyperplanes can correctly classify all of the given data tuples. Intuitively, however, we expect the hyperplane with the largest margin to be more accurate at classifying future data tuples than the hyperplane with the smaller margin. This is why (during the learning or training phase); the SVM searches for the hyperplane with the largest margin; that is, the maximum marginal hyperplane (MMH)" [20]. The associated margin gives the largest separation between classes. "Getting to an informal definition of margin, the shortest distance from a hyperplane to one side of its margin is equal to the shortest distance from the hyperplane to the other side of its margin” [20], where the sides of the margin are parallel to the hyperplane. A separating hyperplane can be written as

$$W . X + b = 0 \qquad (2.1)$$

Where W is a weight vector, namely, W = {w_1, w_2, …… , w_n}; n is the number of attributes and b is a scalar, often referred to as a bias. To aid in visualization, let’s consider two input attributes, A1 and A2, as in Figure 2.5(b). Training tuples are 2-D, e.g., X = (x1, x2), where x1 and x2 are the values of attributes A1 and A2, respectively, for X. By considering b as an additional weight, w0, the above separating hyperplane can be rewritten as

$$w_0 + w_1 x_1 + w_2 x_2 = 0. \qquad (2.2)$$

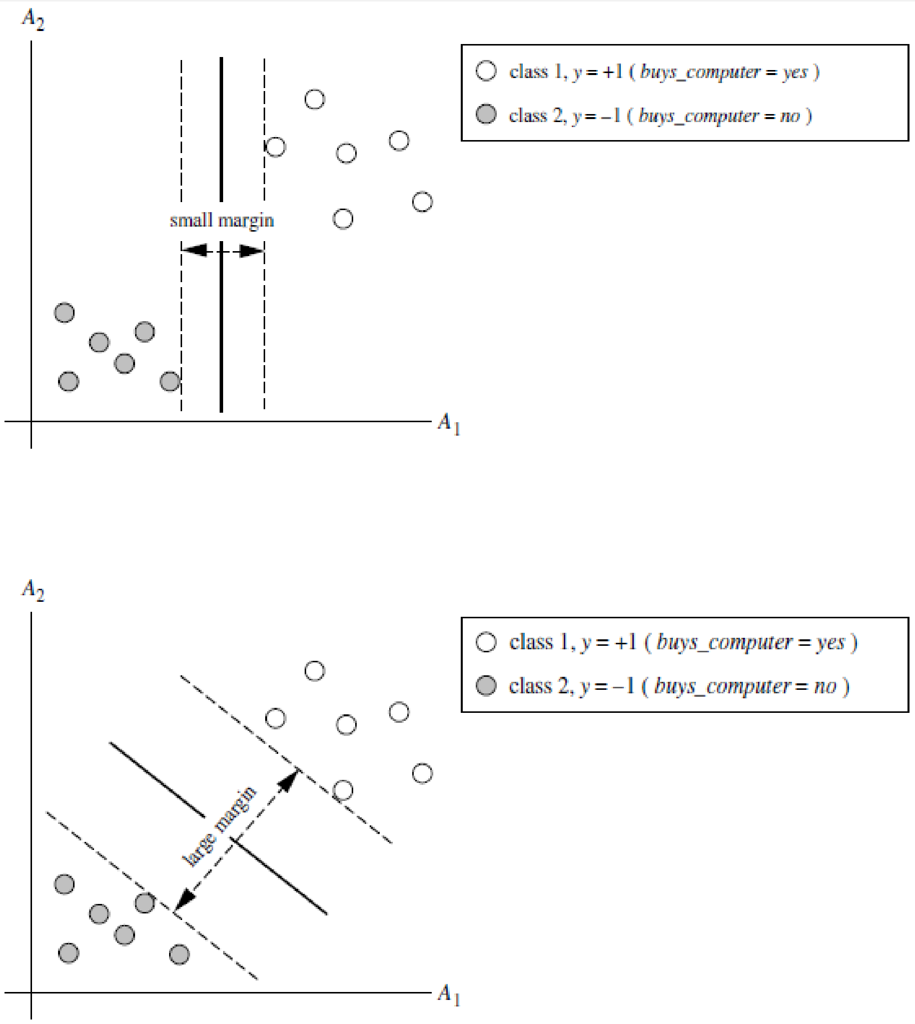

A_2
class 1, $y = +1$ (*buys_computer = yes*)
class 2, $y = -1$ (*buys_computer = no*)
small margin
A_1
A_2
class 1, $y = +1$ (*buys_computer = yes*)
class 2, $y = -1$ (*buys_computer = no*)
large margin
A_1

Fig 2.5 Two possible separating hyperplanes and their associated margins [20].

The weights can be adjusted so that the hyperplanes defining the sides of the margin can be written as

H1 : $w0+w1x1+w2x2 \geq 1$ for $yi = +1$, and (2.3)

H2 : $w0+w1x1+w2x2 \leq -1$ for $yi = -1$. (2.4)

That is any tuple that falls on or above H1 belongs to class +1, and any tuple that falls on or below H2 belongs to class -1. Combining the two inequalities of equations (2.3) and (2.4), we get yi(w0 + w1x1 + w2x2) ≥ 1, $\forall i$ (2.5)

Any training tuples that fall on hyperplanes H1 or H2 (i.e., the sides defining the margin) satisfy equation (2.5) and are called support vectors. That is, they are equally close to the (separating) MMH. Essentially, the support vectors are the most difficult tuples to classify and give the most information regarding classification. From the last formulate, a new formulate for the size of the maximal margin can be obtained. The distance from the separating hyperplane to any point on H1 is $\frac{1}{||W||}$, where ||W|| is the Euclidean norm of W, that is $\sqrt{W \ \ .W^9}$. By definition, this is equal to the distance from any point on H2 to the separating hyperplane. Therefore, the maximal margin is$\frac{2}{||W||}$ .So, how does an SVM find the MMH and the support vectors? Using some fancy math tricks, Equation (2.5) can be rewritten as

If W = {w1,w2,……,w_n} then $\sqrt{W\,.W}$ = $\sqrt{w1^2 + \ w2^2 +, \dots \dots ., +wn^2}$.

(convex) quadratic optimization problem. Such fancy math tricks are beyond the scope of this thesis. Advanced readers may be interested to note that the tricks involve rewriting equation (2.5) using a Lagrangian formulation and then solving for the solution using Karush-Kuhn Tucker (KKT) conditions. Once the support vectors and MMH have being found, a trained support vector machine is reached. The MMH is a linear class boundary, and so the corresponding SVM can be used to classify linearly separable data. Such a trained SVM referred as a linear SVM. Once a trained support vector machine has been got, how to use it to classify test? (new tuple). Based on the Lagrangian formulation mentioned above, the MMH can be rewritten as the decision boundary

d(X^T) = $\sum_{i=1}^{l} y_i \propto_i X_i X^T + \ b_0$. (2.6).

where yi is the class label of support vector Xi; X is a test tuple; αi and b0 are numeric parameters that were determined automatically by the optimization or SVM algorithm and l is the number of support vectors. [20].

For linearly separable data, "the support vectors are a subset of the actual training tuples (although there will be a slight twist regarding this when dealing with nonlinearly separable data, as we shall see)" [20]. Given a test tuple, X^T, plug it into Equation (2.6), and then check to see the sign of the result. This tells us on which side of the hyperplane the test tuple falls. If the sign is positive, then X^T falls on or above the MMH, and so the SVM predicts that X^T belongs to class +1 (such representing buys computer=yes). If the sign is negative, then X^T falls on or below the MMH and the class prediction is -1 (representing buys computer=no). Notice that the Lagrangian formulation of the problem (Equation(2.6)) contains a dot product between support vector Xi and test tuple X^T. This will prove very useful for finding the MMH and support vectors for the case when the given data are nonlinearly separable, as described in section 2.4.1.2. Before moving on to the nonlinear case, there are two more important things to note. The support vectors are the essential or critical training tuples, they lie closest to the decision boundary (MMH). If all other training tuples were removed and training was repeated, the same separating hyperplane would be found. Furthermore, the number of support vectors found can be used to compute an (upper) bound on the expected error rate of the SVM classifier, which is independent of the data dimensionality. An SVM with a small number of support vectors can have good generalization, even when the dimensionality of the data is high.

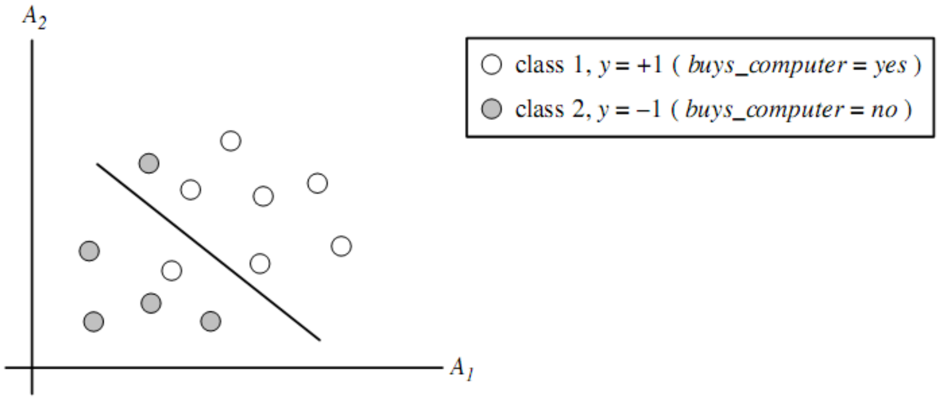

A_2
A_1
○ class 1, $y = +1$ (*buys_computer = yes*)
● class 2, $y = -1$ (*buys_computer = no*)

Fig2.6 A simple 2-D case showing linearly inseparable data. Unlike the linear separable data of Figure 2.3, here it is not possible to draw a straight line to separate the classes. Instead, the decision boundary is nonlinear. [20]

2.3.1.2The Case When The Data Are Linearly Inseparable

Section 2.6.1.1 talk about linear SVMs for classifying linearly separable data, but what if the data are not linearly separable, as in Figure 2.6. In such cases, no straight line can be found that would separate the classes. The linear SVMs would not be able to find a feasible solution here. "Now what the good news is that the approach described for linear SVMs can be extended to create non-linear SVMs for the classification of linearly inseparable data (also called non-linearly separable data, or nonlinear data, for short). Such SVMs are capable of finding non-linear decision boundaries (i.e., nonlinear hyper surfaces) in input space." So, the question is how to extend the linear approach?.Obtaining a nonlinear SVM can be done by extending the approach for linear SVMs as follows. There are two main steps. In the first step, transform the original input data into a higher dimensional space using a non-linear mapping. Several common nonlinear mappings can be used in this step are beyond the scope of this thesis. Once the data have been transformed into the new higher space, the second step searches for a linear separating hyperplane in the new space. We again end up with a quadratic optimization problem that can be solved using the linear SVM formulation. The maximal marginal hyperplane found in the new space corresponds to a nonlinear separating hyper surface in the original space [20]. The complexity of the learned classifier is characterized by the number of support vectors rather than the dimensionality of the data. Hence, SVMs tend to be less prone to overfitting than some other methods [20]. "The concept of overfitting is very important in data mining. It refers to the situation in which the induction algorithm generates a classifier which perfectly fit the training data but has lost the capabilities of generalizing to instances not presented during training. In other word, instead of learning, the classifier just memorizes the training instances". [29].

The researchers in [16] select the non-linear SVM method to deal with electricity metered customer's fraud detection to predict fraudulent customers and customer with anomaly electricity consumption (load profile). The model increases the fraud detection hit rate of 3% achieved by fraud detection team to 60% achieved by SVM model and just of 42% in [17]. SVM works as black box without extract any rule about fraudulent consumption but in this research the goal is to predict if the consumption profile has fraud or not. So SVM is one of the selected classification methods to analyze the thesis’ underlying business datasets.

2.3.2 Neural Network

The second selected mining technique is ANN which is applicable when working with the non-linear problems as done in [44]. This technique used successfully in business fraud detection, and defined as "artificial neural networks (ANNs) are widely used for fraud detection" [30], "a neural network (NN), in the case of artificial neurons called artificial neural network (ANN) or simulated neural network (SNN), is an interconnected group of natural or artificial neurons that uses a mathematical or computational model for information processing based on a connectionistic approach to computation. In most cases, an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network. In more practical terms, neural networks are non-linear statistical data modeling or decision making tools" [36]. Furthermore, "The inspiration for neural networks was the recognition that complex learning systems in animal brains consisted of closely interconnected sets of neurons. Although a particular neuron may be relatively simple in structure, dense networks of interconnected neurons could perform complicated learning tasks such as classification and pattern recognition" [6]. The topologies of neural networks or neural network architectures formed by organizing nodes into layers and linking these layers of neurons with modifiable weighted interconnections. In recent years, neural network researchers have incorporated methods from statistics and numerical analysis into their networks. Being a nonlinear mapping relation from the input space to output space, neural networks can learn from the given cases and summarize the internal principles of data even without knowing the potential data principles ahead. It can adapt its own behavior to the new environment with the results of formation of general capability of evolution from present situation to the new environment [6].

From the aspect of the pure theory, the nonlinear neural networks method is superior to the statistical methods in the application for credit card fraud detection such as in [1]. It is sometime unusual in the practice research even though the common advantages of the neural networks as a possible result of usage of improper network structure and learning computing method. On the other hand, there are still many disadvantages of the neural networks, such as the difficulty to confirm the structure, the efficiency of training, excessive training, and so on. The researchers in

[1] conduct a comparison between three classification methods were tested for their applicability in fraud detection, i.e. decision tree, neural networks and logistic regression. The three methods are compared in terms of their predictive accuracy. Neural network classifier has the best accuracy in testing results.

