

PRESCRIPTION FRAUD DETECTION VIA DATA MINING: A METHODOLOGY PROPOSAL

A THESIS

SUBMITTED TO THE DEPARTMENT OF INDUSTRIAL ENGINEERING

AND THE INSTITUTE OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

By

Karca Duru Aral

July, 2009

I certify that I have read this thesis and that in my opinion it is full adequate, in scope and in quality, as a dissertation for the degree of Master of Science.

Prof. İhsan Sabuncuoğlu (supervisor)

I certify that I have read this thesis and that in my opinion it is full adequate, in scope and in quality, as a dissertation for the degree of Master of Science.

Prof. Halil Altay Güvenir (co-supervisor)

I certify that I have read this thesis and that in my opinion it is full adequate, in scope and in quality, as a dissertation for the degree of Master of Science.

Assoc. Prof. Oya Ekin Karaşan

I certify that I have read this thesis and that in my opinion it is full adequate, in scope and in quality, as a dissertation for the degree of Master of Science.

Asst. Prof. Ayşegül Altın

Approved for the Institute of Engineering and Science

Prof. Mehmet Baray

Director of Institute of Engineering and Science

ABSTRACT

PRESCRIPTION FRAUD DETECTION VIA DATA MINING: A METHODOLOGY PROPOSAL

Karca Duru Aral

M.S. in Industrial Engineering

Advisors: Prof. İhsan Sabuncuoğlu, Prof. Halil Altay Güvenir

July, 2009

Fraud is the illegitimate act of violating regulations in order to gain personal profit. These kinds of violations are seen in many important areas including, healthcare, computer networks, credit card transactions and communications. Every year health care fraud causes considerable amount of losses to Social Security Agencies and Insurance Companies in many countries including Turkey and USA. This kind of crime is often seem victimless by the committers, nonetheless the fraudulent chain between pharmaceutical companies, health care providers, patients and pharmacies not only damage the health care system with the financial burden but also greatly hinders the health care system to provide legitimate patients with quality health care. One of the biggest issues related with health care fraud is the prescription fraud. This thesis aims to identify a data mining methodology in order to detect fraudulent prescriptions in a large prescription database, which is a task traditionally conducted by human experts. For this purpose, we have developed a customized data-mining model for the prescription fraud detection. We employ data mining methodologies for assigning a risk score to prescriptions regarding Prescribed Medicament- Diagnosis consistency, Prescribed

Medicaments' consistency within a prescription, Prescribed Medicament- Age and Sex consistency and Diagnosis- Cost consistency. Our proposed model has been tested on real world data. The results we obtained from our experimentations reveal that the proposed model works considerably well for the prescription fraud detection problem with a 77.4% true positive rate. We conclude that incorporating such a system in Social Security Agencies would radically decrease human-expert auditing costs and efficiency.

Keywords: Fraud Detection, Prescription Fraud, Data Mining

ÖZET

VERİ MADENCİLİĞİ TEKNİKLERİ İLE REÇETE USULSÜZLÜKLERİNİN TESPİTİ: BİR YÖNTEM ÖNERİSİ

Karca Duru Aral

Endüstri Mühendisliği Yüksek Lisans

Tez Yöneticisi: Prof. Dr. İhsan Sabuncuoğlu, Prof. Dr. Halil Altay Güvenir

Temmuz, 2009

Her yıl, sağlık, bankacılık, bilgi işlem ve iletişim gibi bir çok önemli alanda görülen usulsüz işlemler önemli miktarda para, zaman, bilgi ve emek kaybına sebep olmaktadır. Sağlık alanında görülen usulsüzlükler, aralarında Türkiye ve Amerika Birleşik Devletleri'nin de olduğu birçok ülkede sosyal güvenlik kurumları ve özel sağlık sigortası şirketlerine ciddi zararlar vermekte ve sağlık sistemlerini olumsuz etkilemektedir. Uygulayanlar ve uygulanmasına göz yumanlar tarafından zararsız olarak algılanan sağlık sistemi usulsüzlükleri, ilaç firmaları, sağlık hizmeti sağlayıcıları, hastalar ve eczaneler arasındaki yasadışı bir ağ üzerinden yürütülmektedir. Bu usulsüz faaliyetler sosyal güvenlik kurumlarına yalnızca finansal zararlar vermeye kalmayıp, sağlık sistemlerinin daha etkin ve kaliteli işleyebilmesinin önünde büyük bir engel oluşturmaktadır. Sağlık harcamalarının yarısına

yakınının ilaç harcamaları üzerine olduđu ÷lkemizde, reçete usulsüzlükleri de önemli bir sorun teşkil etmektedir.

Bu çalışma, veri madenciliğıyle büyük reçete veri tabanlarında usulsüzlük denetimi yapılması konusunda bir yöntem araştırmasını amaçlamıştır. Çalışmanın amacı, halihazırda uzmanlar tarafından rasgele seçim yoluyla yapılan reçete usulsüzlüğü denetiminin etkin bir otomasyon sistemiyle sağlanması için özelleştirilmiş veri madenciliğı teknikleri geliştirilmesidir.

Bu amaçla, her reçeteye İlaç-Tanı uyumu, İlaç-Yaş uyumu, İlaç-İlaç uyumu, İlaç-Cinsiyet uyumu ve Tani-Fiyat uyumuyla ilgili risk notları vermek üzere oluşturulan teknikler, sonrasında gerçek reçeteler kümesi üzerinde denenmiştir. Test kümesindeki usulsüz reçetelerin %77.4'ünü yakalayan sistem, reçete usulsüzlük denetimi açısından tatmin edici bulunmuş ve önerilen yöntemin sosyal güvenlik kurumları tarafından kullanılmasının uzman denetim masraflarının azaltılması ve denetim etkinliğinin artırması açısından uygun olabileceğı sonucuna varılmıştır.

Anahtar Kelimeler: Veri madenciliğı, usulsüzlük denetimi, Reçete Usulsüzlükleri

Acknowledgement

I would like to start by thanking my supervisors Prof. İhsan Sabuncuoğlu and Prof. Halil Altay Güvenir for sharing their experience and deep knowledge with me throughout the study for this thesis.

I send my greatest appreciations to my thesis jury for sharing their time and thoughts with me.

I appreciate the help of M.D. Çağdaş Baran in labeling a test set for performance evaluations of the proposed model.

From the beginning of this study I had one person in mind to dedicate it, my dear father, Temel Nusret Aral. Unfortunately, he had been diagnosed with stage-IV small-cell lung cancer by the time I started my graduate studies. So strong in character, he fought not only with cancer but also with chemo and radiotherapies for 14 months until he passed away on November 13, 2008. Being a creative, artistic architect, I believe him to be a regrettable early loss not only for my family, and me but also for the all human kind. Thank you daddy, for raising me up the way I am, for being my father... Your love and memory stays here with me.

All I wish for this study is for it to help reduce the burden on health care systems by identifying and hindering criminal acts. Thus, this study might lead to creating an extra

budget for spending on cancer research; therefore increasing the chances that small-cell lung cancer research receives some support since this area is largely ignored by authorities. Please note that small-cell lung cancer is the deadliest of all cancer types, nonetheless receives the least of funds since those who are diagnosed with it are thought to have chosen to have cancer by choosing to smoke. I hope one day humanity will find a way to defeat this most cruel illness.

I would like to thank my mother, Fadime Aral, who had been along with my father all through the way in this fight and being strong for him, for my sister and me. I also thank my sister, İmge Su Aral, who is a joy and had been a joy making us smile even in the worst of times.

I am deeply grateful to Mr. Emre Doğukaya for being a shoulder for me through this hard time. Most probably, I wouldn't have had the courage to pursue a master's degree if it wasn't for him.

I also want to thank Miss. Nazar Tüysüzoğlu for all her support and sincere consideration in the times we have spent together.

TABLE OF CONTENTS

Chapter 1: INTRODUCTION	1
Chapter 2: LITERATURE REVIEW	7
2.1. FRAUD DETECTION LITERATURE BY SUBJECT	8
2.2. AVAILABLE DATA FOR FRAUD DETECTION	10
2.3. INCORPORATED METHODOLOGIES	10
2.3.1. Supervised Approaches	11
2.3.2. Semi-Supervised Approaches	13
2.3.3. Unsupervised Approaches	14
2.3.4. Hybrid Approaches	14
2.4. OUTLIER DETECTION	16
2.5. HEALTH CARE FRAUD DETECTION	20
2.6. PERFORMANCE MEASURES	25
Chapter 3: MATHEMATICAL FORMULATIONS	27
3.1. DATA STRUCTURE	28
3.2. REVISED METHODOLOGIES	30
3.2.1. Frequent Item Set Mining/Association Rule Learning	30
3.2.2. Infrequent Itemset Mining	30
3.2.3. Clustering	31
3.3. METHODOLOGICAL DESIGN	33
3.4. RISK FORMULATIONS	35
3.4.1. Risk Formulation for Categorical Features	36

3.4.2.Risk Formulation for Ordered Features.....	39
Chapter 4: APPLICATION AND COMPUTATIONAL RESULTS.....	43
4.1.OFFLINE PROCESSING	44
4.2.ONLINE PROCESSING.....	46
4.3.OFFLINE FRAUD DETECTION RESULTS	48
4.4.ONLINE FRAUD DETECTION RESULTS.....	51
4.5.PERFORMANCE EVALUATION.....	55
Chapter 5: CONCLUDING REMARKS AND FURTHER RESEARCH DIRECTION	58
BIBLIOGRAPHY	61
APPENDIX	69
A. Sample Model Output.....	70

LIST OF FIGURES

2.4. Figure 1: Scatter Plot of a sample on x and y coordinates.....	18
3.3. Figure 2: Flow chart of the integrated offline and online systems	34
3.4. Figure 3: Examples of computational effectiveness of the risk formulation.....	37
4.2. Figure 4: Prescription Auditing Tool User Interface.....	47
4.4. Figure 5: Inserting a Prescription to the Prescription Auditing Tool	52
4.4. Figure 6: Validation Message Box	53
4.4. Figure 7: Database Update Notification	53
4.4. Figure 8: Riskiness Levels Screen.....	54

LIST OF TABLES

1. Table 1: Health Care Spending In Turkey by years	3
4.3. Teble 2: Prescription Example-1	49
4.3. Table 3: Prescription Example-2	50

In the loving memory of my father...

*In the cross-roads of universe and time,
I met a blessed soul,
I feel blessed that he was my father.*

Chapter 1

INTRODUCTION

Fraud is the abuse of a profit organization's system without necessarily leading to direct legal consequences [1]. Fraud constitutes a critical problem in many areas like health care, banking, insurance, and telecommunications. The fraudulent minority creates a big burden to the society to finance the fraudulent transactions. Any effort aiming to debug the fraudulent transactions in the above-mentioned businesses and probably in many other ones, is named as a fraud detection process. Due to the complexity and enormity of the modern business systems, criminals may and do discover safety gaps and use them to steal data or to defraud somebody. Even if a fraud type is discovered by the authorities and safety regulations are managed, the criminals seek and find other fraudulent ways and thus shift behavior over time. Manual detection conducted by human experts is very expensive even to debug any fraud that has been committed; can't detect all fraudulent transactions of a certain type; can't be managed to detect the fraudulent behavior the moment it is attempted to be committed and lack the ability to detect the shifts and trends in fraudulent behavior.

If we are to classify the fraudsters abusing an organization, according to their nature, we see that a business can be swindled by its managers, its employees or by the third parties. These external third parties are generalized by three types as organized, criminal, and average [1].

Average fraudsters are those who are not a part of an organized crime group, and have a tendency to commit fraudulent acts in an occasional manner. Even though these types of fraudsters are risky enough to be detected, organized or individual criminal fraudsters are more likely to cause more harm to the business system that is affected. These kinds of fraudsters are committing their fraudulent acts in an organized manner, often involved with identity theft and change behavior over time to get through the detection systems and new regulations. Considering the large businesses, it is highly costly to manually check all transactions and activities. So, it can be said that the enormous databases of these large businesses should be detected by customized data mining algorithms, and then the riskiest transactions identified can be inspected by human experts.

Fraud can also be grouped to be application or transaction fraud. In the application case, identity theft of falsified identity information is involved, whereas in the transactional case, a legitimate user/account information is abused by criminals.

We can summarize the problems involved with fraud detection as below [3]:

- Class distributions meaning the proportions between illegitimate transactions and legitimate transactions fluctuate.
- Different types of fraud can affect a business.
- Different styles of fraud have different behavioral characteristics in nature like being a one-time crime, being seasonal or being occasional.
- These characteristics can shift by time.
- Fraudsters change behavior to get through any new detection system and modify fraud styles.

Fraud detection, being part of the overall fraud control, should automate and help to reduce the manual parts of a screening/checking process [1]. For large businesses, it is intuitive that data mining incorporated systems are one of the best tools for fraud detection if

not the only; since large business generate large databases on which human auditing is inefficient.

Data mining is about finding insights which are statistically reliable, previously unknown, and actionable from data [2]. This data must be available, relevant, adequate, and clean. Also, the data mining problem must be well-defined, cannot be solved by query and reporting tools, and guided by a data mining process model [4].

Health care systems are among the largest of business systems in many countries. Being a business where enormous amounts of money cycles through, health care systems are very attractive targets for the above mentioned types of fraudsters. Below table illustrates the dimensions of the health care system in Turkey [5]:

(billion YTL)	2002	2007	2008
Total Social Insurance Spending	7.6	20	24
Total Medicament Spending	4.3	8.6	10.5
Total Hospital Spending	2.8	10.3	13
State Hospital Payments by SGK	1.8	6.4	7.5

Table 1: Health Care Spending In Turkey by years

According to Turkish Health Care Syndicate 2008 Health Care Report, fraud in health care has boomed recently. Having seen a yearly exponential increase in spending, health care systems' abuse is becoming more and more critical. In 2008, health care fraud was committed principally in Van, Eskişehir, Erzurum, Siirt, Adana, Bursa, Zonguldak, Diyarbakır, and many other cities even in the Head Center of the Tuberculosis Fighting Department. These fraudulent acts were in the form of fake medicament reports, fake invoices, billing Social Security Agency (SSA) for examinations, and treatments that were not

rendered. The total cost of these fraudulent acts being millions of TL, about 300 people were arrested regarding fraud charges [5].

According to General Accounting Office of the USA, annual health care expenditures in USA have approached two trillion dollars, which is 15.3% of the Gross Domestic Product by 2007 [56].

The optimistic estimates are that at least 3% of health care expenditures which adds up to be \$60 billion are lost due to fraud in USA. Other estimates are around 10% or \$170 billion for this lost amount [57].

Fraud and abuse are not only widespread and very costly in United States' and Turkey's health-care systems but also are very destructive in many other countries. Examples for fraud in a healthcare system would be billing for services and goods that are not rendered, performing medically unnecessary operations or prescribing unnecessary medicaments. Abuse involves charging for services that are not medically necessary, that do not conform to professionally recognized standards, or are unfairly priced. Some examples for abusive behaviors would be performing a laboratory test on large numbers of patients when only a few should have it or x-raying those without the definite need. Abuse and fraud are similar; nonetheless it is not possible to prove that the abusive acts were done with intent to deceive the insurer.

Prescription fraud is one of the types of health care fraud that has been commonplace in Turkey and constitutes an enormous burden on the Social Security Agency and the private insurance companies. This type of fraud compromises of excessive medicament prescription, and disunity of patients' features with the prescribed medicaments. The perception of the society that the prescription fraud is a victimless crime make it even more widespread and strengthen the fraudulent chain between the pharmaceutical companies, physicians, pharmacies, and patients. The orthodox manual detection is conducted by a committee of assigned medical doctors in the Social Security Agency. When inspecting a hospital, the

human expert goes through a small sample of the prescriptions associated by the hospital and then the agency charges the hospital by paying the amount acquired by multiplying the proportion of the fraudulent claims seen in the sample and the total cost of the prescriptions issued by the hospital in that inspection era. This method is both costly to conduct and does not guarantee any efficiency coefficient.

When considering the immense amount of data associated with the health care system, it is trivial that any system dealing with prescription fraud should be automated and fail-safe to a considerable degree. Since nearly half of the spending of the SSA is on the medicament which was around 10.5 billion TL in 2008 [5], we see that the cost of the fraudulent prescriptions to the Social Security Agency is not tolerable. Thus, any system should be able to find the prescriptions that constitute a certain fraud probability assessed by the user. This probability should be such that the system functions with minimum amount of false negatives, that is to say minimum amount of fraudulent prescriptions being left undetected. This probability coefficient should be determined considering the human expert revising necessary for the output of the system. That way, the cost efficiency should be maintained.

Having revised the necessities of a cost effective system, we can conclude that such a system should incorporate the appropriate data mining methodologies enabling an automated, rapid, and efficient online structure that can be integrated with the electronic online provision systems already in use.

Our proposition for such a system is based on certain risk measurements calculated for each of the patient's features compared to the common practice. We also propose to detect each pair of medicaments in the prescription since a pair may be in contraindication. An alike risk measurement is taken for each medicament pair in the prescriptions. The risk measurement is based on the assumption that, fraudulent behavior related to a certain feature is rare when considering the total data set. The data set we work on is a set of 26,419 real world prescriptions.

Next chapter is on the literature review conducted for both general fraud detection, health care fraud detection methodologies, and outlier detection, which is largely employed for fraud detection. This survey indicates that there are three main types of fraud detection techniques proposed for health care. These are supervised, unsupervised, and hybrid systems of two. Since we work on a data set without any prior knowledge on prescriptions' label to be fraudulent or not, our proposed system is an unsupervised one.

The third chapter focuses on the data structure in hand, the revised methodologies in the literature to see if those are applicable to our problem, our methodology proposal, and the related risk formulations.

The forth chapter gathers the application outputs and the computational results. We briefly go over the offline and online processing as well as their outputs. We give the results of applications on the real world data with an emphasis on the offline and online applications. The validations of the system's efficiency are conducted by validation study comparing the system outputs with those of a human expert labels.

We will conclude by a revision of the proposed system and further study directions.

Chapter 2

LITERATURE REVIEW

There are various resources relating to fraud detection. Fraud detection being a relatively large field, most of the papers on this subject consider outlier detection as a primary tool. Nonetheless, health care fraud detection studies are limited. When we come to the more specific field of prescription fraud detection, we see that there is no other study in this particular field. In this chapter, we focus on fraud detection, outlier detection and health care fraud detection studies in the literature.

As stated earlier, fraud detection automates transaction investigation efforts. In this regard, evolved and customized algorithms for the data in hand to check are the best possible answer for this business critical problem. We see that, the studies in this field mainly comprise of artificial intelligence, data mining, expert systems, fuzzy logic, statistics and visualization. The main shortcomings of data mining-based fraud detection research are that the lack of publicly available real data for experimentations and the lack of published methods. Even though studies continue for more effective solutions, there are commercially available data mining software with the claim to be competent to detect fraud in many sector specific cases.