2.3.3 K-Nearest-Neighbor

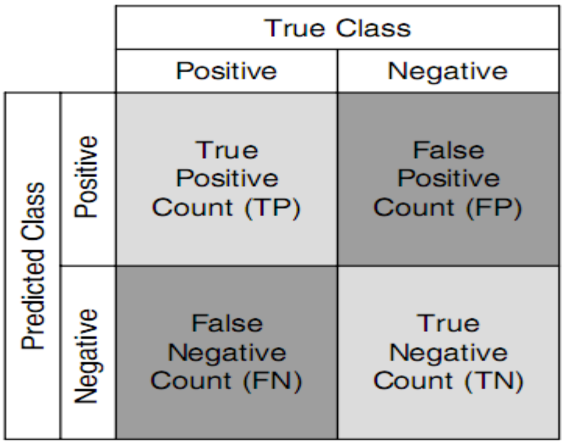
The last selected mining method is KNN, which also supports non-linear problem. This section reviews some point about KNN. "The k-nearest-neighbor method was first described in the early 1950s. The method is labor intensive when given large training sets, and did not gain popularity until the 1960s when increased computing power became available. It has since been widely used in the area of pattern recognition" [7]. "In pattern recognition, the k-nearest neighbor algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space. k-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally, and all computations is deferred until classification. The k-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of its nearest neighbor." [38]. The Nearest-neighbor classifiers are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it. The training tuples are described by n attribute. Each tuple represents a point in a n-dimensional space. In this way, all of the training tuples are stored in a n-dimensional pattern space. When given an unknown tuple, a k-nearest-neighbor classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. These k training tuples are the k nearest neighbors of the unknown tuple. “There are many methods for determining whether two observations are similar. For example, the Euclidean or the Jaccard distance. Prior to calculating the similarity, it is important to normalize the variables to a common range so that no variables are considered to be more important” [43]. “Closeness” is defined in terms of a distance metric, such as euclidean distance. The euclidean distance between two points or tuples, say, X1 = (x11, x12, , , x1n) and X2 = (x21, x22, , , , x2n), is

$$\text{dist}(X1, X2) = \sqrt{\sum_{i=1}^{n}(x_{1i} \quad - x_{2i})^2} \qquad (2.7)$$

In other words, for each numeric attribute, we take the difference between the corresponding values of that attribute in tuple X1 and in tuple X2, square this difference, and accumulate it. The square root is taken of the total accumulated distance count. How to determine a good value for k, the number of neighbors?. This can be determined experimentally. Starting with k = 1, we use a test set to estimate the error rate of the classifier. This process can be repeated each time by incrementing k to allow for one more neighbor as done in with the datasets of this research. The k value that gives the minimum error rate may be selected. In general, the greater the number of training tuples is, the greater the value of k will be (so that classification and prediction decisions can be based on a larger portion of the stored tuples). Nearest-neighbor classifiers use distance-based comparisons that intrinsically assign equal weight to each attribute [7].

2.4 Measuring Performance

After model building, knowing the power of model prediction on a new instance, is very important issue. Once a predictive model is developed using the historical data, one would be curious as to how the model will perform on the data that it has not seen during the model building process. One might even try multiple model types for the same prediction problem, and then, would like to know which model is the one to use for the real-world decision making situation, simply by comparing them on their prediction performance (e.g., accuracy). To measure the performance of a predictor, there are commonly used performance metrics, such as accuracy, recall..etc. First, the most commonly used performance metrics will be described, then some famous estimation methodologies are explained and compared to each other. "Performance Metrics for Predictive Modeling In classification problems, the primary source of performance measurements is a coincidence matrix (classification matrix or a contingency table)" [7]. Figure 2.7 shows a coincidence matrix for a two-class classification problem. The equations of the most commonly used metrics that can be calculated from the coincidence matrix is also given in Fig 2.7.



		True Class	
		Positive	Negative
Predicted Class	Positive	True Positive Count (TP)	False Positive Count (FP)
	Negative	False Negative Count (FN)	True Negative Count (TN)

Figure 2.7 a simple confusion matrix. [7]

As being seen in Fig 2.7, The numbers along the diagonal from upper-left to lower-right represent the correct decisions made, and the numbers outside this diagonal represent the errors. "The true positive rate (also called hit rate or recall) of a classifier is estimated by dividing the correctly classified positives (the true positive count) by the total positive count. The false positive rate (also called a false alarm rate) of the classifier is estimated by dividing the incorrectly classified negatives (the false negative count) by the total negatives. The overall accuracy of a classifier is estimated by dividing the total correctly classified positives and negatives by the total number of samples. Other performance measures, such as recall (sensitivity), specificity and F-measure are also used for calculating other aggregated performance measures (e.g., area under the ROC curves)" [7]. ROC defined as "a receiver operating characteristic (ROC), or simply ROC curve, is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate), at various threshold settings. TPR is also known as sensitivity, and FPR is one minus the specificity or true negative rate." [48].

$$\text{True Positive Rate} = \frac{TP}{TP + FN} \tag{2.8}$$

$$\text{True Negative Rate} = \frac{TN}{TN + FP} \tag{2.9}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.10}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.11}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.12}$$

Accuracy estimating for a classifier induced by supervised learning algorithms is important for some reasons. First, it can be used to estimate its future prediction accuracy, which could determine the error rate of prediction and imply the level of confidence one should have in the classifiers' output in the prediction system. Second, it can be used for choosing the best classifier from a given classifiers set. Lastly, "estimation accuracy can be used to assign confidence levels to multiple classifiers so that the outcome of a combining classifier can be optimized" [7]. The following subsection provides an explanation for one of the most popular estimation methodologies used for classification models, which applied successfully in this thesis.

2.4.1 The K-Fold Cross Validation

When the amount of data for training and testing is limited, t holdout method reserves a certain amount for testing and uses the remainder for training. In practical terms, it is common to hold out one-third of the data for testing and use the remaining two-thirds for training. Of course, the selected third for testing or training may be unlucky: The sample used for training or testing might not be representative. "In order to minimize the bias associated with the random sampling of the training and holdout data samples in comparing the predictive accuracy of two or more methods, one can use a methodology called k-fold cross validation. In k-fold cross validation, also called rotation estimation, the complete data set is randomly split into k mutually exclusive subsets of approximately equal size. The classification model is trained and tested k times. Each time it is trained on all but one folds and tested on the remaining single fold" [7]. The cross validation overall accuracy model calculated by simply averaging the k individual accuracy measures (Eq. 2.13).

$$CVA = \frac{1}{k}\sum_{i=1}^{k} A_i \qquad (2.13)$$

where CVA stands for cross validation accuracy, k is the number of folds used, and A is the accuracy measure (e.g., hit rate, sensitivity, specificity, etc.) of each fold. Since the cross-validation accuracy would depend on the random assignment of the individual cases into k distinct folds, a common practice is to stratify the folds themselves. In stratified k-fold cross validation, the folds are created in a way that they contain approximately the same proportion of predictor labels (i.e., classes) as the original dataset. Empirical studies showed that stratified cross validation tends to generate comparison results with lower bias and lower variance when compared to regular cross-validation. The k-fold cross validation is also called 10-fold cross validation, because the k taking the value of 10 has been the most common practice. In fact, empirical studies showed that ten seem to be an optimal number of folds A pictorial representation of the k-fold cross validation where k = 10 is given in Figure 2.8 The methodology (step-by-step process) that one should follow in performing k-fold cross validation is as follows:-

Step 1: The complete dataset is randomly divided into k disjoint subsets (i.e., folds) with each containing approximately the same number of records. Sampling is stratified by the class labels to ensure that the proportional representation of the classes is roughly the same as those in the original dataset.

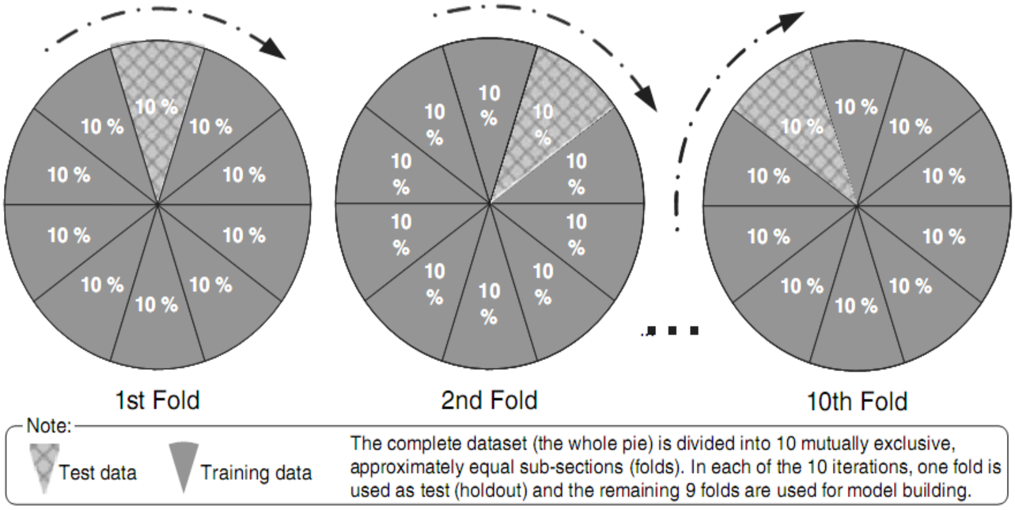

10 %
1st Fold
2nd Fold
10th Fold
Note:
Test data
Training data
The complete dataset (the whole pie) is divided into 10 mutually exclusive, approximately equal sub-sections (folds). In each of the 10 iterations, one fold is used as test (holdout) and the remaining 9 folds are used for model building.

Fig 2.8 Pictorial representation of 10-fold cross validation

Step 2: For each fold, a classifier is constructed using all records except the ones in the current fold. Then the classifier is tested on the current fold to obtain a cross-validation estimate of its error rate. The result is recorded.

Step 3: After repeating the step 2 for all 10 folds, the ten cross-validation estimates are averaged to provide the aggregated classification accuracy estimate of each model type.

The 10-fold cross validation does not require more data compared to the traditional single split (2/3 training, 1/3 testing) experimentation. In fact, in data mining community, for methods-comparison studies with relatively smaller datasets, k-fold type of experimental methods are recommended. In essence, the main advantage of 10-fold (or any number of folds) cross validation is to reduce the bias associated with the random sampling of the training and holdout data samples by repeating the experiment 10 times, each time using a separate portion of the data as the holdout sample. The down side of this methodology is that, one needs to do the training and testing for k times (k = 10 in this study) as opposed to only once. The leave-one-out methodology is very similar to the k-Fold cross validation where the value of k is set to 1. That is, every single data point is used for testing at least once on as many models developed as there are a number of data points. Even though this methodology is rather time consuming, for some small datasets it is a viable option. [7].

2.5 SUMMARY

Load profiling or consumption profiling play a big rule in the detection of the NTLs by determining the pattern of consumers and the way they consume water or energy. NTLs are difficult to measure because they are often unaccounted by the system operators and have no recorded information. Reducing NTLs is crucial for distribution companies. Water and energy consuming almost handled by a financial billing system to mange and calculate the delivered amount and save information records about customers. A lot of researchers have devoted a significant amount of effort in studying NTLs and FFD and analyze their underline business. Using Classification techniques is a popular trend in this kind of fraud. Real world problem tend to be non-linear, so this thesis select three non-linear classification techniques and evaluated them by 10-fold cross validation which is one of the best evaluation methods.

CHAPTER 3

RELATED WORK

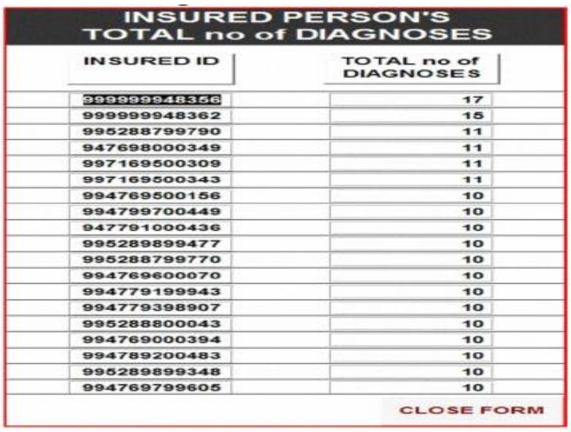
A lot of published papers about fraud detection using data mining techniques has been discussed and reviewed. Several groups of researchers have devoted a significant amount of effort in studying FFD from the same data, so they choose it to build the prediction model which can compare the transaction information with the historical trading patterns to predict the probability of a current transaction, and provide a scientific basis for the intelligent authorized anti-fraud strategy, or refuse to authorize and launch investigations to suspicious transactions. This chapter introduces the state-of the art for the applied techniques in fraud detection in some domains like banks, energy and health. The chapter addressing various technologies used in fraud detection, which are RDBMS, Data mining (unsupervised learning) and Data mining (supervised learning). Finally, the weak points of the related works were presented and discussed.

3.1 Credit Card Fraud Detection

The researchers in [1] apply three classification methods to credit card fraud detection problems, i.e. decision tree, neural network and logistic regression on a transactional database with more than 40 fields from year 2005 to 2006 to generate a predictive model, because of a nondisclosure agreement, they reveal a few variables, which are common data schema used by most banks. The data used were already labeled by the bank as fraud or non-fraud records. After that they evaluated the three models by the lift chart metric to measure the performance of targeting models in classification applications. The result state that the neural network model provides higher left than a logistic regression and decision tree.

3.2 Prescription Fraud Detection

In [2] The researchers design and apply a database application to record and control the prescription data in a health care center, the control included some restrictions and constraints to prevent fraudulent actions. The restrictions categorized as administrative to detect missing or invalid data and medical rules to check if there is a correspondence between prescribed drugs and the diagnoses that appears in the prescriptions. The application applied in 20000 real prescriptions data record. Finally, an interesting result from statistical analysis was acquired such 2% of an insured person had a high number of different diagnoses during the same year as shown in Figure 1 and that need further investigation. A powerful relational database helps in applying data mining techniques with real correct results.



INSURED PERSON'S TOTAL no of DIAGNOSES

INSURED ID	TOTAL no of DIAGNOSES
999999948356	17
999999948362	15
995288799790	11
947698000349	11
997169500309	11
997169500343	11
994769500156	10
994799700449	10
947791000436	10
995289899477	10
995288799770	10
994769600070	10
994779199943	10
994779398907	10
995288800043	10
994769000394	10
994789200483	10
995289899348	10
994769799605	10

Figure 3.1.Number of different diagnoses per insured person during the same year [2].

3.3 Financial Fraud Detection

In [8] the researchers from China and U.S.A analyzed and studied the used data mining techniques in 18 firms as a study reference, then they proposed a generic framework for data mining based on financial fraud detection as in Figure 3.2.