2.1. FRAUD DETECTION LITERATURE BY SUBJECT

When we group the studies regarding fraud detection, we see that most of the studies group around administrative fraud detection, credit card fraud detection, telecommunications fraud detection and insurance fraud detection.

Internal fraud meaning the loss due to acts of a type intended to defraud, misappropriate property or circumvent regulations, the law or company policy, excluding diversity / discrimination events, which involves at least one internal party [6]. This type of fraud being stated to be one of the operational risks by the Basel Committee is a big problem involving accounting, financial statement and occupational fraud. There are studies in the literature to pinpoint internal fraud by Lin *et al.*, (2003) proposing a Fuzzy Neural Network for Assessing the Risk of Fraudulent Financial Reporting; by Bell and Carcello, (2000); by Fanning and Cogger, (1995) proposing a neural network approach; by Summers and Sweeney, (1998) focusing on an empirical analysis on misstated financial statements; by Beneish, (1997) proposing a model providing assessments of the likelihood of manipulation in financial reports; by Green and Choi, (1997) proposing another neural network for assessing the risk of management fraud. Kim *et al.*, (2003) focuses on an anomaly detection approach for fraud detection in retail sector. For this, implementing features of the human immune system is proposed.

When we consider the insurance fraud; home insurance is a field studied by Bentley, (2000), proposing fuzzy rules and by Von Altrock, (1997), proposing fuzzy logic. Crop insurance is studied by Little *et al.*, (2002) applying data mining methodologies. Automobile insurance is studied by Phua *et al.*, (2004) in which a classification of skewed data is made; by Viaene *et al.*, (2004) proposing implementing Naive Bayes for claim fraud diagnosis; by Brockett *et al.*, (2002) proposing Principal Component Analysis for fraud classification; by Stefano and Gisella, (2001) proposing a fuzzy expert system; by Belhadji *et al.*, (2000) proposing a system based on the systematic use of fraud indicators; and by Artis *et al.*, (1999) on modeling types of automobile insurance fraud behavior.

For medical/health care insurance fraud detection, there are a number of studies in the literature. Yamanishi *et al.*, (2004) propose applying finite mixtures. Major and Riedinger, (2002) define a hybrid knowledge and statistical-based system named as EFD as in Electronic Fraud Detection. Williams, (1999) proposes evolutionary hot spots data mining. He *et al.*, (1999) incorporates a hybrid system of genetic algorithms and k-nearest neighbor method. Cox, (1995) proposes a fuzzy system. Yang, W. and Hwang S. (2006) propose a process-mining framework for health care fraud detection. Ortega *et al.*, (2007) incorporate a data mining methodology on multilayer perceptron neural networks.

There are two types of credit fraud detection, one is on screening credit applications and the other is the credit card transactions. Wheeler and Aitken, (2000) describe an application of case-based reasoning for eliminating fraud in credit approval process. For credit card transactions there are studies by Fan, (2004); Chen *et al.*, (2004); Chiu and Tsai, (2004); Foster and Stine, (2004); Kim and Kim, (2002); Maes *et al.*, (2002); Syeda *et al.*, (2002); Bolton and Hand, (2001); Bentley *et al.*, (2000); Brause *et al.*, (1999); Chan *et al.*, (1999). They propose systematic data selection, support vector machines, a web-based scheme, predictive modeling, neural classifiers, Bayesian and neural networks, parallel granular neural networks, unsupervised profiling methods, fuzzy Darwinian detection, neural data mining, and distributed data mining, respectively.

For telecommunications fraud, there is subscription abuse and phone call abuse. Cortes *et al.*, (2003); Cahill *et al.*, (2002); Rosset *et al.*, (1999), have worked on subscription fraud, proposing dynamic graphs, data mining, rule based induction, respectively. Kim *et al.*, (2003); Burge and Shawe-Taylor, (2001); Moreau *et al.*, (1999); Murad and Pinkas, (1999) have studied the phone- call fraud problems. These studies focus on support vector machines, unsupervised neural network, hybrid systems and unsupervised profiling, respectively.

Other less studied fraud detection problems are e-business, government tax and customs' fraud. Barse *et al.*, (2003) and McGibney and Hearne, (2003) focus on video websites and voice-over-ip telecommunications fraud, respectively. Bhargava *et al.*, (2003) and

Sherman, (2002) propose methodologies to monitor online sellers and online buyers, respectively, by automated systems. Bonchi *et al.*, (1999) studied tax fraud. Shao *et al.*, (2002) has worked on customs' fraud.

2.2. AVAILABLE DATA FOR FRAUD DETECTION

The literature survey by Phua *et al.*, 2004, reveals that publicly available data for fraud detection is limited to a small automobile insurance database screened by the same author. This survey indicates that telecommunications and credit fraud detection are the domains where large databases with many attributes can be found. Whereas for insurance and internal fraud, studied databases are limited. There are even studies on 100 examples available. Nonetheless, attribute numbers for the insurance and internal fraud studies can be as high as 150. The paper with largest database on insurance fraud is by Williams, 1999, with 40000 examples [1,24].

The employed attributes in the literature are either binary, numerical, categorical or a combination of those. The attributes for medical insurance databases are patient demographics (age and sex), treatment details (services), and policy and claim details (benefits and amount) [1].

Data mining methodologies in the literature either use training data with fraud/legitimate labels, examples of legal transactions or data with no labels to indicate fraud or legitimacy.

2.3. INCORPORATED METHODOLOGIES

We can group the existing methodologies of fraud detection as being supervised, unsupervised, or as being hybrids of the above. These data mining methodologies are described as:

2.3.1. Supervised Approaches

Supervised algorithms are trained by previously labeled training set of fraudulent and legitimate transactions. Then, the algorithms allocate mathematical methodologies to assign scores of similarity with the fraudulent profiles. The most popular applications of supervised algorithms are Neural networks and support vector machines (SVMs).

Kim *et al.* (2003) define SVM ensembles with for telecommunications subscription fraud. Barse *et al.* (2003) propose a multi-layer neural network to handle synthetic database of Video-on-Demand. For credit card fraud detection Syeda *et al.* (2002) propose fast rule generation by fuzzy neural networks on parallel machines. A feed-forward Radial Basis Function (RBF) neural network with three-layers was proposed by Ghosh and Reilly (1994). This neural network was trained in two phases. It was used to assigning risk scores to new credit card transactions in every two hours.

Maes *et al.* (2002) conducted a comparison study between neural networks and Bayesian networks. This study incorporates the STAGE algorithm for Bayesian networks. Back propagation algorithm was used to train the neural networks. The results indicate that even though Bayesian networks are more accurate and needs a short training time, they are slower to be applied for new instances. Such a Bayesian Network was developed by Ezawa and Norton (1996), which has four stages and two parameters. This paper asserts that all the methods of regression, nearest neighbor, and neural networks are too slow for their data in hand. Decision trees were also problematic with some discrete variables in the dataset. Viaene *et al.* (2004) propose AdaBoosted naive Bayes (fully independent boosted Bayesian network) with weight of evidence formulation for scoring. When compared with unboosted and boosted naive Bayes, the proposed method had slightly better accuracy.

Some other methodologies are decision trees, rule induction, and case-based reasoning. Fan (2004) introduced systematic data selection to mine concept-drifting, possibly insufficient, data streams. The paper proposed a framework to select the optimal model from

four different models (based on old data chunk only, new data chunk only, new data chunk with selected old data, and old and new data chunks). The selected old data is the examples which both optimal models at the consecutive time steps predict correctly. The cross validated decision tree ensemble is consistently better than all other decision tree classifiers and weighted averaging ensembles under all concept-drifting data chunk sizes, especially when the new data chunk size of the credit card transactions are small. With the same credit card data as Fan (2004), Wang *et al.* (2003) demonstrates a pruned classifier C4.5 ensemble which is derived by weighting each base classifier according to its expected benefits and then averaging their outputs. The authors show that the ensemble will most likely perform better than a single classifier which uses exponential weighted average to emphasize more influence on recent data.

Rosset *et al.* (1999) presents a two-stage rules-based fraud detection system which first involves generating rules using a modified C4.5 algorithm. Next, it involves sorting rules based on accuracy of customer level rules, and selecting rules based on coverage of fraud of customer rules and difference between behavioral level rules. It was applied to a telecommunications subscription fraud. Bonchi *et al.* (1999) use boosted C5.0 algorithm on tax declarations of companies. Shao *et al.* (2002) apply a variant of C4.5 for customs fraud detection. Case-based reasoning (CBR) was used by Wheeler and Aitken (2000) to analyze the hardest cases which have been misclassified by existing methods and techniques. Retrieval is performed by threshold nearest neighbor matching. Diagnosis utilize multiple selection criteria (probabilistic curve, best match, negative selection, density selection, and default) and resolution strategies (sequential resolution-default, best guess, and combined confidence), which analyze the retrieved cases. The authors claim that CBR had 20% higher true positive and true negative rates than common algorithms on credit applications.

As for the statistical modeling, Foster and Stine (2004) employ least squares regression and stepwise selection of predictors. They assert that traditional statistical methods are effective to be used for fraud detection. Belhadji *et al.* (2000) propose the cooperation of human experts for choosing best indicators (attributes) for fraud detection. Then, they

calculate conditional probabilities of fraud for each indicator. As the third step, Probit regressions are used to define the most important indicators. Probit regression is used for fraud prediction on automobile property damages. The flexible thresholds are adjustable for customization regarding any company's fraud policy. In another study on automobile insurance data, Artis *et al.* (1999) make a comparison between multinomial logit model (MNL) and nested multinomial logit model (NMNL) on a classification problem. Both models provide estimated conditional probabilities for the three classes.

Some other techniques are expert systems, association rules, and genetic algorithms. Nonetheless the papers on the above mentioned techniques do not make efficiency or effectiveness comparisons with any existing techniques. Major and Riedinger (2002) have created an expert system to detect medical insurance fraud in which expert knowledge is integrated with statistical techniques. Pathak *et al.* (2003), Stefano and Gisella (2001) have studied fuzzy expert systems. Chiu and Tsai (2004) define Fraud Patterns Mining (FPM) algorithm, transformed from *Apriori*, in order to mine credit card data. Bentley (2000) proposes genetic programming with fuzzy logic for data classification on real home insurance claims and credit card transaction data.

2.3.2. Semi-supervised Approaches

Kim *et al.* (2003) present a five-step fraud detection method. First, rules are generated randomly by association rules algorithm *Apriori*; then rules are applied on legitimate labeled transaction database and any rule matching this data is eliminated. Third, the rules that are left are used for screening. Any rules, which cannot detect any anomalies, are eliminated. Fourth, any rule that can detect anomalies are refined by tiny random mutations. In the last step, the successful rules are retained. This proposed methodology is tested on retail transaction processing system internal fraud data.

For telecommunications fraud detection, Murad and Pinkas (1999) employ profiling. Profiling is attained by summarizing the calls daily and by overall levels of normal behavior

of each account. Clustering algorithm with cumulative distribution distance function is used to define the common daily profiles. When the profile's call duration, destination, and quantity exceed the threshold and standard deviation of the profile an alert is raised.

2.3.3. Unsupervised Approaches

In the area of telecommunications fraud detection, Cortes *et al.* (2001) study temporal evolution of large dynamic graphs. The graphs are built up by the sub graphs named as Communities of Interest (COI). Exponential weighted average method is used to update sub graphs daily. COIs are built up by the mobile phone accounts using call quantity and durations. The study had revealed specifications of the telecommunication fraudsters. Burge and Shawe and Taylor (2001) use a recurrent neural network to identify account behavior profiles.

In medical insurance domain, Yamanishi *et al.* (2004) presented the unsupervised SmartSifter. The algorithm can work with categorical and continuous variables. SmartSifter investigates statistical outliers by Hellinger distance. On automobile insurance data, Brockett *et al.* (2002) propose employing Principal Component Analysis of RIDIT scores on rank-ordered categorical attributes.

In credit card transactions, Bolton and Hand (2001) present Peer Group Analysis in screening inter-account behavior changes by time by comparing the cumulative weekly mean amount between the account in question and the similar accounts. Bolton and Hand (2001) present Break Point Analysis to screen intra-account behavior changing over time. This method is used to detect any significant peaks in spending of an account. The *t*-test is used to rank the accounts.

2.3.4. Hybrid Approaches

Supervised Hybrids

There are studies in the literature integrating supervised algorithms like neural networks, Bayesian networks, and decision trees. Chan *et al.* (1999) try to combine naive Bayes, C4.5, CART, and RIPPER classifiers. The results give better efficiency on credit card transactions. Phua *et al.* (2004) propose back propagation neural networks, naive Bayes, and C4.5 as classifiers. A single meta-classifier is used to identify the best base classifier among those, and then integrate the base classifiers' predictions on automobile insurance claims. Ormerod *et al.* (2003) use a rule generator to adjust the weights of the proposed Bayesian network. Kim and Kim (2002) define a decision tree algorithm to classify the data in hand. They use a weighting function to compute fraud density, and then a back propagation neural network is used to generate a weighted risk score on credit card transactions. He *et al.* (1999) use genetic algorithms to compute optimal weights of the attributes, then the k -nearest neighbor algorithm is employed to classify the general practitioner (GP) dataset.

Supervised/Unsupervised Hybrids

Labeled data is used for supervised and unsupervised hybrids in telecommunications fraud detection. Cortes and Pregibon (2001) propose the use of daily updated telecommunication account summaries (signatures). The fraudulent labeled signatures are then added to the training set. This set is then used for training the supervised algorithms such as tree, slipper, and model-averaged regression. The authors assert that fraudulent calls have nature of late night activity and long call durations. Cortes *et al.* (2003) propose a graph-theoretic method. This method is used to visually detect fraudulent international calls. Cahill *et al.* (2002) compute a risk score to each call regarding its similarity to fraudulent profiles and dissimilarity to the account's signature. The signatures are updated with low-score calls. In this updating process, recent calls are given more weight than older calls.

Moreau *et al.* (1999) indicate that supervised neural network and rule induction algorithms perform better than two types of unsupervised neural networks in identifying the shifts between short and long term account behavior profiles. They used AUC as the performance measure.

There are studies in which unsupervised approaches are used to classify the insurance data into clusters for incorporating supervised approaches. A three step procedure is proposed by Williams and Huang (1997) in which: k -means is employed for cluster detection, C4.5 is used for decision tree rule induction, and domain knowledge, then statistical summaries and visualization tools are utilized for rule evaluation. Williams (1999) employs a genetic algorithm for the second step to generate rules. This enables the user to explore the rules. in automobile injury claims, Brockett *et al.* (1998) propose a technique using Self Organizing Maps (SOM) for clustering just before employing back propagation neural networks

On medical providers' claims, He *et al.* (1997), use hybrids of back propagation neural networks and SOMs in order to screen the classification results. Brause *et al.*, (1999) present RBF neural networks for screening the outputs of association rules for credit card transactions.

2.4. OUTLIER DETECTION

Outlier detection methods have been suggested for numerous applications, such as credit card fraud detection, clinical trials, voting irregularity analysis, data cleansing, network intrusion, severe weather prediction, geographic information systems, athlete performance analysis, and other data mining tasks [60].

Outlier detection methods in the literature are:

- Univariate methods
- Multivariate methods

Another classification of outlier detection methods is between:

- Parametric (statistical) methods and
- Nonparametric methods that are model-free

Assuming a known underlying distribution of the observations, statistical parametric methods are based on statistical estimates of unknown distribution parameters. Outliers are defined to be the observations that deviate from the model assumptions by statistical methods. Papadimitriou *et al.*, (2002) state that statistical parametric methods are often unsuitable for high-dimensional data sets and for arbitrary data sets without prior knowledge of the underlying data distribution.

Data mining methods are in the class of non-parametric outlier detection methods. These methods are called to be *distance-based*. Being able to work on large databases, these methods mostly incorporate local distance measures [62,63,64,65,66,55]

Clustering techniques constitute another class of non-parametric outlier detection methods. Points in small sized clusters are labeled to be outliers.

Detection methods for spatial outliers compose another class of non-parametric outlier detection methodologies. Such methods look for locally deviating instances regarding the neighbor observations.

Earlier studies in univariate outlier detection literature are built on the assumption of the data being identically and independently distributed from a known distribution [60] Another assumption under which many other works are conducted is that distribution parameters and the type of the expected outliers are known as well. It is trivial that these assumptions do not hold for real world data in general.

For observations with multivariable attributes, multivariate analysis is performed. This enables to identify the interactions among different variables. Ben Gal, I. (2005) gives a simple example as illustrated in Figure 1 for this need. This is an example of points plotted on two dimensions on x and y axis. It is intuitive that there is one outlier in this case, which is a multivariate outlier but not univariate. If we consider each attribute (x and y coordinates)

separately, the outlier point is not an outlier in either of the measures. Therefore, outlier detection should consider the relationships between any two attributes.

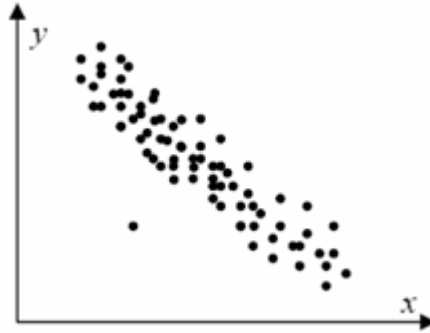


Figure 1: Scatter Plot of a sample on x and y coordinates

Thus, multivariate outlier detection procedures can either be statistical or data mining based. As stated earlier, statistical methods are based on the assumption of known distributions. The statistical methods of multivariate outlier detection try to identify the outliers as being the points that are stated being away from the center of the data distribution. The measurement of this distance can be done by a number of different distance measures. The effectiveness of the outlier detection procedure can be increased by incorporating robust estimates as in one-dimensional procedures. The most prevalent estimates are the distribution mean and the variance-covariance [67].

As stated earlier, data mining methodologies are mostly non-parametric, not needing the assumption of prior knowledge of the data distribution model. Data-mining methodologies are formed to handle large databases that are multi-dimensional. The sub groups of this category are:

- Distance-based methods,
- Clustering methods,
- Spatial methods.

As a non-parametric data mining methodology, distance-based methods were first proposed by Knorr and Ng (1997, 1998). The first definition of a distance based outlier is that being so if the observation in question is placed at a distance r from at least a β percent of the rest of the observations. Acuna and Rodriguez (2004) underline that this definition is problematic since it necessitates the determination of r and β and a ranking for the outliers. The time complexity of the algorithm based on this definition is $O(pn^2)$ where number of the attributes (features) is p and n is the size of the sample. Therefore, it is impractical to be used in large datasets. Another problem of this definition is that, if the data set has dense and sparse regions, definition turns to be inadequate as stated in Ramaswamy *et al.*, (2000) and Papadimitriou *et al.*, (2002).