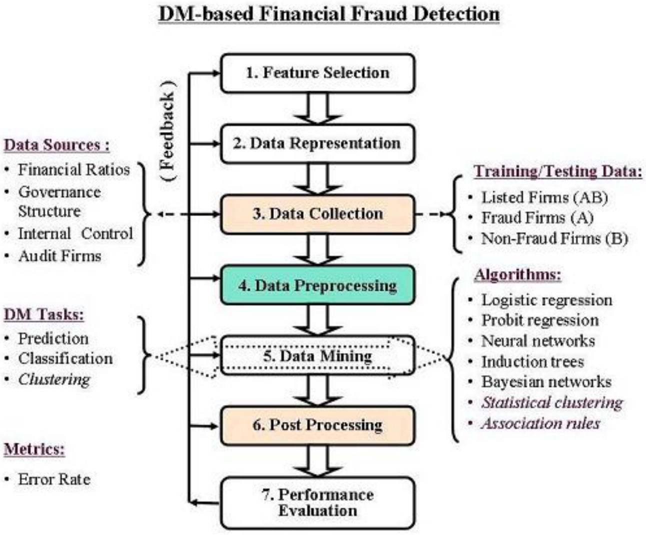


Figure 3.2 A Generic framework for DM-based FFD as in [8]

It is clear that classification methods are the dominant techniques in this generic frame work.

3.4 Expressway Toll Fraud Detection

In [14] The researchers apply anti-fraud analysis under the circumstance of the expressway network toll database system, because of more than 100,000,000 Yuan toll fraud amount in 21 areas of province's western part, eastern part, the Pearl river Delta and northern part in China, they examine the distance-based outlier mining technique to detect toll frauds in the expressway toll database system that diary tracks and records the toll transactions. The frauds mainly fall into driver fraud, toll collector fraud, and joint fraud by both of them. The researchers identify the suspicious data from a vast amount of database by using the outlier mining method.

3.5 Detection Of Abnormalities And Fraud In Customer Consumption

In [21] The researchers applied the non-supervised mining SOM (self-organizing map) technique in Brazil In which allows the identification of the consumption profile historically registered for a consumer, and its comparison with present behavior, and shows a possible fraud detection technique that used and applied in high-voltage electricity consumption transactions as shown in Fig 3.

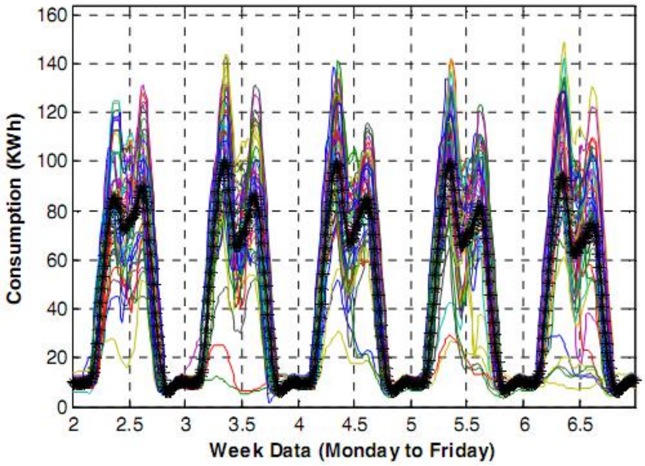


Figure 3.3 Graphic with the weeks of Cluster 1(44 weeks) [21]

It represents the consumption average within 44 weak and the compared with successor period of time shown in Fig 4.

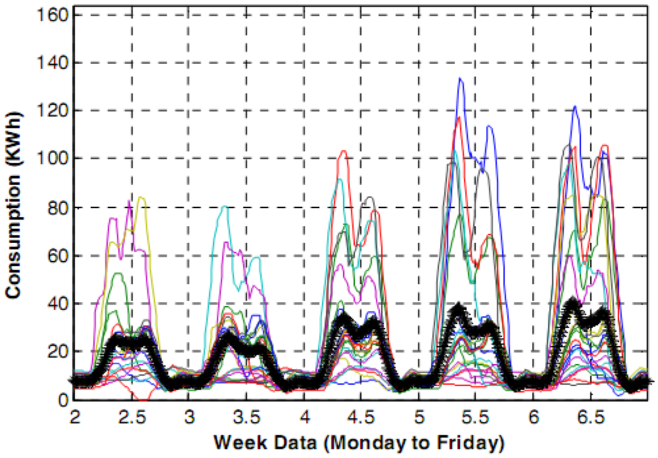


Figure 3.4. Graphic with the weeks of Cluster 2 (24 weeks) [21]

It’s clear that there is a change in a customer's consumption behavior alert auditors about something fraudulent or need further investigation**.**

3.6 Non Technical Loss Detection For Metered Customers In Power Utility

In [16] the researchers design and develop a fraud detection system for electricity metered customers, the data were collected from the power utility in kula-lambour , after feature extraction and selection, they define the normal load profile dataset and the fraudulent load profile dataset representing them as in Fig [3.6]

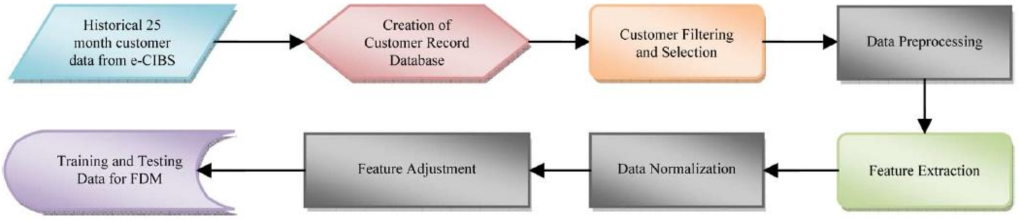

Historical 25 month customer data from e-CIBS
Creation of Customer Record Database
Customer Filtering and Selection
Data Preprocessing
Feature Extraction
Data Normalization
Feature Adjustment
Training and Testing Data for FDM

Fig3.5 Proposed framework for processing e-CIBS data for C-SVM training and validation [16]

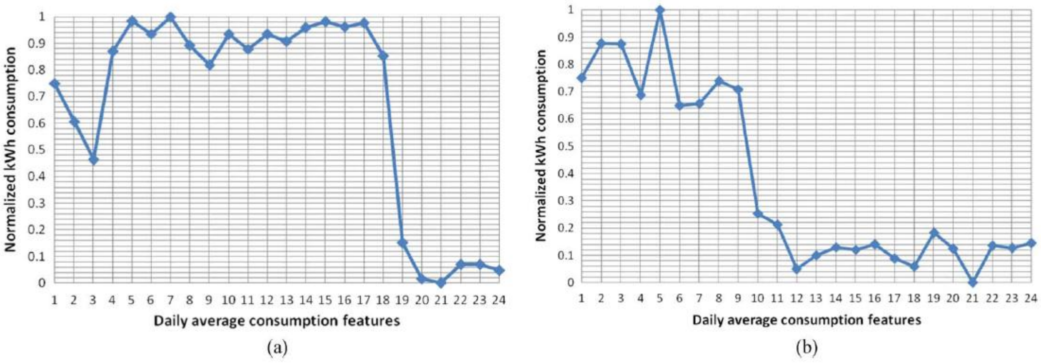


Figure 3.6 Normalized load profiles of two typical fraud customers over a period of two years.

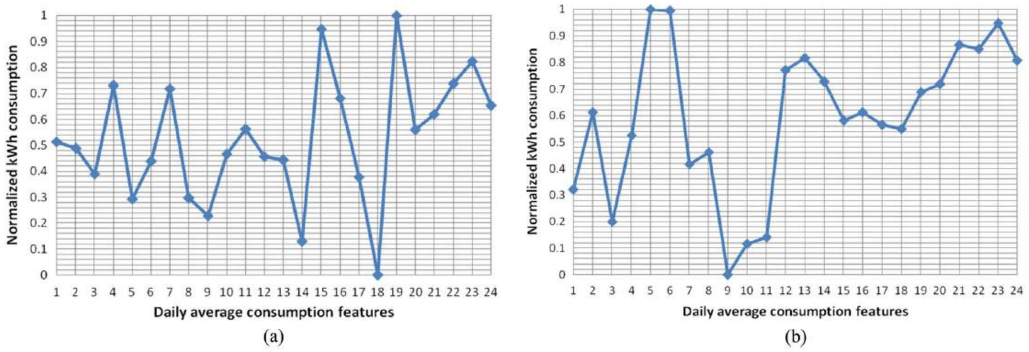


Figure 3.7. Normalized load profiles of two normal customers over a period of two years. [16]

By applying the SVM classification method, they improve the hit rate in fraud detection from the manual detection by the fraud team with 3% hit rate to 60% by the SVM model.

3.7 SUMMARY

Using RDBMS technology as in [2] is limited and works as a database constraint. SQL queries are slow and not considered as knowledge discovery science. The scientific methodology that applied in [21] for high electricity consuming fraud use a descriptive model in which cannot predict the state of new consuming instances. In [16] the researchers build a fraud detection model for a business deal with electricity consuming in kula-Lambur they improve the manual detection from 3% to 60% by intelligent model, they build the model in a small learning data set just 0.2 % of the filtered data.

CHAPTER 4

MODEL DEVELOPMENT

This chapter provides the methodology proposed for the fraud detection and the development of the fraud detection model which involves three stages: (I) data preprocessing, (ii) classification engine development, (iii) model testing and evaluation. The model applied using the SVM mining technique and compared with other two selected techniques (ANN and KNN).

4.1 Research Approach

The research approach is divided into two categories: (I) research methodology. (II) Work experimentation. The overall work completed for the applied model includes the experimentation methodology. The following sections give the overall view of the research model used for the purpose of the research study and experimental methodology.

4.1.1 Model Development Methodology

The proposed model is about detection customer fraudulent behavior within customers   consumptions historical profiles in Gaza city. The overall model methodology outlined in the thesis is given in Figure 4.1.

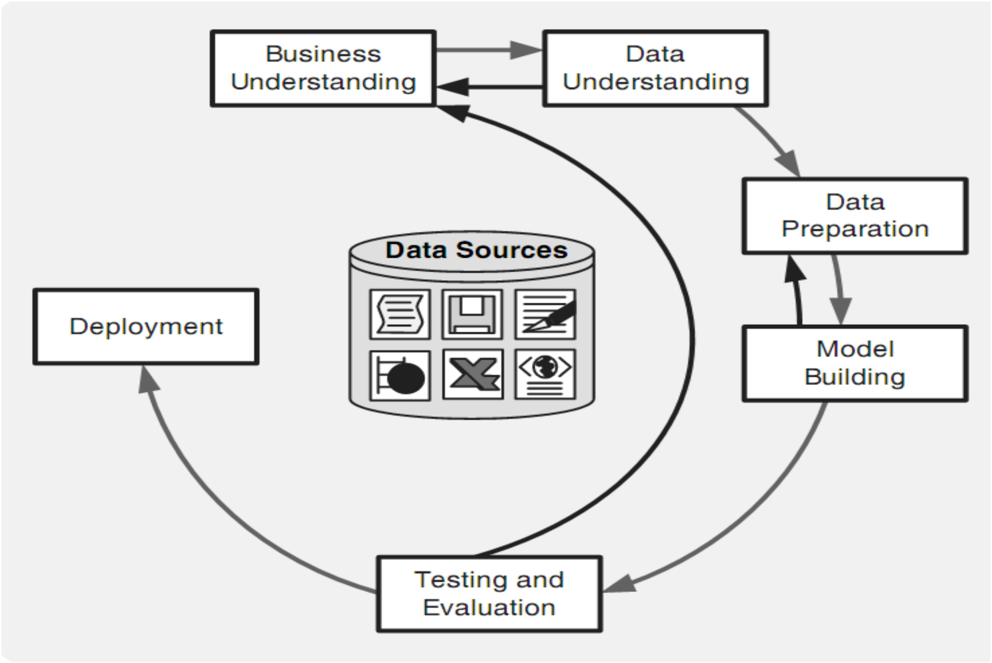


Figure4.1. Phases of CRISP-DM reference mode as in [34][3]

4.1.2 Research Methodology

The research methodology proposed in order to develop an intelligent fraud detection model for detection of water theft activities is shown in Figure 4.2.

The model methodology is embedded within the overall research methodology as in Figure 4.2.

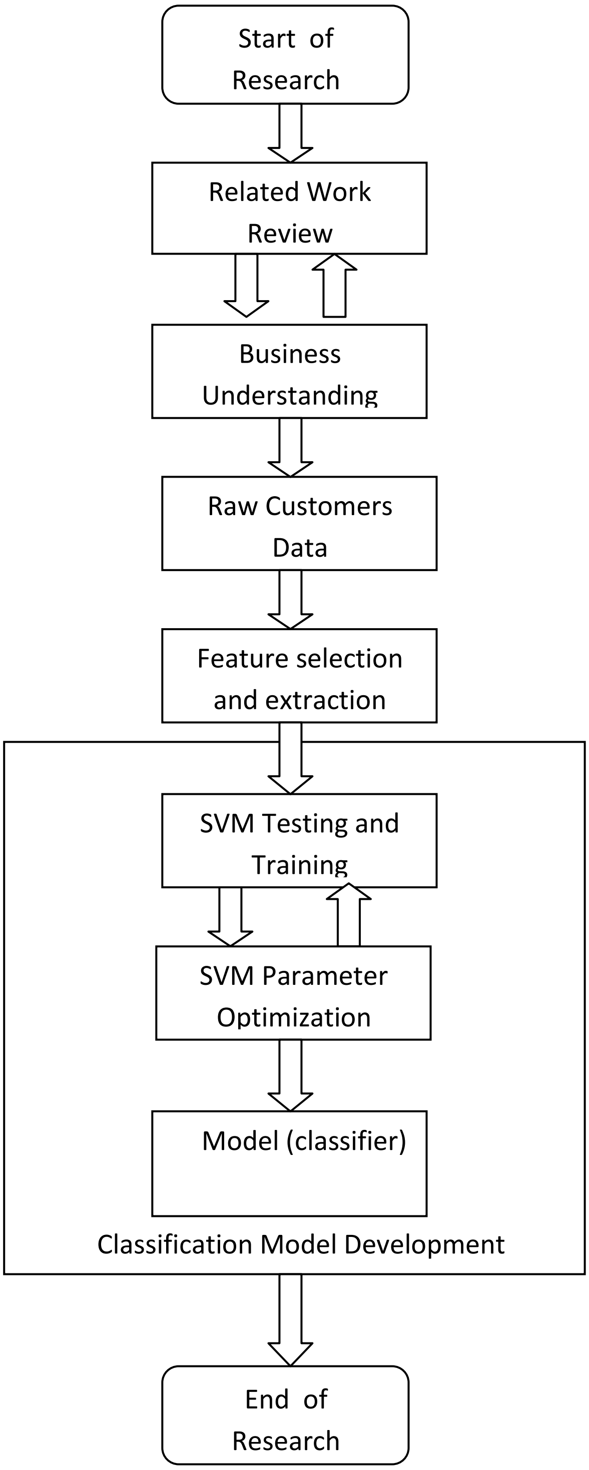


Figure 4.2 Thesis Research Methodology.