Another definition of outliers is presented by Ramaswamy *et al.*, (2000). This definition can be stated as: given two integers v and l ($v < l$), outliers are defined to be the top l sorted observations having the largest distance to their v^{th} nearest neighbor. A critique of this latter definition is that the information on the closer neighbors is lost when considering only the v^{th} nearest neighbor. Another way to overcome this drawback is to define outliers to be the observations of which the *average distance* to the v^{th} nearest neighbors is large. Acuna and Rodriguez, 2004 points out that the drawback of this latter definition to be the longer computational time needed.

As for clustering based methods, these methods define small-sized clusters including a cluster of one observation to be outlier. *Partitioning around medoids* (PAM), *Clustering large applications* (CLARA) by Kaufman and Rousseeuw, (1990); and a *fractal-dimension* based method by Barbara and Chen, (2000) are among the examples for such methods. We need to underline that designed for clustering; these methodologies aren't really for outlier detection. As stated in Papadimitriou *et al.* (2002), the outlier detection criteria are usually implicit and clustering procedures do not easily convey these criteria.

The next group of non-parametric methodologies being spatial methods, these are relevant with the above-mentioned clustering methods. This methodology's applications are mostly seen in domains that convey spatial information, like, ecology, geographic information systems, transportation, climatology, location-based services, public health and public safety.

A spatial outlier is a spatially referenced object whose non-spatial attribute values are significantly different from the values of its neighborhood [72]. The authors classify the spatial methodologies by two groups of *quantitative tests* and *graphic approaches*.

The first group of methods presents tests to identify spatial outliers from the rest of database. Two representative approaches in this category are the Scatter plot by Haining, (1993) is an example of this group of methods.

Graphic methods employ visualization of spatial data, highlighting the outliers. Haslett *et al.*, 1991, Panatier, 1996 propose Variogram clouds and pocket plots for graphic outlier detection, respectively. Multidimensional scaling (MDS) was presented by Schiffman *et al.* (1981). Like in a map, MDS illustrates the analogy of observations. Penny and Jolliffe (2001) present metric and non-metric MDS reformulations.

The method for detecting spatial outliers in graph data set by Shekhar *et al.* (2001, 2002) employs the distribution property of the difference between the value of an attribute and the average attribute value of the neighbors. Shekhar, *et al.*, (2003) present an approach to make a comparison between spatial outlier-detection methods.

2.5. HEALTH CARE FRAUD DETECTION

There are three groups of health care fraud detection studies. The first group uses supervised methodologies. The second group incorporates unsupervised algorithms. The third group uses multiple methodologies of one or both of the first two.

Predictive supervised algorithms examine all previously labeled transactions to mathematically determine how a standard fraudulent transaction looks like by assigning a risk score [1]. As a supervised fraud detection method, neural networks seem to be the most popular ones. Here are the supervised fraud detection studies in the literature:

Ormerod *et al.* (2003) propose the usage of a Bayesian Network. They present a Mass Detection Tool (MDT) for detection of medical insurance fraud. Ethnography is the core element of the algorithm for specifying needs and process, capture expertise, and design an interface for triggering fraud indicators while capturing unexpected anomalies detected by claims handlers. The MDT uses a dynamic Bayesian Belief Network of fraud indicators, The system employs automated knowledge updating to keep up with dynamically changing fraud, adding new indicators that emerge from patterns of repeated anomalies [52].

Chan CL *et al.* (2001) introduces a Fuzzy Bayesian Classifier. This research combines the Bayesian classifier and the Fuzzy Set Theory to create a new data mining methodology. Bayesian classifier, based on Bayesian inference, is one of the data mining techniques that can be used for classification problems. Bayesian classifier classifies by incorporating all features influencing the classification result. The Bayesian classifier, having a good power of interpretation of the result, needs to associate different probability distributions when dealing with continuous attributes, which increases the complexity of the computation. To overcome this, Fuzzy set theory is exercised to transform the continuous attributes into discrete ones. This system is then used in analyzing health insurance fee data. 80% of the data set was used to train the Fuzzy Bayesian Classifier and then, the system was tested on the 20% of the data. The true positive rate (sensitivity) of the classifier is 0.639 and the true negative rate (specificity) is 0.968 [53].

Ortega *et al.* (2007) describes another medical claim fraud/abuse detection system based on data mining used by a Chilean private health insurance company. The proposed detection system employs multilayer perceptron neural networks (MLP). The entities involved in the medical fraud problem are as: medical claims, affiliates, medical professionals

and employers. The multilayer perceptron neural network is trained for the features of all these entities. The proposed fraud detection system is shown to detect 75 fraudulent cases per month [58].

Unsupervised algorithms are relatively less incorporated as a health care fraud detection tool. There were two examples seen in the literature survey in the process of making this thesis.

The first one is the Electronic Fraud Detection (EFD) proposed by Major *et al.* (2002) which introduces the usage of an expert system. Electronic Fraud Detection (EFD) is designed to assist the Investigative Consultants in the Managed Care & Employee Benefits Security Unit of the Travelers Insurance Companies in the detection of frauds committed by health care providers [23]. The database that EFD is designed for has never been investigated manually and it has few positive examples. In order to get through these problems, EFD incorporates two levels of knowledge discovery techniques. In the first level, in order to highlight the unusual provider behavior, EFD incorporates expert knowledge and statistical information assessment. The 27 behavioral heuristics employed in EFD are used to screen and to measure the provider behavior in question. The rules seek to identify providers which seem to deserve a human expert investigation. Then, new rules are built by machine learning in order to enhance the screening efforts. Pilot operations had been carried out to analyze 22,000 health care providers. Then a prototype system got implemented in SAS Institute's SAS System, AICorp's Knowledge Base Management System, and Borland International's Turbo Prolog.

The second unsupervised methodology is proposed by Yamanishi *et al.* (2004) and named as SmartSifter which is an outlier detection engine addressing the problem from the viewpoint of statistical learning theory. The proposed methodology, SmartSifter works online to identify outliers, incorporating the online unsupervised training of a finite mixture model on the information in hand. Every time there is a new entry to the system, SmartSifter runs to

learn the new probabilistic model. The output of the SmartSifter is score given for the new entry. In the case of a high score, the new entry is said to have a high risk to be an outlier. The superiority of SmartSifter are identified to be it being adaptive to changing sources of data; its output being a score which has a easily understandable meaning; it being computationally inexpensive; and it being able to work with both categorical and continuous variables. SmartSifter's experimental applications have been shown to identify meaningful rare cases in real-life health insurance pathology data from Australia's Health Insurance Commission [22].

Hybrid methods consist of a combination of supervised and unsupervised methods or concatenating two or more methods in one of the supervised/unsupervised groups.

Williams *et al.* (1999) present the hot spots methodology. The proposed methodology incorporates a multi-strategy in an interactive approach to identify important nuggets. First, the methodology employs data mining. Then, the system screens the outcoming models in order to identify the important nugget. The system is then used on insurance and fraud applications[24].

He, H. *et al.* (1999) studied the medical fraud detection problem regarding the General Practitioners (GP). The features to classify GP profiles are weighted by genetic algorithms. Then, these weights are imposed in K-Nearest Neighbor algorithm in order to detect practice profiles. Then, the practice profiles are classified by the Majority Rule and the Bayesian Rule. The results are found successful in classifying GP practice profiles in a test dataset. This study is said to open the way towards its application in the medical fraud detection at Australia's Health Insurance Commission (HIC) as a routine application [25].

There are some commercial products and publicly available products in the market that claim to be effective medical fraud detection tools. These are:

- SPSS : Clementine 10
- Karypis Lab: CLUTO
- SAS Institute : Enterprise Miner (EM)

Being a data-mining tool for large-scale databases, Clementine 10 employs “anomaly detection” feature that permits it to be used for fraud detection. The incorporated anomaly detection scheme can simplify analysis and scoring, improve insight, and facilitate the use of these insights in operational deployments. Nonetheless, we haven’t encountered any successful application of the software to medical fraud detection problem.

CLUTO being freely available is a family of computationally efficient and high-quality data clustering and cluster analysis programs and libraries that are well suited for low- and high-dimensional data sets and for analyzing the characteristics of the various clusters. CLUTO is utilized for clustering data sets arising in many diverse application areas including information retrieval, customer purchasing transactions, web, GIS, science, and biology, thus can be used for fraud detection.

SAS Enterprise Miner streamlines the data mining process to create accurate predictive and descriptive models based on analysis of vast amounts of data from across the enterprise. There are organizations using SAS data mining software to detect fraud, anticipate resource demands, increase acquisitions and curb customer attrition. The software provides multiple advanced predictive and descriptive modeling algorithms, including market basket analysis, decision trees, gradient boosting, neural networks, linear and logistic regression, and more.

SAS EM and CLUTO have been applied to a large real-life health insurance dataset [62]. Experimental results indicate that CLUTO is faster than SAS EM while SAS EM provides more useful clusters than CLUTO.

2.6. PERFORMANCE MEASURES

In order to conclude that a study is successful, some performance measures should be defined and fulfilled. As revealed by Phua, 2004, many studies consider the possible cost savings or profits to be the success indicators. Phua *et al.*, 2004; Chan *et al.*, 1999; Fawcett and Provost, 1997 define explicit cost. Wang *et al.*, 2003 employ benefit models. For telecommunications fraud, Cahill *et al.*, 2002, outlines scoring an instance (a phone call in this case) by dividing the similarity measure of it to known fraud examples divided by the dissimilarity measure of it to known legal examples.

The unsymmetrical nature of the fraud imposes false positive and false negative error costs to unequal. These costs are unstable, changing by time and changing from example to example. Since a false negative example can be highly costly and a false positive error only costs for the human expert screening time, a false negative error is mostly more costly than a false positive error.

Therefore, lately, supervised algorithms based fraud detection methodologies no longer use assessments on true positive rate (correctly detected fraud divided by actual fraud) and accuracy at a chosen threshold (number of instances predicted correctly, divided by the total number of instances). Some employ Receiver Operating Characteristic (ROC) analysis (true positive rate versus false positive rate). Viaene *et al.* (2004), seek to maximize the Area under the Receiver Operating Curve (AUC). Caruana and Niculescu-Mizil (2004) argues that the most effective way to assess supervised algorithms is to use one metric from threshold, ordering, and probability metrics; and they justify using the average of mean squared error, accuracy, and AUC.

Lee and Xiang (2001) define entropy, conditional entropy, relative conditional entropy, information gain, and information cost for semi-supervised methodology assessment.

Ghosh and Reilly, 1994, illustrate other measurements like the speed of detection defined by detection time over time to alarm, the number of types of fraud revealed by the system and the format of the detection being online or offline.

When considering insurance fraud detection, some human expert involvement is imposed. Von Altrock (1995) asserted their system to perform better than human experts. Brockett *et al.* (2002) and Stefano and Gisella (2001) have found their work to be successful even to consistent human expert outcomes. Belhadji *et al.*, 2000 and Williams, 1999, both defend the role of human experts in a fraud detection system.

Chapter 3

MATHEMATICAL FORMULATIONS

Since most of the fraud detection papers focus on nonlinear, black-box supervised algorithms as neural networks, we can assert that less complex, reliable and faster algorithms are needed for such a research. Given that our database does not have fraudulent and legitimate labels for the transactions, our only data mining option for fraud detection is an unsupervised approach.

For auditing medical transactions, it is obvious that we need two tools. One is for batch screening/auditing and the other is for online/on time transaction control. This imposes building up two systems that are working interactively. Clearly, the online system needs to incorporate strategies to overcome the need for re-processing the whole batch of prescriptions in every new transaction. Besides these needs, the data structure, and size are the other design considerations. We try to fulfill these requirements under the assumption that the fraudulent cases are outliers in the database.

Since many outlier detection algorithms in the literature are designed for the specific problem in hand, and since there is no other prescription fraud detection work available in the literature, we need to consider our design needs and try to build up an efficient tool for the problem.

3.1. DATA STRUCTURE

Since data mining tools are developed regarding data structure, as well as the dimension and the size of the database in hand, first we need to analyze our database structure. We work on a database of 87,785 prescribed drugs in 2007 and 2008, stored in Excel 2007 spreadsheet format. The initial database provided us with 9 features (attributes). These 9 features are:

- Commercial name of the prescribed drug,
- Barcode number of the drug,
- Prescription number,
- Patient number,
- Age,
- Sex,
- ATC code of the drug,
- ATC name of the drug,
- Diagnosis for which the drug is prescribed.

Anatomical Therapeutic Chemical (ATC) codes were initially recommended by the WHO Regional Office for Europe in 1981. In the ATC system, drugs are divided into fourteen main groups (1st level), with one pharmacological/therapeutic subgroup (2nd level). The 3rd and 4th levels are chemical/pharmacological/therapeutic subgroups and the 5th level is the chemical substance. The complete classification of *Simvastatin* illustrates the structure:

C: Cardiovascular System
C10: Serum Lipid Reducing Agents
C10A: Cholesterol and Triglyceride Reducers
C10AA: HMG CoA Reductase Inhibitors
C10AA01: Simvastatin

When examining the transactions, we noticed that the ATC code and ATC names given in the database are not compatible with the above scheme. Moreover, there were drugs that were given different ATC codes when recurred. Thus, we decided to eliminate these attributes that could have been very helpful in classifying the drugs. Therefore, we decided to add *active principle ingredient* of the medicament to our database. For this, we have used the price lists of 2007 published by the Health Care Ministry, which is an extensive resource on commercially available drugs including the active ingredients, pharmacy and depot prices. We have matched the ingredients and the prices by an Excel Visual Basic Macro code. Nonetheless, the list lacked the information on about 2500 prescribed drugs that were left blank for price and ingredient features after executing the Macro code. We have inserted the related prices on these manually. Unfortunately, even manually we were not able to identify some active ingredients of some drugs that are not currently in the market. Thus, we decided not to use active ingredients as a feature in the database.

The next step was to decide on the features to involve in the procedure. It is trivial that age and sex are critical features that medical doctors consider in prescribing medicaments. Diagnosis is the core feature to judge the fraudulency of a prescription. Price should give us an idea of the level of spending per prescription, thus it is also important. So, the features that we are to consider in prescription fraud detection are:

- Commercial name of the prescribed drug,
- Market price of the prescribed drug,
- Prescription number,
- Age,
- Sex,
- Diagnosis for which the drug is prescribed,

3.2. REVISED METHODOLOGIES

In the the search for an efficient algorithm for prescription fraud detection, we have incorporated some of the existing methodologies for this problem. Among these algorithms are: association rules, infrequent item set mining and k-means.

3.2.1. Frequent Item Set Mining/Association Rule Learning

Association rule learning, being a popular method in data mining for revealing interesting relations within databases, was the first data mining methodology that we studied for the prescription fraud detection. The executable *Apriori* by Chriatian Borgelt was used as the data-mining tool. This tool is also incorporated in the commercial data-mining tool Clementine by SPSS. As for the application in our database, we defined the frequency threshold very low for the algorithm to give us infrequent item sets. After having run the code on the prescription database for the item sets of ATC codes in prescriptions, the results were found insignificant by the medical doctor Çağdaş Baran.

3.2.2. Infrequent Item Set Mining

Infrequent item set mining is a new algorithm for minimal infrequent item set mining. This algorithm aims to identify the rare item sets seen in a large database [81]. The algorithm *Minit* presented by Haglin and Manning (2007), is designed to serve as a tool for mining those rare item sets. We have made use of Minit's open-source code for mining the infrequent item sets in our database.

Consider a feature specification for patients as for sex and age interval. Let the set F be the set of the prescriptions that this specified group of patients has been prescribed. Assuming that finding a rare item set of drugs in this dataset of prescriptions would mean that this rare item set of drugs are from a fraudulent transaction, running Minit on the set F would give us those fraudulent transactions in the data set.

We have made several trials for incorporating this methodology for the 2007 prescriptions data that has 57,128 drugs. We have conducted trials on the domains that were the meaningful classifications for our database:

- Women,
- Men,
- Infants,
- Children,
- Adolescents,
- Men adults,
- Women adults,
- Women adolescents,
- Men adolescents.

The trials revealed such a big amount of infrequent item sets that we need to consider using other scalable methodologies for prescription fraud detection on our large database.

3.2.3. Clustering

Clustering is a methodology for partitioning the observations in hand into k clusters. Since the nature of our database do not impose a particular k , we need to define an algorithm in which the number of clusters is flexible. Here, we try to cluster the prescriptions, thus the observations to work on are prescriptions that are defined to be item sets of medicaments.

For this purpose we have developed a novel algorithm. First let us define similarity and quality measures. We define the similarity function as: $sim(p_1, p_2) = |p_1 \cap p_2| / |p_1 \cup p_2|$, where p_1 and p_2 are item sets. On the other hand, our quality measure is $\sqrt{\sum_i sim(c, p_i)^2}$, where c is the centroid of an item set.

Please note that these measures are novel in the sense that they are modifications of notations already in use in the literature.

To initialize, we define the first observation to be the centroid. Then we calculate the quality regarding this first centroid. In the next iteration, we list all items in the data set and rank them according to their occurrence frequencies. We take the first ranking item, put it in the centroid and calculate the quality measure. Then we take the second ranking and add this item to the centroid and calculate the quality regarding this new centroid, and so on. If an item in the list decreases the previously calculated quality, we define the centroid to be the one defined in the previous step. After having updated the centroid, we now arrange the clusters. We calculate the similarity between the observations and (item sets) and the defined centroid for each observation. For the first iteration if the similarity measure between the item set and the centroid is bigger than zero, then this item set rests in the initial cluster. If the similarity is zero for an item set, this item set defines a new centroid.

After having completed the initialization, we enter the second iteration. In this iteration, we recalculate the quality measures for each cluster as we did in the first iteration. That way a new centroid is defined for each of the clusters. Then, we recalculate the similarity measures between the item sets and centroids. If a similarity measure is below the threshold, the item set defines a new cluster; else it remains in the existing cluster. The iterations terminate if there is no new cluster to create.

With this algorithm, we aimed to create a non pre-defined number of clusters of the dataset. This way we hoped to find fraudulent prescriptions (outliers), which can be defined to be small clusters. Unfortunately, this clustering approach based on quality measures, proved to be sub-optimal in the tests. That is the reason we consider other approaches here.

3.3. METHODOLOGICAL DESIGN

As stated in the previous section, we have a domain of 6 dimensions, meaning that we have 6 different features to consider for this database which are; prescription number, medicament name, diagnosis, age, sex, and price. If we are to find the fraudulent transactions, it is clear that we are involved with a multivariate study. Nonetheless, if we explicate the nature of the data in hand, we see that the correlated features are:

- Medicament and Diagnosis,
- Medicament and Age,
- Medicament and Sex,
- Diagnosis and the total cost of drugs prescribed for this diagnosis,
- Medicament and Medicament interactions in a prescription.