In this research study, the TOW detection approach illustrated in Figure 4.2 uses historical customer billing data of MOG customers and transforms the data into the required format for the classifier, by data preprocessing and feature extraction. Gaza customers are represented by their consumption profiles over a period of time. These profiles are characterized by means of patterns, which significantly represent their general behavior, and it is possible to evaluate the similarity measure between each customer and their consumption patterns. This creates a global similarity measure between normal and fraud customers as a whole. The identification, detection and prediction are undertaken by the SVM, which is the intelligent classification method and the best evaluated technique for the same underlying business in many scientific papers [16][17][18]. The fraud detection model developed in this thesis forms the basis of this research study. SVM training and testing is implemented using the software LIBSVM which is a library for SVMs [3] embedded in the RapidMiner software [32]. The computer used for training the SVM model was Dell Optiplex 990 workstation with Windows 7, core i3 processor with 4 GB of RAM. The following sections will further discuss the details of the processes outlined in the fraud detection model methodology in Figure 4.1,4.2.

4.2 Business Understanding And Data Collection

In business understanding phase, a lot of data about customers and water consuming collected and analyzed carefully. Some information was collected from the employees work in the same domain to understand that each customer has a water service agreement with a water meter installed in his building, periodically each three months or six months, the meter reader employee read the water meter to record the consumption per that period. After collecting all customers' readings, the employee who responsible for accounting, issue the monthly service invoices to be distributed to all customers according to a location number.

Data collection is one of the most important phases in the research project. It included studying the underlying business, data understanding and gathering information from the team who responsible for water theft in MOG. The water consumption data that used in this project and research study was identified by the TOW department team experts who have inspected customer premises and meter installations giving fraud cases and by billing system experts giving the historical customer water usage and billing information. The historical customer billings and consumption data were collected to train the proposed intelligent model to learn and differentiate between normal and suspicious consumption patterns. The data used for training the SVM was collected from MOG’s billing system, “Customer billing data are considered more reliable than any data that can be collected by survey techniques.” [50], which includes Gaza customer's invoices and historical transaction.

There are 62665 monthly invoices include just 30,348 water consumption customer invoice within 12 years about monthly historical consumptions. Twelve years of 30,348 monthly consumption represent 4.3 million historical records. The number of fraud customers is 2700 customers according to DWTC (before applying any filtering restriction). The data acquired for a period of 144 months, represents consumptions from 03/2000 to 02/2012. The data was obtained in the oracle database format. Two types of customer data were collected, which are: The customer information billing system (consumption profiles) and the customer's water irregularities data (fraud cases). Thesis topic is about the water consumption fraud detection, consumption profile feature is used in designing prediction models in [35], the water consumption of citizens collected and metered in a way like electricity consumption (load profile) which is used in a lot of research papers as in [21][10][16][17][22]. The water consumption profile used in [45] to build a water demand categorical prediction model. So for these reasons, customer's water consumption profile selected as a major feature in the proposed prediction model.

4.2.1 Customer Information Billing System Data ECBIS

The customer information and billing data based on Oracle database collected from the computer center of MOG. About 4.3 million records represent 144 months extracted in 144 tables; each table represents 30,348 customer consumption. Some features were extracted from the historical monthly consumption data table which has several attributes as listed in table 4.1.

Table 4.1 The attributes that extracted from historical water consumptions data.

Column	Description
Agreement_id	The customer account no
Service_type	The billing service type (water or electricity)
Meter_no	Customer water meter number
Reading_date	Water consumption Reading date
Previous_Reading	Meter Previous reading
Reader_id	The meter reader code to identify the meter reader name
Meter_status	To record if the meter normal or destroyed and othe states
Calc type	Record if the consumption automatic averaged or real quantity by reader
Current reading	Meter current reading
Batch_no	Reading batch or file no
Consumption_qty	Consumption quantity
Location_id	The location number to identify the building

Within data gathering and business understanding, another calculated feature was added to table 4.1 to enter feature selection phase as shown in table 4.2.

Table 4.2 The list of calculated attributes (to be evaluated in feature selection phase).

Column	Description
Building_agr_count	Water consumption accounts counts in the same building
Persons_per_building	Number of persons per building
Persons_per_unint	Number of persons per one unit
Payment_count_pct	The percentage of monthly paid voucher according to number of invoices
Paid_voucehr_count_pct	The percentage of paid vouchers according to invoices count

4.2.2 Customers Water Breaches Data (CWBD)

The data of water breaches customers collected by the team in DWTC who responsible for water theft in Gaza city when they conduct an inspection campaign. The data they collected last 12 years consist of 2700 fraud customer with the attributes listed in table 4.3.

Table 4.3 list the attributes of fraudulent cases.

Column	**Description**
Agreement_id	The customer account no.
Breach date	The date of water breach
Breach cost	The money needed to pay for that breach
Location_id	The building and street number

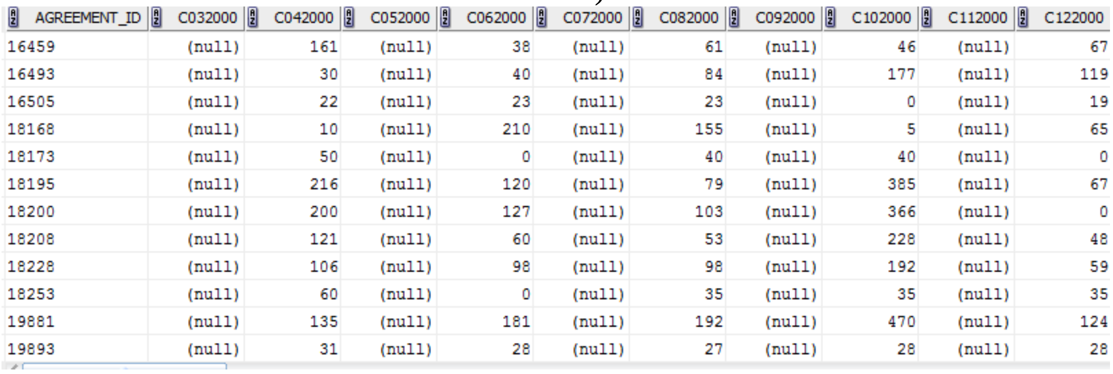
4.3 Data Preprocessing

Data preprocessing is the first major stage involved in the development of the fraud detection system. Data preprocessing involves data mining techniques in order to transform raw customer data into the required format, to be used by the classifier for detection and identification of fraud consumption patterns. Different data mining techniques have been applied to preprocess the ECIBS and CWBD data.

4.3.1 ECIBS And CWBD Data Preprocessing

As mentioned in section 4.2.1 about ECIBS data, which contains 4.3 million record represents 144 months extracted in 144 tables, each table represents 30348 customer consumption. Using SQL statements, all tables unified in one table with customer information and other 144 attribute about monthly consumption as indicated in fig 4.3.

Fig 4.3 144 column represent monthly customers consumptions (03/2000 to 02/2012).



AGREEMENT_ID	C032000	C042000	C052000	C062000	C072000	C082000	C092000	C102000	C112000	C122000
16459	(null)	161	(null)	38	(null)	61	(null)	46	(null)	67
16493	(null)	30	(null)	40	(null)	84	(null)	177	(null)	119
16505	(null)	22	(null)	23	(null)	23	(null)	0	(null)	19
18168	(null)	10	(null)	210	(null)	155	(null)	5	(null)	65
18173	(null)	50	(null)	0	(null)	40	(null)	40	(null)	0
18195	(null)	216	(null)	120	(null)	79	(null)	385	(null)	67
18200	(null)	200	(null)	127	(null)	103	(null)	366	(null)	0
18208	(null)	121	(null)	60	(null)	53	(null)	228	(null)	48
18228	(null)	106	(null)	98	(null)	98	(null)	192	(null)	59
18253	(null)	60	(null)	0	(null)	35	(null)	35	(null)	35
19881	(null)	135	(null)	181	(null)	192	(null)	470	(null)	124
19893	(null)	31	(null)	28	(null)	27	(null)	28	(null)	28

The agreement_id indicates the account number of the customer. As customer data is confidential to the utility so in order to protect the privacy of the customers, "Customer Names" have been omitted.

In the first ten years (3/2000 to 9/2010) the metered water consumption data was collected by the meter reader every two months to issue the bill and after that became every one month, so that there are null values in columns (empty cells) have no consumption reading. To fill these null values, a simple equation was applied on these columns by filling the previous null consumption value with half consumption of the next two month column value, as an example for one null column.

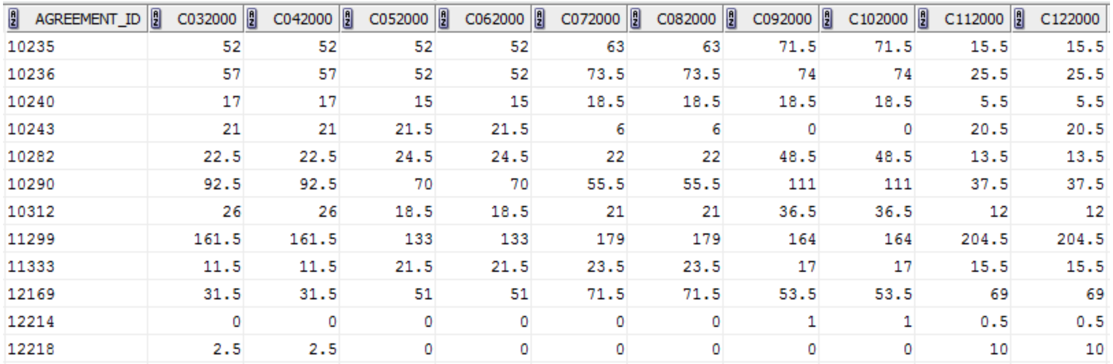
$$C_n = \frac{C_{n+1}}{2}$$

C_n Is the column value with n index number.

C_{n+1} Is the next column value, n is the odd month number 3,5,…9

After applying that equation, the data set will be filled as seen in figure 4.4

Fig 4.4 monthly customer consumption (03/2000 to 02/2012) after filling consumption columns.



AGREEMENT_ID	C032000	C042000	C052000	C062000	C072000	C082000	C092000	C102000	C112000	C122000
10235	52	52	52	52	63	63	71.5	71.5	15.5	15.5
10236	57	57	52	52	73.5	73.5	74	74	25.5	25.5
10240	17	17	15	15	18.5	18.5	18.5	18.5	5.5	5.5
10243	21	21	21.5	21.5	6	6	0	0	20.5	20.5
10282	22.5	22.5	24.5	24.5	22	22	48.5	48.5	13.5	13.5
10290	92.5	92.5	70	70	55.5	55.5	111	111	37.5	37.5
10312	26	26	18.5	18.5	21	21	36.5	36.5	12	12
11299	161.5	161.5	133	133	179	179	164	164	204.5	204.5
11333	11.5	11.5	21.5	21.5	23.5	23.5	17	17	15.5	15.5
12169	31.5	31.5	51	51	71.5	71.5	53.5	53.5	69	69
12214	0	0	0	0	0	0	1	1	0.5	0.5
12218	2.5	2.5	0	0	0	0	0	0	10	10

The two data sets ECIBS and CWBD are combined into one dataset with all attribute in the two datasets, "The important type of complex data structure arises from the integration of different databases. In modern applications of data mining, it is often necessary to combine data that come from different sources." [31]. Another new calculated attribute titled "FRAUD_STATE" has been added in which each customer founded in CWBD labeled with 'YES', the rest of other customers take "NO" value. This combined dataset was preprocessed to remove unwanted customers and smoothen out noise and other inconsistencies, in order to extract only useful and relevant information required. Data preprocessing as a part of Data understanding and preparation as shown in Figure 4.1 and illustrated in Figure. 4.5.