Since there is no correlation between the rest like age and sex; we do not need to get involved with this cross-feature. Now, let's consider the interactions between diagnosis and age as well as diagnosis and sex. There can be specifications like pediatric diagnoses or women illnesses. Then shall we consider these cross-features? The answer is no, since any such diagnosis should convey specific medicaments in the prescription. These specific medicaments should reveal any mismatching between the diagnosis and age or sex. These arguments transform our domain of 6 dimensions to sub-domains of 2 dimensions which are illustrated by the above mentioned interactions. Therefore, our problem is refined to deal with five two-dimensional spaces. Working with incidence and risk matrices which are to be defined in the next sections and having two parts of consideration as online and offline processing, our methodology's flow chart is as:

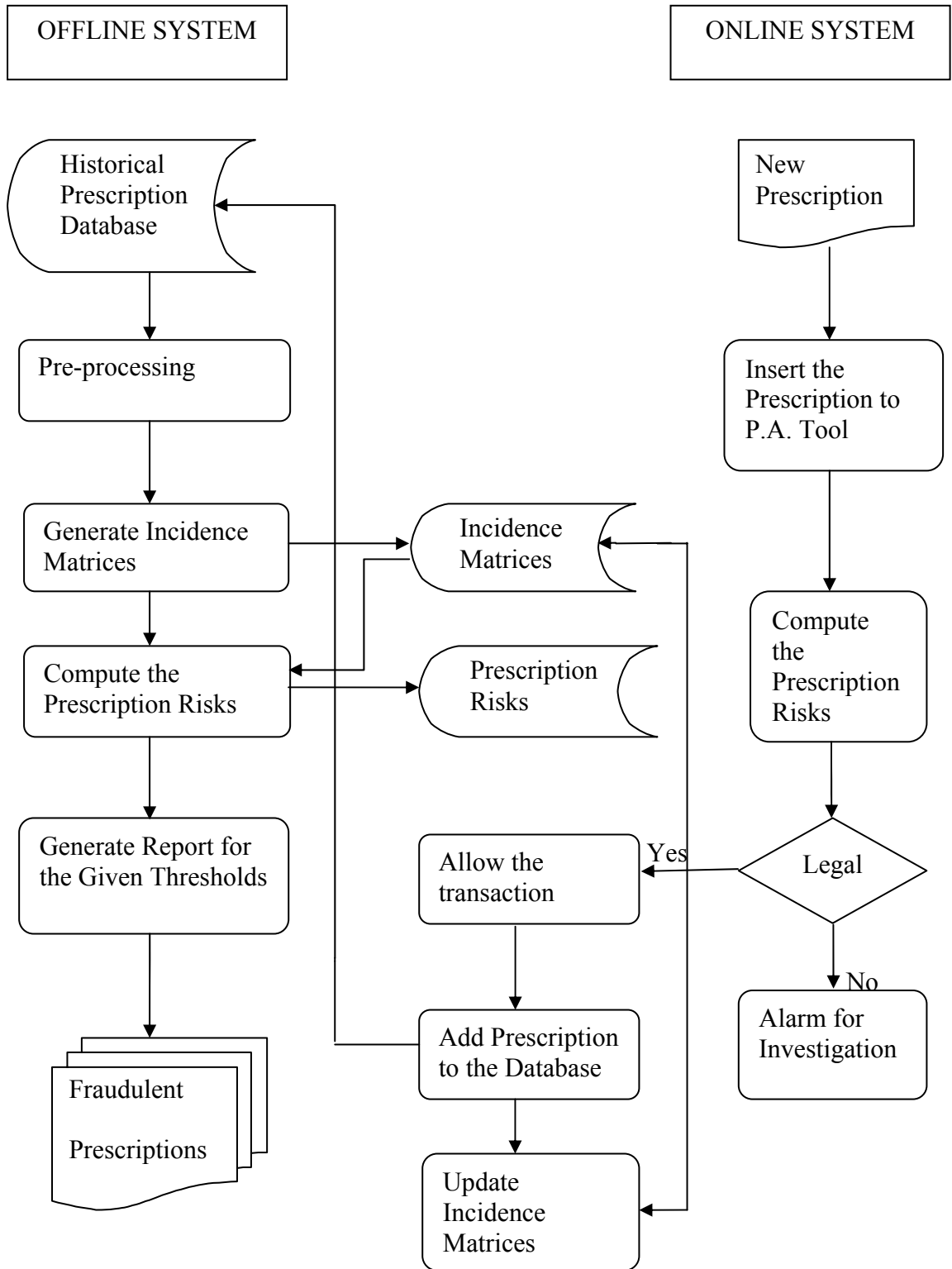


Figure 2: Flow chart of the integrated offline and online systems

3.4. RISK FORMULATIONS

To fulfill the design requirements, we introduce the Risk Formulations methodology for prescription fraud detection. The methodology consists of identifying the domains to apply risk formulations and then gradually constructing incidence matrices on which to calculate the risks.

The features that are involved with prescriptions are the prescription number, diagnosis, patient id, age of the patient, sex of the patient, practitioner id, the health care institution, and the prescribed medicaments. Our data included all but the practitioner and health care institution domains. Thus, we have concluded that the fraudulent cases that can be identified in our database are:

- Medicament and age mismatch (e.g. prescribing a pediatric medicine to a grown up),
- Medicament and sex mismatch (e.g. prescribing a birth control pill to a male),
- Medicament and diagnosis mismatch (e.g. in prescribing a antibiotic for a simple common cold),
- Medicament and medicament mismatch (e.g. in prescribing two drugs that are contra- indicating),
- Diagnosis and cost mismatch (e.g. in prescribing medicaments of hundred liras of total cost for a simple flu).

After having identified the domains to look for fraud, we build up the appropriate methodology under the assumption that fraudulent cases are rare in a large database of prescriptions issued by different health care institutions. Thus, what we need to do is detecting outliers in a way that the least common observations in one of the above domains are labeled to be fraudulent prescriptions. Nonetheless, when we consider the ordered age feature, a prescription in which a pediatric drug issued to a 15 year-old should not give a risk equal to

the risk of a prescription where the same pediatric drug is issued to a 45 year-old. Therefore, we introduce a modified version of risk formulation for ordered features like age.

3.4.1. Risk Formulation for Categorical Features

Sex, diagnosis, and prescription medicaments are non-ordered features, meaning that one can neither measure the entities listed in any of those nor make a grandeur comparison between those entities.

Consider the data set we work on. First of all, we build up the incidence matrices for the categorical features. These matrices hold the information regarding the number of times an instance shows up in the data set. In what follows, we describe how incidence matrices are created for each domain.

a) Medicament – Sex Domain

Let i represent a certain medicament and j represent the sex that it is issued to. Consider the Medicament – Sex incidence matrix denoted by MS . Note that the sex feature have two entities: female and male. Thus the size of this matrix is $2 \times (\text{the number of medicaments})$. Let's initialize $MS(i,j) = 0$ for all i and j . We would increment $MS(i,j)$ by one every time we encounter a case where the medicament i is issued to the sex j .

In order to compute the $risk_{MS}(i, j)$ which is the fraud likelihood of the cases in which the i^{th} medicament is prescribed for the j^{th} sex, we take the maximum of the i^{th} row of MS denoted by $Max_{MS}(i)$. Thus, $Max_{MS}(i)$ is the number of times medicament i is issued to the sex that is most issued to, thus it indicates the sex that the medicament i should be normally prescribed to in the cases where there is large gap between the $Max_{MS}(i)$ and $MS(i,j)$. Having identified those, the risk formulation is:

$$risk_{MS}(i, j) = \frac{\exp(-MS(i, j)/Max_{MS}(i)) - \exp(-1)}{1 - \exp(-1)} \quad (\text{Eq. 3.1})$$

Then, the risk matrix of the Medicament and Sex domain can be defined as:

$$MSR(i, j) = risk_{MS}(i, j)$$

Above formulation employs exponential function in order to receive a steep risk function since the formulation needs to return high indicators of fraud risk for small values of $MS(i, j)/Max_{MS}(i)$, which is the ratio of (i, j) incidence over the $Max_{MS}(i)$. Meaning that the function's sensitivity to detect fraud increases as the ratio $MS(i, j)/Max_{MS}(i)$ becomes smaller given that the derivative of $\exp(-x)$ increases as x gets smaller. Let us illustrate this with an example. Consider the medicament A which is an osteoporosis medicament for women and B which is an ordinary flu medicament. Let A be prescribed to 2 men and 102 women. Let B be prescribed to 55 women and 50 men. Then, the calculated risks for A would be 0.9693 if prescribed for men and 0 if prescribed for women. The risks for B would be, 0.0554 if prescribed for men and 0 for women. As illustrated in the Figure 3 below, exponential function detects well that the medicament A is a drug for women by giving a high risk value for A when given to men; whereas, there is no obvious sex distinction for B.