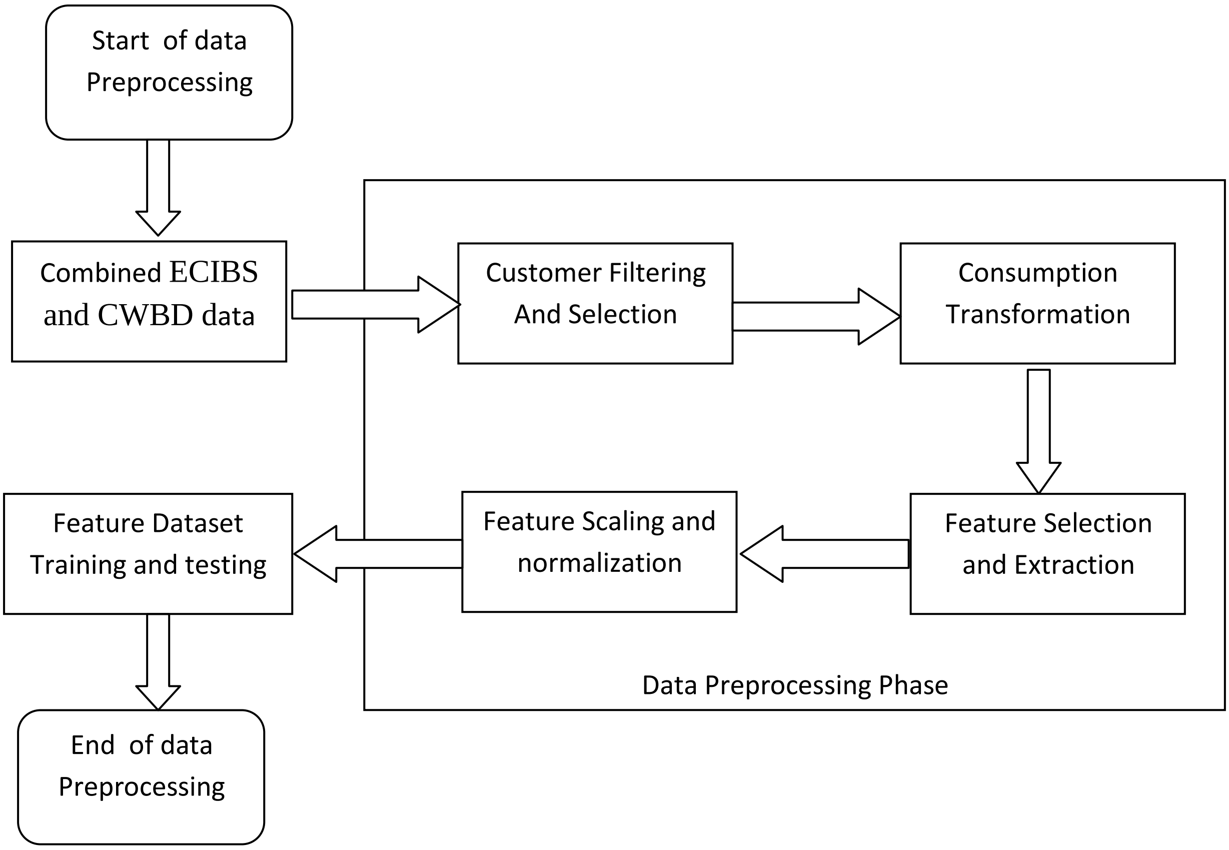

Start of data Preprocessing
Combined ECIBS and CWBD data
Customer Filtering And Selection
Consumption Transformation
Feature Selection and Extraction
Feature Scaling and normalization
Feature Dataset Training and testing
End of data Preprocessing
Data Preprocessing Phase

Fig 4.5: Flowchart of the ECIBS and CWBD data preprocessing

As shown in Figure 4.5, four major steps are involved in the preprocessing of ECIBS and CWBD data, which are as follows:

1. Customer Filtering and Selection
2. Consumption Transformation
3. Feature Selection and Extraction
4. Feature Normalization

The following sections further discuss in detail the four steps involved for ECIBS and CWBD data preprocessing.

4.3.1.1 Customer Filtering And Selection

As the ECIBS and CWBD data acquired from MOG in a raw format where about 62,665 customer accounts with 2700 fraudulent accounts, according to computer center and DWTC, therefore, in order to extract relevant and functional information, only customers with complete and useful data were selected from Gaza city for the classification model development. Since, the data acquired is in the form of a database, the Structured Query Language (SQL) applied to satisfy the criteria as listed:

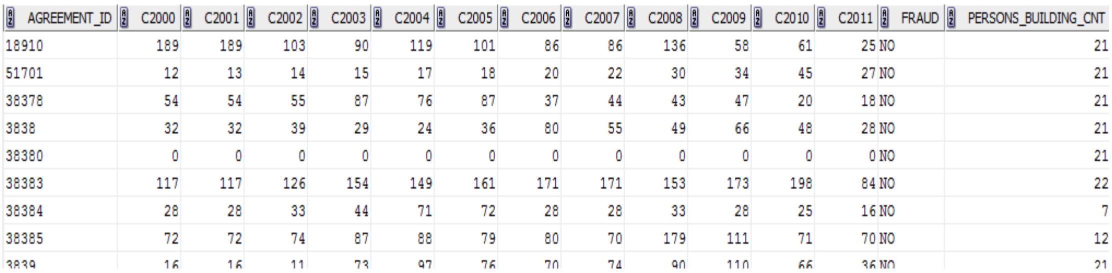
1. Remove repeating customers in the monthly ECIBS and CWBD data.
2. Remove customers having no consumption (i.e., zero cubic meter) throughout the entire
144 months period (Have no water service).
3. Remove customers have ending or suspending their service agreement
(consumption type = 50, 60 and 61).

After performing filtering and data reduction on the Gaza city data, only 28,845 customer records remained from all 62,665 customer Invoices and 660 fraudulent cases from original 2700. Even though approximately 50% of customers were removed after applying the filtering conditions mentioned previously, the amount of the remaining customers was more than sufficient for SVM training and model development according to several research paper sas in [16][17][18].

4.3.1.2 Consumption Transformation

Real world datasets tend to be noisy and inconsistent. Therefore, to overcome these problems, data mining techniques using statistical methods were applied to the ECIBS and CWBD data, in order to remove noise and other inconsistencies. Customers whose service agreement date began after the beginning of the consumptions period (3/2000 to 2/2012), have a null water consumption value across missing months, so all null values of this state have been updated to zeros.

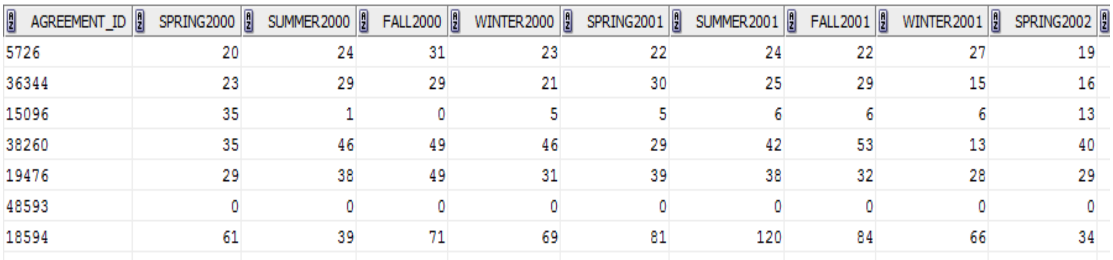
To experiment more than dataset, two new datasets have been created as in figures 4.6 and 4.7. The dataset in figure 4.6 represents the yearly customer consumption, so a transformation function has been applied to the 144 monthly consumptions to take the average of each 12 months for each year.



AGREEMENT_ID	C2000	C2001	C2002	C2003	C2004	C2005	C2006	C2007	C2008	C2009	C2010	C2011	FRAUD	PERSONS_BUILDING_CNT
18910	189	189	103	90	119	101	86	86	136	58	61	25	NO	21
51701	12	13	14	15	17	18	20	22	30	34	45	27	NO	21
38378	54	54	55	87	76	87	37	44	43	47	20	18	NO	21
3838	32	32	39	29	24	36	80	55	49	66	48	28	NO	21
38380	0	0	0	0	0	0	0	0	0	0	0	0	NO	21
38383	117	117	126	154	149	161	171	171	153	173	198	84	NO	22
38384	28	28	33	44	71	72	28	28	33	28	25	16	NO	7
38385	72	72	74	87	88	79	80	70	179	111	71	70	NO	12
3839	16	16	11	73	97	76	70	74	90	110	66	36	NO	21

Fig 4.6. The averaged yearly customer's consumptions.

Within the business understanding and data analysis and according to [11] the consumption quantity defers across year season. So a transformation function was applied to the 144 monthly consumption attribute to distribute every year on four season consumption by taking the average of each season to represent the season consumption. Each year became with four consumption value. The months of 6,7,8 represent Summer, Fall (9,10,11), winter (12,1,2) and spring (3,4,5).After transformation the number of attribute became (144/12)/4 = 48 attributes in the produced data set shown in figure 4.7.



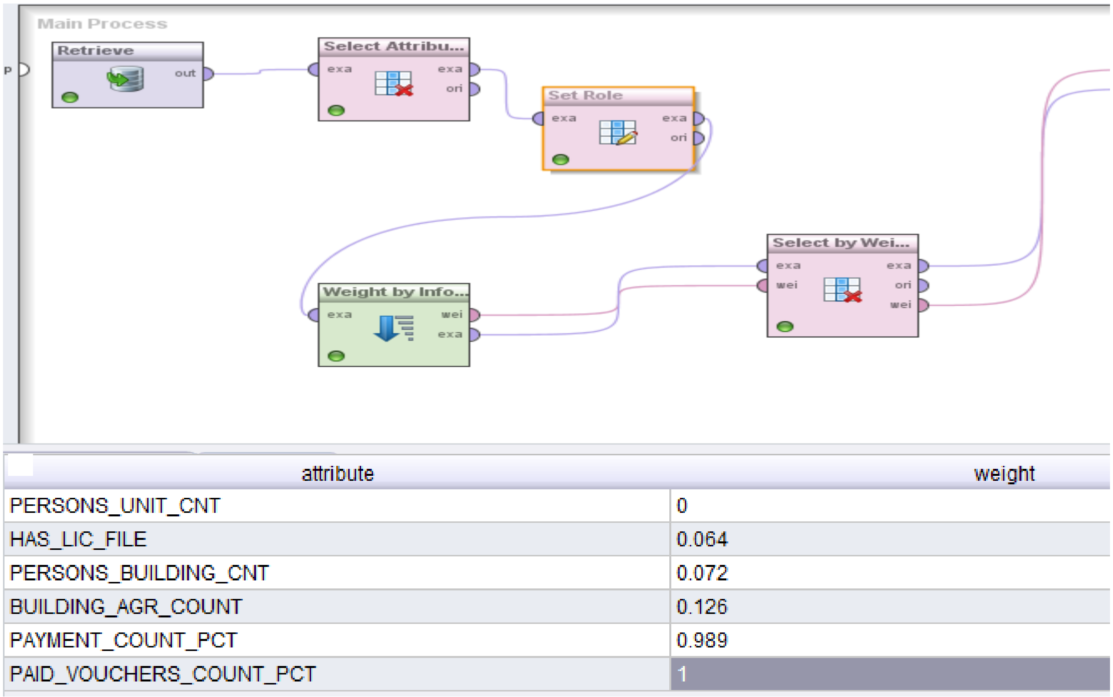
AGREEMENT_ID	SPRING2000	SUMMER2000	FALL2000	WINTER2000	SPRING2001	SUMMER2001	FALL2001	WINTER2001	SPRING2002
5726	20	24	31	23	22	24	22	27	19
36344	23	29	29	21	30	25	29	15	16
15096	35	1	0	5	5	6	6	6	13
38260	35	46	49	46	29	42	53	13	40
19476	29	38	49	31	39	38	32	28	29
48593	0	0	0	0	0	0	0	0	0
18594	61	39	71	69	81	120	84	66	34

Fig 4.7 The averaged four seasons customer's consumptions for twelve years.

4.3.1.3 Feature Selection And Extraction

Particular features were selected from the ECIBS data, in order to extract only useful and relevant information required for training the classification model. Since the proposed fraud detection approach applies consumption information of customers (load profiles) to the detection of fraud activities, therefore, the consumptions features are crucial part for pattern recognition in this research study.

Other features were evaluated for selection based on feature selection using information gain method, which is a popular method used to reduce the dimensionality [47]. From table 4.2 three features which are considered to be useful to the problem of NTL detection were evaluated and appeared in the top of evaluation results as shown in Fig 4.8.



attribute	weight
PERSONS_UNIT_CNT	0
HAS_LIC_FILE	0.064
PERSONS_BUILDING_CNT	0.072
BUILDING_AGR_COUNT	0.126
PAYMENT_COUNT_PCT	0.989
PAID_VOUCHERS_COUNT_PCT	1

Fig 4.8. The evaluation result after applying feature selection on the Extracted features of the unified dataset ECIBS and CWBD.

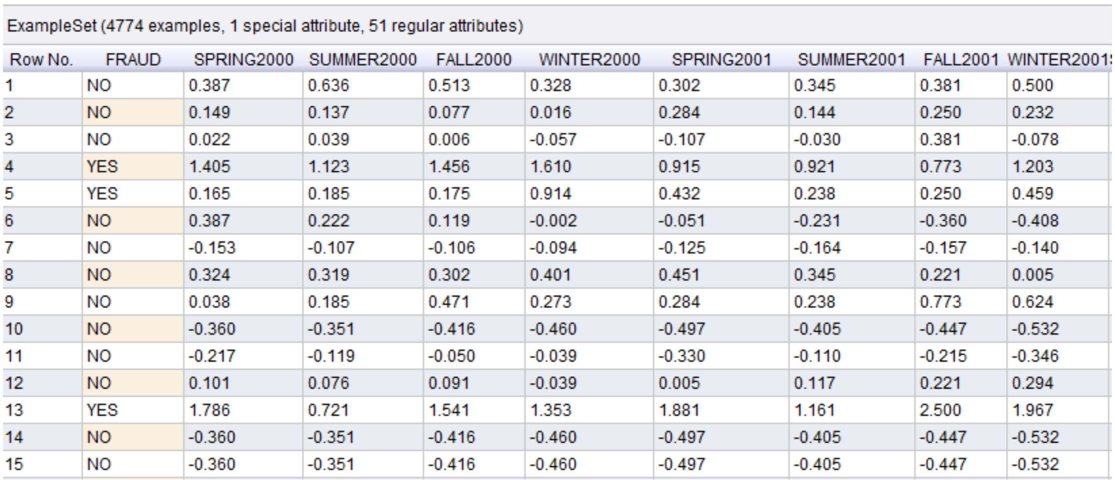
After evaluation, the top three features were selected to be with the consumption profile features in the training phase as shown in table 4.4.

Table 4.4 The list of the top three calculated attributes weights in feature selection phase.

Column	Description
Building_agr_count	Water consumption accounts counts in the same building
Payment_count_pct	The percentage of monthly paid voucher according to number of invoices
Paid_voucehr_count_pct	The percentage of paid vouchers according to invoices count

4.3.1.4 Feature Normalization

The selected features have a different scale, so, in order for the feature data to fit the SVM classification model properly [16][17][18], all feature data (44 features for 44 seasons) were represented using a normalized scale. All data was linearly scaled using the z-transformation. Figure 4.9 represents the normalized features in columns and the customers indicated by the rows.



ExampleSet (4774 examples, 1 special attribute, 51 regular attributes)

Row No.	FRAUD	SPRING2000	SUMMER2000	FALL2000	WINTER2000	SPRING2001	SUMMER2001	FALL2001	WINTER2001
1	NO	0.387	0.636	0.513	0.328	0.302	0.345	0.381	0.500
2	NO	0.149	0.137	0.077	0.016	0.284	0.144	0.250	0.232
3	NO	0.022	0.039	0.006	-0.057	-0.107	-0.030	0.381	-0.078
4	YES	1.405	1.123	1.456	1.610	0.915	0.921	0.773	1.203
5	YES	0.165	0.185	0.175	0.914	0.432	0.238	0.250	0.459
6	NO	0.387	0.222	0.119	-0.002	-0.051	-0.231	-0.360	-0.408
7	NO	-0.153	-0.107	-0.106	-0.094	-0.125	-0.164	-0.157	-0.140
8	NO	0.324	0.319	0.302	0.401	0.451	0.345	0.221	0.005
9	NO	0.038	0.185	0.471	0.273	0.284	0.238	0.773	0.624
10	NO	-0.360	-0.351	-0.416	-0.460	-0.497	-0.405	-0.447	-0.532
11	NO	-0.217	-0.119	-0.050	-0.039	-0.330	-0.110	-0.215	-0.346
12	NO	0.101	0.076	0.091	-0.039	0.005	0.117	0.221	0.294
13	YES	1.786	0.721	1.541	1.353	1.881	1.161	2.500	1.967
14	NO	-0.360	-0.351	-0.416	-0.460	-0.497	-0.405	-0.447	-0.532
15	NO	-0.360	-0.351	-0.416	-0.460	-0.497	-0.405	-0.447	-0.532

Fig 4.9: The normalized features over a period of 44 seasons.

The normalizations were applied on all three data sets (Yearly, seasonally and monthly).

4.4 Classification Engine Development

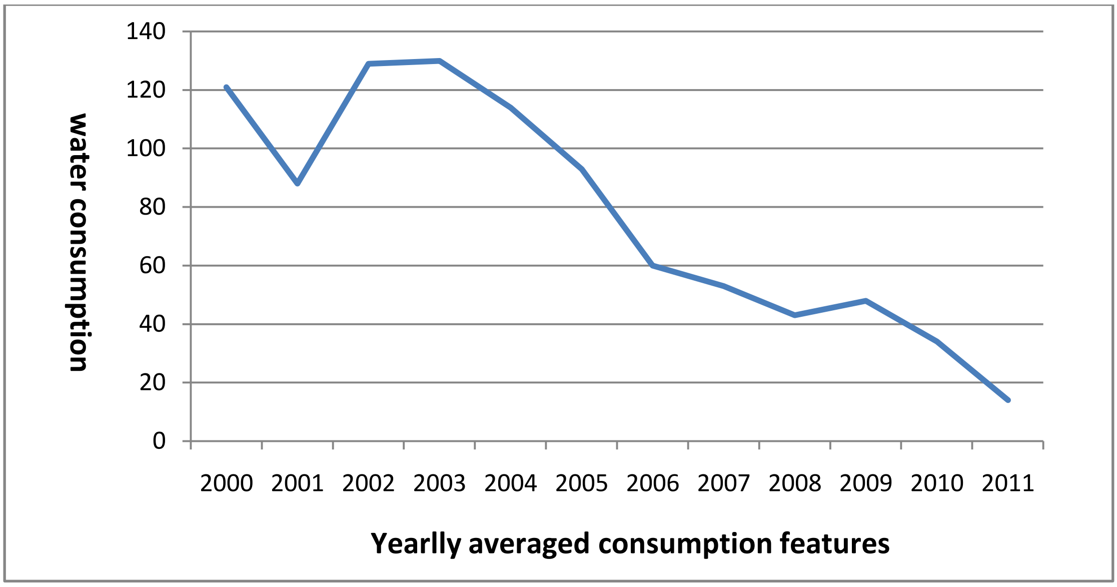
The development of the classification engine, namely the SVM model, is the main focus of this project and research study. Development of the classification engine involves: load profile inspection for detection of normal and fraud customers, training and development of the SVM classifier, SVM parameter tuning, class weight adjustment and SVM training and testing. The following sections will further discuss in detail the development of the SVM engine.

4.4.1 Load Profile Inspection

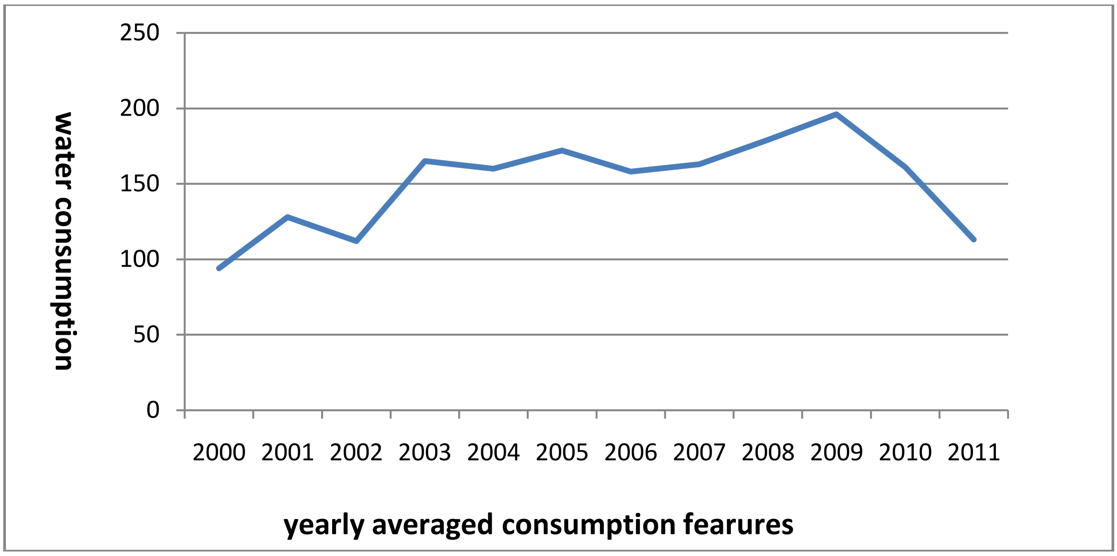
Load profiles, i.e., the 44 seasonally averaged water consumption features of the customers were inspected to retrieve samples, in order to build the SVM model for the purpose of training. As in this study, a 2-class SVM classifier is used to represent two different types of customer load profiles, therefore, load profile samples used to build the SVM classifier were extracted from the preprocessed Gaza city data as shown in Table 4.1. The load profiles inspected were extracted and classified into two different categories according to their behavior, i.e., fraud or normal consumption (non-fraud).

From the 28,845 filtered customers in the Gaza city data, only 660 customers were identified and detected as Theft of Water (TOW) cases by (MOG DWTC team) in the past twelve years. Utilizing the TOW information as the class label for the features results in an unbalanced dataset, as there are only 660 TOW cases from the total 28,845 customers, i.e., only 2.2% customers are fraud while the remaining are good. Therefore, in this scenario, an unbalanced class ratio is achieved [17][25], for which the accuracy does not prove to be accurate anymore. So all fraud customers were selected to train SVM model to identify the fraud customer's load profile and some of the non-fraud remaining customers with no TOW cases were selected as the researchers did in [1][17].The remaining customers with no TOW were identified as normal customers by manual labeling. Customers with no fraud activities (No TOW) and the customers with fraud activities (TOW) form the backbone for the development of the SVM model.

Manual inspection was performed on a large sample of the remaining 28,185 customers. Customers' load profiles have been inspected and filtered in which abrupt changes appear clearly, indicating fraud activities, abnormalities and other irregularities in consumption characteristics. Only 4114 customer's load profiles were identified with no presence of abrupt or sudden drops relating to fraudulent events. These 4114 customers' load profiles selected as normal consumption in order to train the SVM classification model to identify the normal customer load profile. Figure 4.10 indicates the load profile of one typical fraud customers and another one normal consumption load profile over a period of twelve years (the beginning of the new billing system was since twelve years).



(a.)



(b.)

Fig 4.10: Load profiles of one typical fraud customer (a) and another normal customer load profile (b) over a period of twelve years.

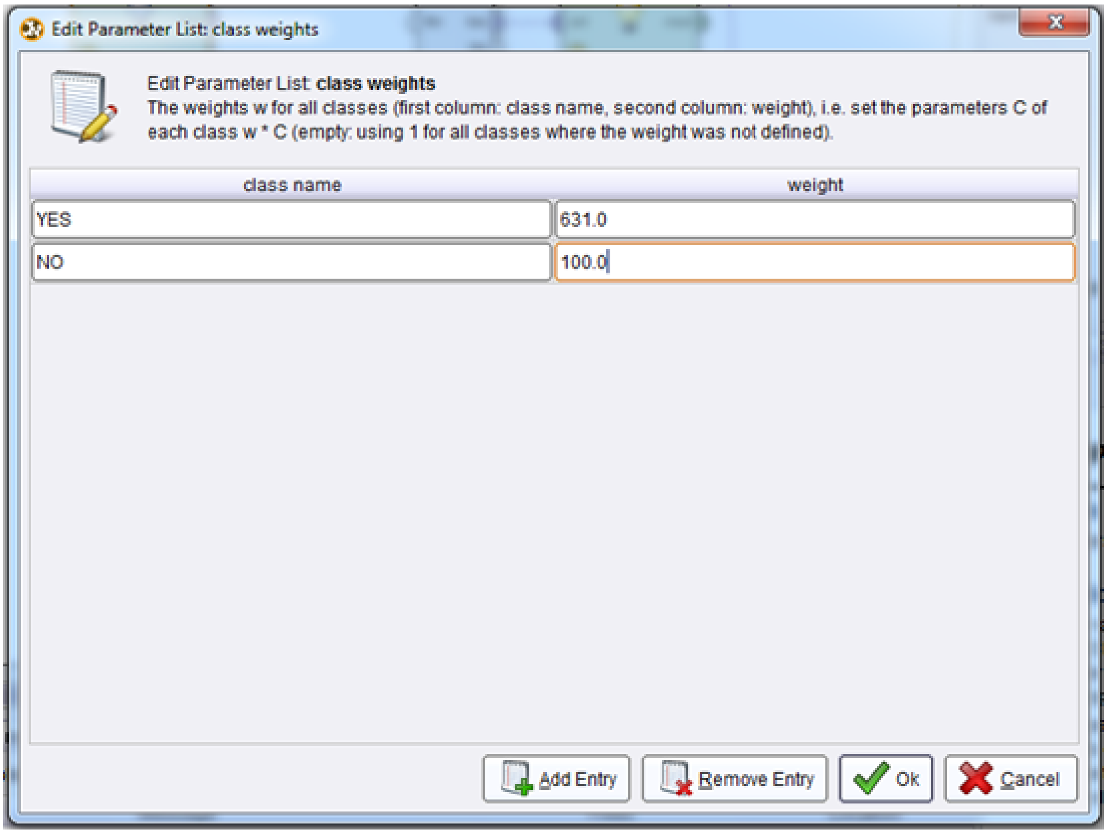
Note that, the drop in Figure 4.10 (b) is a general drop caused since 2009 when MOG change the policy of water meter reading from every two months to every 6 months.

4.4.2 SVM Development

After load profile inspection, a new data set created with 4114 normal profiles and 660 fraud profile with total 4774 in which represent 15% of the overall metered water dataset to enter the model development phase and used to train the SVM classifier in order to create an SVM model. The SVM model development consists of a few stages, which includes: weight adjustment, parameter optimization, and SVM training and testing. The following sections will further discuss in detail the development stages of the SVM model.

4.4.2.1 Weight Adjustment

The ratio between the two classes of samples is still unbalanced, Class one (no fraud) having 4114 samples and Class 2 (with fraud) having 660 samples; therefore, fortunately SVM classifier technique has a parameter set used to weight and balance the samples' ratio. The weights are adjusted by calculating the sample ratio for each class. This is achieved by dividing the total number of classifier samples with the individual class samples. In addition, class weights are also multiplied by a weight factor of 100 in order to achieve satisfactory weight ratios for training. Fig 4.11 as in rapid miner.


Edit Parameter List: class weights

Edit Parameter List: **class weights**
The weights w for all classes (first column: class name, second column: weight), i.e. set the parameters C of each class w * C (empty: using 1 for all classes where the weight was not defined).

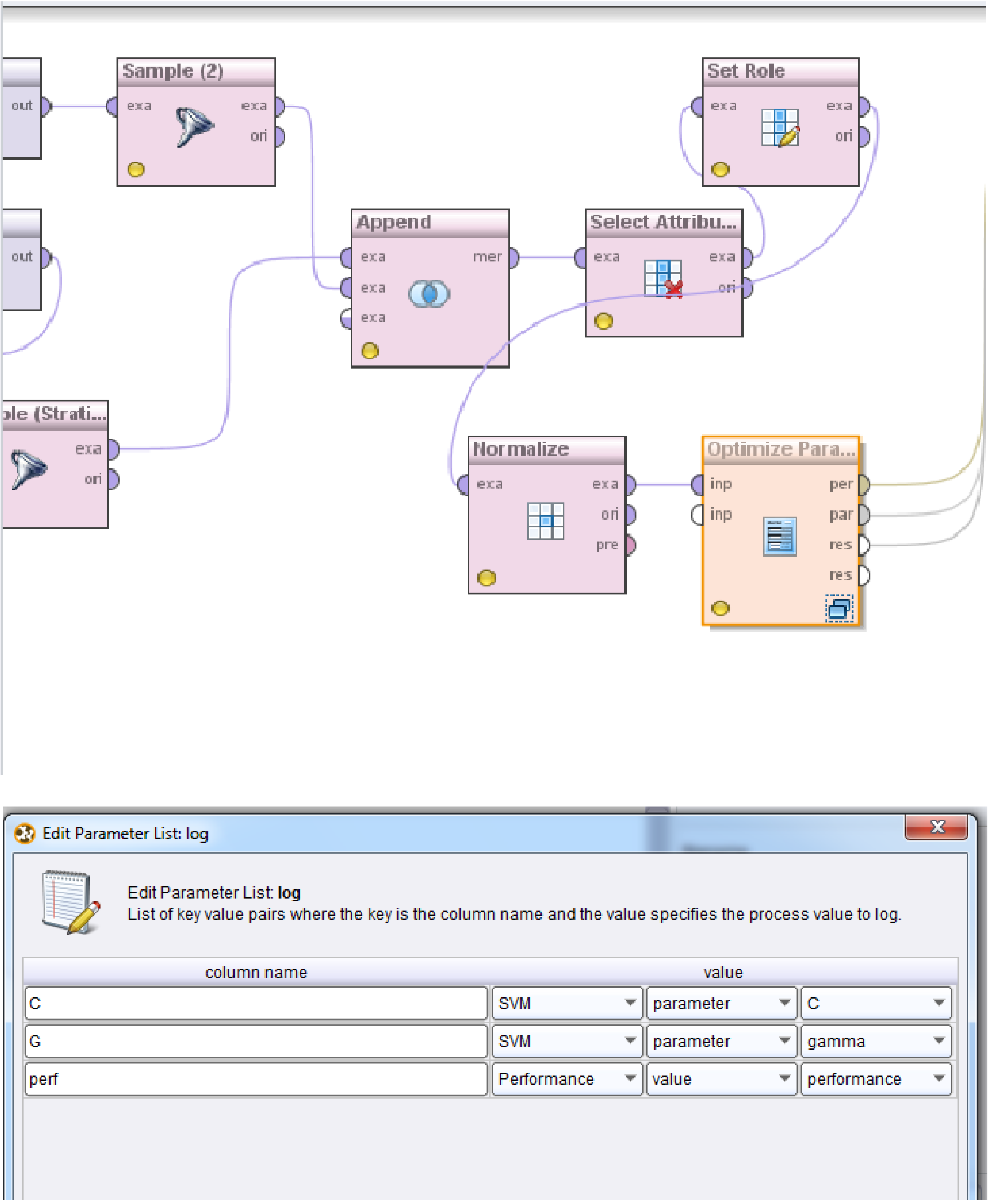
class name	weight
YES	631.0
NO	100.0

Fig 4.11 SVM Label classes weighted as 631% for YES and 100% for NO.

The other techniques (ANN and KNN) do not have a parameters to balance class labels, so the under-sampling technique was applied to balance classes, under-sampling is a popular method in addressing the class imbalance problem, which eliminates training examples of the over-sized class until it matches the size of the other class. Since it discards potentially useful training examples, the performance of the resulting classifier may be degraded. Nevertheless, some studies have shown that under-sampling is effective in learning with imbalanced datasets [52][53][54][56], sometimes even stronger than over-sampling, especially on large datasets[55][53].

4.4.2.2 Parameter Optimization

To find the optimal values of the SVM parameters, Three parameters were defined and entered as inputs to the grid search technique. The Grid Search method proposed by Hsu et al. in [4] is used for SVM parameter optimization. In the grid search method, exponentially growing sequences of parameters (C, Gamma) are used to identify SVM parameters obtaining the best 10-fold CV accuracy [17]. The parameters C, Gama and performance tested and evaluated within a ranged values from 0 to 100 using rapid miner software as shown in Fig 4.12.



Edit Parameter List: log

Edit Parameter List: **log**
List of key value pairs where the key is the column name and the value specifies the process value to log.

column name	value		
C	SVM	parameter	C
G	SVM	parameter	gamma
perf	Performance	value	performance

Fig 4.12 SVM parameters optimizations using grid search technique.

The maximum performance achieved when c and gamma parameters equal zeros (default value).

4.4.2.3 SVM Training And Testing

After weight adjustment and parameter optimization, all 4774 samples are trained in order to build an SVM model (classifier). In order to facilitate fair comparison between classifiers, the same number of model inputs was adopted for the experiment [51], RapidMiner software version 5.2 [32] used to conduct training. As indicated in Figure 4.13.

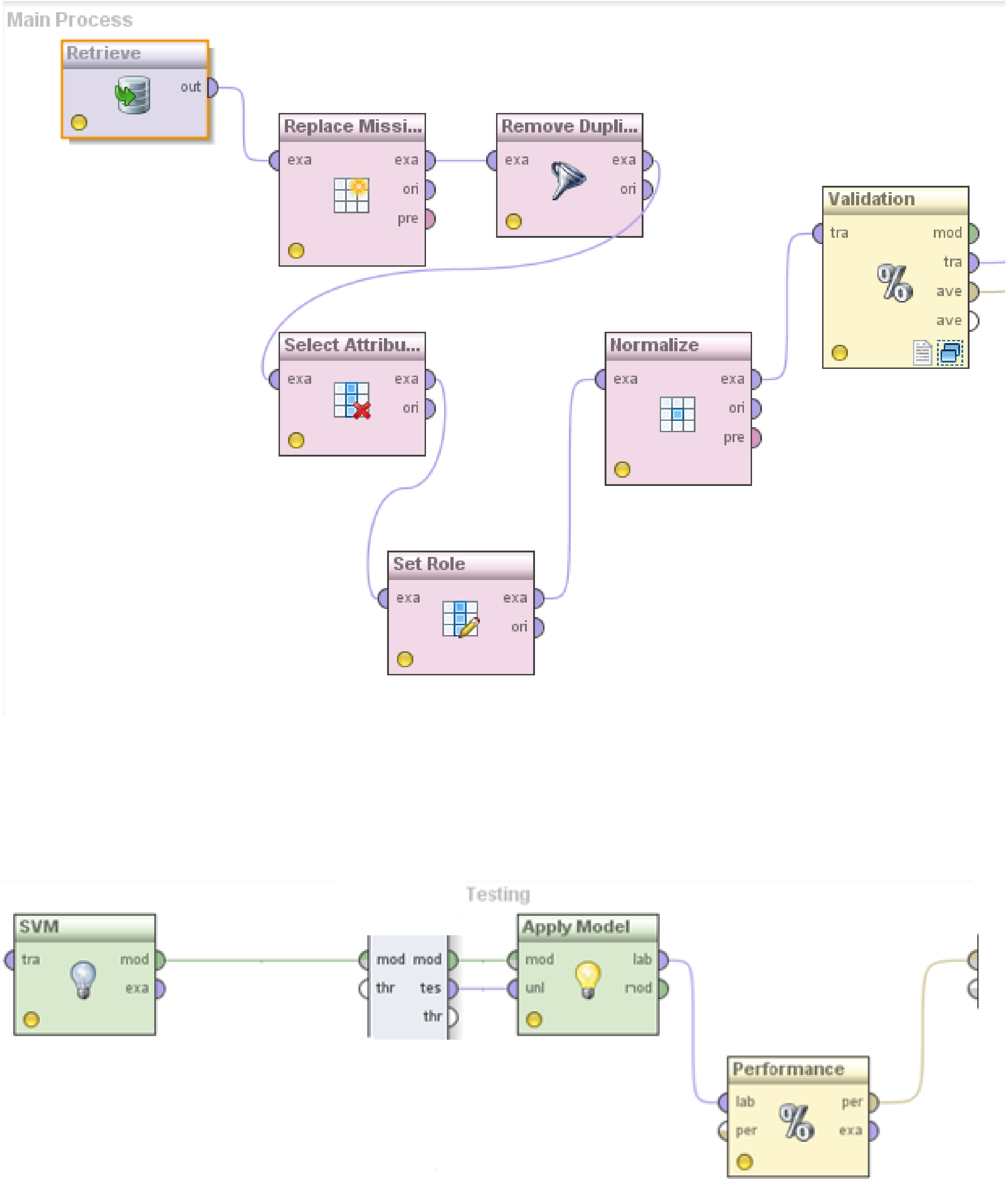


Fig 4.13 SVM Classifier training and testing using cross validation.

In order to classify customers as fraud or normal. Ten fold Cross validation used to test and evaluate three classifiers which are SVM, Neural Network and KNN. Each classifier trained and tested on three data sets structure, yearly consumption data set (Yearly_DS), season's consumptions data set (Seasonally_DS) and monthly consumption data set (Monthly_DS). Each Data set include 4114 records with NO fraud class and 660 records with YES fraud class. Each classifier applied twice on each data set, one with just consumptions profile features and the other with just consumptions profile features in addition to the selected attribute in table 4.4. ANN and KNN classifiers haven't any parameter to weight classes in order to balance data set as in SVM classifier, so balancing data is applied by performing random sampling on the larger class. The final experimentation results shown in chapter 5.

4.5 SUMMARY

This chapter expressed the implemented methodology to apply the fraud detection model on three datasets, two major stages were involved in the development of intelligent fraud detection model which include (i) Data preprocessing (ii) classification engine development. The data preprocessing sub chapter illustrated data mining techniques used for preprocessing the raw customer information and billing data for feature selection and extraction. The sub chapter, classification engine development illustrated the feature selection, parameter optimization, development of the SVM classifier and the SVM training and testing and lastly, the SVM model evaluated with two other classification techniques which are ANN and KNN.

CHPTER 5

EXPERIMENTATION RESULTS

This chapter presents the evaluation results between the three selected classification techniques. Tables (5.1, 5.2, 5.3) explain the experimental results after applying and testing the SVM model on the yearly averaged consumptions dataset, seasonally averaged dataset and monthly averaged dataset. After that, SVM classification results compared with the other two mentioned classifiers (ANN and KNN). Each data set has been trained and tested twice per each classifier. As mentioned in 4.4.2.3 section, once with just consumptions load profile attributes (averaged consumptions attributes within a period of time) and the last with all attributes (averaged consumptions within a period of time plus top features in table 4.4).

5.1 Yearly Consumption Dataset Experimentation Results.

Unbalanced data sets result in high accuracy, but the researchers did not prove that with unbalanced datasets as in [17][25], so this study focused on balanced data sets.
The results denote that SVM classifier gets 87.62 best accuracy score when dealing with just consumptions profile attributes. The second classifier is ANN with a 76.80 accuracy score. The biggest score for the fraud detection hit rate is achieved by ANN, which is 71.06 using all attributes as shown in table 5.1.

Table 5.1 Yearly consumption data set testing and evaluating results.

Data Set Type	**Classifier**	**Performance**					**Data Set Status**
		Accuracy	**Recall**		**Precision**		
			Yes	**No**	**Yes**	**NO**	
Yearly_DS (Load Profile)	KNN	89.99	45.30	97.09	71.19	91.78	Unbalanced
Yearly_DS (All Selected Attr.)	KNN	89.95	44.09	97.23	71.67	91.63	Unbalanced
Yearly_DS (Load Profile)	KNN	76.43	68.79	83.57	79.65	74.12	Balanced
Yearly_DS (All Selected Attr.)	KNN	74.82	65.61	83.57	78.73	72.18	Balanced
Yearly_DS (Load Profile)	Neural Network	89.80	44.09	97.06	70.46	91.62	Unbalanced
Yearly_DS (All Selected Attr.)	Neural Network	89.84	49.70	96.22	67.63	96.22	Unbalanced
Yearly_DS (Load Profile)	Neural Network	76.80	66.67	86.26	81.94	73.46	Balanced
Yearly_DS (All Selected Attr.)	Neural Network	75.63	71.06	79.89	76.76	74.70	Balanced
Yearly_DS (Load Profile)	SVM	**87.62**	**56.52**	92.56	54.69	93.05	Balanced
Yearly_DS (All Selected Attr.)	SVM	**85.98**	**61.06**	89.94	49.09	93.56	Balanced

5.2 Seasonally Consumption Dataset Experimentation Results

When dealing with seasonally consumptions data sets, the results denote that SVM classifier get the best accuracy score which is 93.76 using all attributes and 81.32 of fraud detection hit rate using just consumptions attribute and 80.61when using all selected attributes. The next classifier is ANN, which has a 88.45 accuracy score with 85.45 fraud detection rate which is the best hit rate score; KNN classifier has the lowest score in accuracy and detection rate as shown in table 5.2.

.

Table 5.2 Seasonally consumptions data set testing and evaluating results.

Data Set Type	Classifier	Performance					Data Set Status
		Accuracy	Recall		Precision		
			Yes	No	Yes	NO	
Seasonally_DS (Load Profile)	KNN	91.58	51.15	98.48	85.20	92.19	Unbalanced
Seasonally_DS (All Selected Attr.)	KNN	91.62	50.30	98.25	82.18	92.49	Unbalanced
Seasonally_DS (Load Profile)	KNN	82.57	75.04	89.01	85.37	80.66	Balanced
Seasonally_DS (All Selected Attr.)	KNN	79.55	70.76	86.62	80.94	78.67	Balanced
Seasonally_DS (Load Profile)	Neural Network	96.43	77.18	99.71	97.86	96.24	Unbalanced
Seasonally_DS (All Selected Attr.)	Neural Network	96.46	76.52	99.66	97.30	96.36	Unbalanced
Seasonally_DS (Load Profile)	Neural Network	88.85	85.45	91.75	89.86	88.07	Balanced
Seasonally_DS (All Selected Attr.)	Neural Network	87.31	84.09	89.90	86.99	87.56	Balanced
Seasonally_DS (Load Profile)	SVM	**93.12**	**81.32**	95.14	74.06	96.76	Balanced
Seasonally_DS (All Selected Attr.)	SVM	**93.76**	**80.61**	95.87	75.78	96.86	Balanced

5.3 Monthly Consumption Dataset Experimentation Results

The last trained and tested dataset denote that SVM classifier get the best accuracy which is 92.29 using just consumption load profile and 80.45 score for the fraud detection hit rate in the balanced monthly consumption dataset. The ANN classifier has the optimal fraud detection rate when dealing with all attribute which is 87.12 as shown in table 5.3.

Table 5.3 Monthly consumptions data set testing and evaluating results.

Data Set Type	Classifier	Performance					Data Set Status
		Accuracy	Recall		Precision		
			Yes	No	Yes	NO	
Monthly_DS (Load Profile)	KNN	91.75	47.42	98.80	86.23	92.20	Unbalanced
Monthly_DS (All Selected Attr.)	KNN	92.15	51.06	98.68	85.97	92.70	Unbalanced
Monthly _DS (Load Profile)	KNN	76.89	70	83.73	82.04	74.34	Balanced
Monthly_DS (All Selected Attr.)	KNN	77.65	70.61	84.64	82.04	74.34	Balanced
Monthly_DS (Load Profile)	Neural Network	95.24	71.52	99.01	92.01	95.63	Unbalanced
Monthly_DS (All Selected Attr.)	Neural Network	95.14	71.52	98.89	91.12	95.62	Unbalanced
Monthly_DS (Load Profile)	Neural Network	84.90	86.21	83.58	83.92	85.91	Balanced
Monthly_DS (All Selected Attr.)	Neural Network	85.13	87.12	83.13	83.70	86.66	Balanced
Monthly_DS (Load Profile)	SVM	**92.29**	**79.55**	93.81	67.14	96.65	Balanced
Monthly_DS (All Selected Attr.)	SVM	**91.86**	**80.45**	95.65	74.79	96.83	Balanced

This study adopted the SVM classifier for the following reasons. First, balancing technique used for ANN and KNN depend on random sampling in which decrease the number of   instances in the training data set to more than the half. Second, the SVM classifier depend on class weighting technique to balance data set without omitting any instance. Third SVM got the maximum accuracy score with balanced data sets in all three datasets. Finally, there is no big difference in the maximum fraud detection score rate between SVM and ANN.

5.4 SVM Classifier Evaluating Results.

The results presented in tables 5.1, 5.2, 5.3 imply that SVM got the best results in all balanced datasets as concluded and illustrated in table 5.4. As being seen in these tables, applying SVM classifier on the seasonally database structure that represent the normal change in water consumptions according to seasonal weather temperature change, SVM got the best accuracy result with all selected features.

Table 5.4 Represents SVM classifier results after training and testing phase.

Data Set Type	**Classifier**	**Performance**					**Data Set Status**
		Accuracy	**Recall**		**Precision**		
			Yes	**No**	**Yes**	**NO**	
Yearly_DS (Load Profile)	SVM	**87.62**	**56.52**	92.56	54.69	93.05	Balanced
Yearly_DS (All Selected Attr.)	SVM	**85.98**	**61.06**	89.94	49.09	93.56	Balanced
Seasonally_DS (Load Profile)	SVM	**93.12**	**81.32**	95.14	74.06	96.76	Balanced
Seasonally_DS (All Selected Attr.)	SVM	**93.76**	**80.61**	95.87	75.78	96.86	Balanced
Monthly_DS (Load Profile)	SVM	**92.29**	**79.55**	93.81	67.14	96.65	Balanced
Monthly_DS (All Selected Attr.)	SVM	**91.86**	**80.45**	95.65	74.79	96.83	Balanced

The calculated attributes increase the accuracy and recall in which define the importance and affection for them in the prediction and determination of fraudulent customer’s profile. So the fraudulent behavior affected with the Paid_vouchers_count_pct attribute that determines how many customer invoices have been totally paid regardless of how many times the customer visits the MOG; thereby, customer committing in paying invoices decrease the probability of fraud.

With Payment_count_pct attribute that determines the percentage of customer payment's count with regard to how many visit he came to MOG to pay the invoices, the fraud probability decreased with a higher percentage. The Building_agr_count attribute that determines the number of customers' accounts per building (invoices per building) has a little effect in defining fraudulent customer profile but less than the other attribute as illustrated in fig 4.3. The consequence of using these three attributes is the increasing in fraud detection hit rate (the true positive rate) to reach 80.61with 93.76 overall accuracy.

5.5 SUMMARY

In spite of, the SVM predict the instance class as a black box namely without induce a descriptive rule explains the attributes that define the class label, this research concentrate on determining if the tuple fraudulent or not in addition to the SVM classifier has a notable number of advantages when compared with the neural networks classifier. Firstly, SVM has non-linear dividing hypersurfaces that give it high discrimination. Secondly, SVM provides a good generalization ability for unseen data classification. In addition, SVM determines the optimal network structure itself, which is not the case with traditional neural networks. With the introduction of the applied SVM SVC technique, the developed model can control the balance between sensitivity and specificity, giving the fraud detection system more flexibility. Thus, for this type of research study, SVM is more practical and favorable, where this control is most needed in the frequent presence of unknown and unbalanced data sets as the researcher done in [17].

CHAPTER 6

CONCLUSION AND FUTURE WORK

In this research study, the SVM has been investigated and applied to the development of the proposed fraud detection model. The main concern of this research study is the application of SVM, namely SVC, for the classification of patterns (load consumptions profiles for customers) into two categories: normal (No) and fraud (Yes) customers. The contributions of the study of thesis research are listed as follows:

6.1 Contributions Of The Research

The main significance and benefits from the research study reported in this thesis are identified as follows.

- Analysis and mining in the business domain of managing customers water consumptions in Gaza municipality . As in [26] the researcher considers the analysis of a new business domain as a research contribute.
- Design and reveal a fraud detection model by using various data resources such as computerized historical water consumption data and a manual customer’s irregularities data registered on paper files.
- A novel, efficient and effective real world classifier model for water consumption profile fraud detection with experiments demonstrate the identification of relevant attributes and relevant periodical consumptions dataset that improve classification accuracy for data, which allows more information for fraudulent customers' profiles.

- The model cover 12 years of customer's historical water consumptions data with 720,000 training data records in which increases the fraud detection hit rate to 80% while the detection rate in [16] is just 60% with a drawback of use only two years customers load profile exported form 10 years customers' load profiles. So the model has better accuracy, better hit rate and larger training data set.

- Design an Intelligent classification model with about 80% fraud detection hit rate rather than human being effort with paper files to detect water thefts and Irregularities in a random successfully hit rate within 1 and 10%.

- The fraud detection model developed in this research study provides utility information tools to water utilities in Gaza city for efficient detection and classification of NTL and WOT activities in order to increase effectiveness of their onsite operation.

- With the implementation of the proposed model, operational costs for water utilities in MOG due to on-site inspection in monitoring NTL and WOT activities will be significantly reduced. This will also reduce the number of inspections carried out at random, resulting in higher fraud detection hit rate.

- Disseminate knowledge and behavior regarding fraudulent consumption patterns is obtained by the use of the proposed model, which is useful for further study and analysis by NTL experts and inspection teams in water utilities in Gaza.

- Lastly, by implementing the proposed fraud detection model, great time saving in detecting and identifying problematic mechanical meters can be achieved by MOG water utility.

6.2 Future Work.

- Day after day, customer irregularities increasing, so more customer irregularities will be collected to make bigger training data set for further testing and evaluation to increase the fraud detection hit rate and improve the proposed model accuracy with more customer data.

- The selected techniques work as a black box, without induce descriptive rules to show the attributes how indicate fraudulent behavior, so dealing with a rule based model is a strong idea to apply in the future work.

- Implement the classification model as a whole system to work in MOG at DWTC and be applicable to work in any municipality or any organization deal with citizen's water consumptions.

- This research study can be expanded by classifying load consumption patterns by respective districts, i.e. in Gaza strip like Nosyrat, Deirbalah…etc. However this approach requires more training data from each region, especially the load consumption patterns of the fraud customers.

- It is also recommended that the developed fraud detection model can be further tested and evaluated on other water distribution utilities in Gaza strip such as coastal municipal water utility.

This research study strongly believes that the fraud detection system can contribute significant improvement to water distribution utilities for the reduction of NTL activities.

7. REFERENCES.

1) AihuaShen, Rencheng Tong “Application of classification Models on credit card Fraud Detection”, 2007.

2) Anastassios Tagaris “Implementation of Prescription Fraud Detection Software Using REDBMS Tools and ATC Coding”, 2009.

3) C.-C. Chang and C.-J. Lin, 2005. LIBSVM: A Library for Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm. Accessed on January 28, 2008.

4) C.-W. Hsu, C.-C. Chang, and C.-J. Lin, “A Practical Guide to Support Vector Classification”, technical Report, department of computer science and Information engineering, national taiwan university, Taipei, 2003.

5) Daivid j. Hand "Data Mining : Statistics and More?",1998.

6) Daniel t. Larose "Discovering knowledge in data an introduction to data mining" , 2005.

7) David L. Olson DursunDelen "Advanced Data Mining Techniques".

8) Dianmin Yue, Xiaodan Wu, Yunfeng Wang, Yue Li "A Review of Data Mining-based Financial Fraud Detection Research", 2007.

9) Dongsong Zhang; Lina Zhou, "Discovering golden nuggets: data mining in financial application",2004.

10) D. Gerbec, Student Member, IEEE, S. Gašperič, Student Member, IEEE, I. Šmon, and F.Gubina, Member, IEEE "An Approach to Customers Daily Load Profile Determination", IEEE 2002.

11) D. S. Kirschen B. D. Pitt ,UMIST “Application of Data Mining Techniques to Load Profiling”, Manchester, UK 1999.

12) E. Byvatov, U. Fechner, J. Sadowski, and G. Schneider, “Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/Nondrug Classification”, Journal of Chemical Information and Computer Science, vol. 43, 2003.

13) E.W.T. Ngai a, Yong Hu b, Y.H. Wong a, Yijun Chen b, Xin Sun "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature",2011.

14) Huang Zhijun and Xia Chuangwen “A Kind of Algorithms for Euclidean Distence-Based outlier mining and its application to expressway toll fraud detection”,2009.

15) Ian H.Witten, Eibe Frank, Mark A. Hall "data Mining practical machine learning tools and techniques" Third edithon,2011.

16) Jawad Nagi, Keem Siah Yap, Sieh Kiong Tiong, Member, IEEE, Syed Khaleel Ahmed, Member, IEEE and Malik Mohamad “Non technical Loss Detection for Metered Customers in Power Utility Using Support Vector Machines”, 2010.

17) Jawad Nagi," An intelligent system for detection of non-technical losses inTenaga Nasional Berhad (TNB) malaysia low voltage distribution network" , master's dissertation, university tenaga national, malaysia, 2009.

18) J. Nagi, K.S. Yap, F. Nagi, S.K. Tiong, S. P. Koh, S. K. Ahmed, " NTL Detection of Electricity Theft and Abnormalities for Large Power Consumers In TNB Malaysia ",IEEE, 2010.

19) Jerome H. Friedman "Data mining and statistics : What's the connection?",1997.

20) Jiawei Han and Micheline Kamber" Data Mining: Concepts and Techniques Second Edition",2005.

21) José E. Cabral, Joao O. P. Pinto “Fraud Detection in High Voltage Electricity Consumers Using Data Mining”.2008.

22) J. R. Galvan, A. Elices, A. Munoz, T. Czernichow, and M. A. Sanz-Bobi, “System for Detection of Abnormalities and Fraud in Customer Consumption” in Proc. of the 12th Conference on the Electric Power Supply Industry, November 2-6, 1998, Pattaya, Thailand.

23) Konstantinos Tsiptsis, Antonios Chorianopoulos" Data Mining Techniques in CRM",2010.

24) Konstantinos Tsiptsis, AntoniosChorianopoulos, ”Data Mining Techniques In CRM”, 2009.

25) Lawrence O. Hall and Ajay Joshi " Building Accurate Classifiers from imbalanced Data Sets ".

26) Longbing Cao, Philip S. Yu, Chengqi Zhang, Huaifeng Zhang “Data Mining for Business Applications”, Springer 2009.

27) Mirjana PejiC Bach, ”Data Mining Applications Detection research Public organizations”, 2003.

28) Nongye, "The handbook of data mining". 2003.

29) Oded Z. Maimon, LiorRokach, “Data Mining with Decision Trees: Theroy and Applications“, 2008.

30) Ou Liu1, Jian Ma2, Pak-Lok Poon1, and Jun Zhang3," On an Ant Colony-Based Approach for Business Fraud Detection, IEEE, 2009.

31) Paolo Giudici , Silvia Figini, "Applied Data Mining For Business and industry", Second Edition, 2009.

32) Rapid Miner 5.1, http://www.rapidminer.com, (2012, October).

33) R. Behroozmand, and F. Almasganj, "Comparison of Neural Networks and Support Vector Machines Applied to Optimized Features Extracted from Patients' Speech Signal for Classification of Vocal Fold Inflammation", in Proc. of the 5th IEEE International Symposium on Signal Processing and Information Technology, 2005.

34) Sumana Sharma, Kweku-Muata Osei-Bryson "Framework for formal Implementation of the business understanding phase of data mining projects",2009.

35) Yuko Tachibana and Mikihiko Ohnari," Prediction Model of Hourly Water Consumption in Water Purification Plant through Categorical Approach " , IEEE, 1999.

36) http://en.wikipedia.org/wiki/Neural_network, (2012,November).

37) http://www.mogaza.org/?page=service, (2012, June).

38) http://en.wikipedia.org/wiki/KNN, (2012, November).

39) http://en.wikipedia.org/wiki/Support_vector_machine, (2012, June).

40) http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/glossary.html, (2012, July).

41) http://www.statsoft.com/textbook/statistics-glossary/l/?button=0#Linear Modeling (online text book, 2012 December).

42) http://en.wikipedia.org/wiki/Nonlinear_system (2012, December).

43) Glenn J. Myatt, “Making Sense of Data A Practical Guide to Exploratory Data analysis and Data Mining” , 2007.

44) WeiXuan Hu, "The Application of Artificial Neural Network in Wastewater Treatment ",IEEE, 2011.

45) Yuko Tachibana and Mikihiko Ohnari, " Prediction Model of Hourly Water Consumption in Water Purification Plant through Categorical Approach ", IEEE, 1999.

46) Dr. M V Krishna Rao, Mr. S H Miller, "Revenue Improvement From Intelligent Metering Systems ", IEEE, 1999.

47) Daniel Morariu, Lucian N. Vintan, and Volker Tresp, " Evaluating some Feature Selection Methods for an Improved SVM Classifier ", International Journal of Intelligent Technology Volume 1 Number 4, 2006.

48) http://en.wikipedia.org/wiki/Receiver_operating_characteristic, (2012, Dec).

49) Innocent E Davidson, “Evaluation And Effective Management Of Non-Technical Losses In Electrical Power Networks”, IEEE, 2002.

50) R. F. Chang and C. N. Lu, “Load Profiling and Its Applications in Power Market ”, IEEE, 2002.

51) Ishmael S. Msiza, Fulufhelo V. Nelwamondo and Tshilidzi “Artificial Neural Networks and Support Vector Machines for Water Demand Time Series Forecasting”, IEEE, 2007.

52) Zhi-HuaZhou, andXu-YingLiu, “Training Cost-Sensitive Neural Networkswith Methods Addressing the Class Imbalance Problem ”, IEEE,2006.

53) N. Japkowicz,“Learning from imbalanced datasets: a comparison of Various strategies,”in Working Notes of the AAAI’00 Workshop on Learning from Imbalanced DataSets, Austin,TX, pp.10–15,2000.

54) N. Japkowicz and S. Stephen,“The class imbalance problem:a systematic study, ”Intelligent Data Analysis, vol.6, no.5, pp.429–450, 2002.

55) C. Drummond and R.C. Holte, “C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling,” in Working Notes of the ICML’03 Workshop on Learning from Imbalanced DataSets, Washington, DC, 2003.

56) M.A. Maloof, “Learning when datasets are imbalanced and when costs Are unequal and unknown,”in Working Notes of the ICML’03Workshop On Learning from Imbalanced DataSets, Washington, DC, 2003.