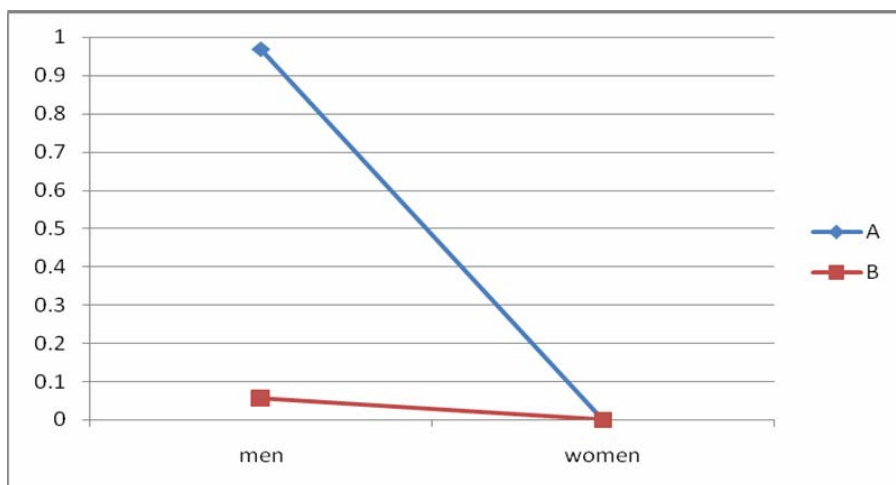


Figure 3: Examples of computational effectiveness of the risk formulation

If $MS(i, j) = Max_{MS}(i)$ there is no risk with this transaction, thus, we subtract $\exp(-1)$ from $\exp(-MS(i, j)/Max_{MS}(i))$ in order for the risk function to be nullified for such legal transactions. Then, we divide the numerator by $1 - \exp(-1)$ in order to attain a scaled value for the fraud risk between 0 and 1. Thus, the result is in the range 0-1 where the higher values indicate high likelihood of fraud and the lower values indicate low likelihood of fraud. The argument is the same for the following formulations.

b) Medicament – Diagnosis Domain

Let i represent a certain medicament and j represent the diagnosis that it is issued with. Consider the Medicament – Diagnosis incidence matrix denoted by MD. The size of this matrix is the number of medicaments * number of diagnoses. Let's initialize $MD(i, j) = 0$ for all i and j . We would increment $MD(i, j)$ by one every time we encounter a case where the medicament i is issued with the diagnosis j .

In order to compute $risk_{MD}(i, j)$ which is the fraud likelihood of the cases in which the i^{th} medicament is prescribed for the j^{th} diagnosis, we take the maximum of the i^{th} row of MD denoted by $Max_{MD}(i)$. Thus, $Max_{MD}(i)$ is the number of times medicament i is issued with the diagnosis that is most issued with, thus carries the information on the diagnosis that the medicament i should be normally prescribed to in the cases where there is large gap between the $Max_{MD}(i)$ and $MD(i, j)$. Then, the risk formulation is as:

$$risk_{MD}(i, j) = \frac{\exp(-MD(i, j)/Max_{MD}(i)) - \exp(-1)}{1 - \exp(-1)} \quad (\text{Eq. 3.2})$$

The risk matrix of the Medicament and Diagnosis domain is defined as $MDR(i, j) = risk_{MD}(i, j)$.

c) Medicament – Medicament Domain

Let i represent a certain medicament and j represent another medicament that it is issued in the same prescription. Consider the Medicament – Medicament incidence matrix denoted by MM . The size of this matrix is (the number of medicaments)*(the number of medicaments). Let's initialize $MM(i,j) = 0$ for all i and j . Consider the Medicament – Medicament incidence matrix denoted by MM . We would increment $MM(i,j)$ by one every time we encounter a prescription where the medicament i and j are issued in the same prescription.

In order to compute the $risk_{MM}(i, j)$ which is the fraud likelihood of the prescriptions in which the i^{th} medicament is prescribed with the j^{th} medicament, we take the maximum of the i^{th} row of MM denoted by $Max_{MM}(i)$. Thus, $Max_{MM}(i)$ is the number of times medicament i is issued with the medicament that it is most issued with, thus carries the information on the medicament that the medicament i should be normally prescribed to in the cases where there is large gap between the $Max_{MM}(i)$ and $MM(i, j)$. Having identified those, the risk formulation is as:

$$risk_{MM}(i, j) = \frac{\exp(-MM(i, j)/Max_{MM}(i)) - \exp(-1)}{1 - \exp(-1)} \quad (\text{Eq. 3.3})$$

The risk matrix of the Medicament and Medicament domain is defined as $MMR(i, j) = risk_{MM}(i, j)$.

3.4.2. Risk Formulation for Ordered Features

In our data set we have Age and Cost as ordered features which need special attention. Here, we define the refined formulations for the categorical features.

a) Medicament*Age Domain

Let i represent a certain medicament that is prescribed to age j . Consider the Medicament – Age incidence matrix denoted by MA. The size of this matrix is (the number of medicaments)*(number of ages). Here, the age range is the age range inferred by the dataset. Let's initialize $MA(i,j) = 0$ for all i and j . We would increment $MA(i,j)$ by one every time we encounter a prescription where the medicament i is issued to a patient of age j .

In order to compute the $risk_{MA}(i, j)$ which is the fraud likelihood of the prescriptions in which the i^{th} medicament is prescribed to a patient at the age j , we take the maximum of the i^{th} row of MA denoted by $Max_{MA}(i)$. Thus, $Max_{MA}(i)$ is the number of times medicament i is issued to the age that is most issued to, thus carries the information on the age that the medicament i should be normally prescribed to in the cases where there is large gap between the $Max_{MA}(i)$ and $MA(i,j)$. Moreover, we also identify the minimum of the i^{th} row of MA denoted as $Min_{MA}(i)$ as well as the minimum and maximum age that the prescription i is issued to in order to calculate an age range for the medicament. Let Max_j and Min_j denote the maximum and minimum of ages that the medicament i is prescribed to, respectively. In other words, $Max_j(i) = \{j : Max_{MA}(i) = MA(i, j)\}$ and $Min_j(i) = \{j : Min_{MA}(i) = MA(i, j)\}$. Then the age range of medicament i is $r_i = Max_j(i) - Min_j(i)$. Thus, the modified risk formulation would be:

$$risk_{MA}(i, j) = \frac{\exp(-(MA(i, j)/Max_{MA}(i)) * (1 - (d_i(j)/r_i))) - \exp(-1)}{1 - \exp(-1)} \quad (\text{Eq. 3.4})$$

where, $V_i = \sum_k ((k * MA(i, k)) / \sum_k MA(i, k))$ and $d_i(j) = |j - V_i|$, and the risk matrix of the Medicament and Age domain is defined as $MAR(i, j) = risk_{MA}(i, j)$.

b) Diagnoses*Cost Domain

Consider a prescription P, in which a number of different diagnoses are seen. Given the structure of our database, we can sum up the total cost regarding each of the corresponding diagnoses in P. Then we calculate which interval this costs fall into. Note that we work with 5 TL intervals for computational efficiency up until 1000 TL. Then for higher amounts, we use the indices 201 (for the interval 1000-1500), 202 (for the interval 1500-2000 TL), 203 (for the interval 2000-2500 TL) and 204 (for the amounts higher than 2500 TL). Let the total cost of the diagnosis D be 23 TL in a certain prescription. Then, the interval this amount falls into is the 5th interval since $5*4 < 23 < 5*5$. Having identified the interval for this diagnosis D in prescription P, we update the Diagnosis-Cost incidence matrix denoted by DC by incrementing $DC(d,5)$ by one. Note that the size of the matrix DC is (the number of diagnosis)*(number of cost intervals which is equal to 204).

In order to compute the $risk_{DC}(i, j)$ which is the fraud likelihood of the prescriptions in which the i^{th} diagnosis is has a total cost in the j^{th} interval, we take the maximum of the i^{th} row of DC denoted by $Max_{DC}(i)$. Thus, $Max_{DC}(i)$ is the number of times diagnosis i is issued to the cost interval that is most issued to, thus carries the information on the interval that the diagnosis i should be normally prescribed. In here, we also identify the minimum of the i^{th} row of DC denoted $Min_{DC}(i)$. We also identify the minimum and maximum cost intervals that the diagnosis i is issued to in order to calculate a price range for the diagnosis. Let Max_j and Min_j denote the maximum and minimum of intervals that the diagnosis i is prescribed to, respectively, where $Max_j(i) = \{j : Max_{DC}(i) = DC(i, j)\}$ and $Min_j(i) = \{j : Min_{DC}(i) = DC(i, j)\}$. Then the cost range of the diagnosis i is $r_i = Max_j(i) - Min_j(i)$. Thus, the modified risk formulation would be as:

$$risk_{DC}(i, j) = \frac{\exp(-(DC(i, j)/Max_{DC}(i)) * (1 - (d_i(j)/r_i))) - \exp(-1)}{1 - \exp(-1)} \quad (\text{Eq. 3.5})$$

Where, $V_i = \sum_k ((k * DC(i, k)) / \sum_k DC(i, k))$ and $d_i(j) = |j - V_i|$, and the risk matrix of the Diagnosis and Cost domain is defined as $MDC(i, j) = risk_{DC}(i, j)$.

Chapter 4

APPLICATION AND COMPUTATIONAL RESULTS

We have coded the above mentioned framework and formulations in Matlab 2008A release. Our data in hand is composed of 87,785 prescribed drugs in 2007 and 2008. The data is in Excel 2007 spreadsheet format having as columns:

- Commercial Drug Name,
- Prescription Number,
- Patient's Age,
- Patient's Sex,
- Diagnosis,
- Market price of the drug.

Commercial drug name, patient's sex, and diagnosis columns are in text style. Prescription number, patient's age, and market price of the drug columns are in numeric style.

4.1. OFFLINE PROCESSING

We have created an m-file which is the programming medium for Matlab, for the offline batch processing of the database. This code on 800 lines, takes the Excel file database and processes it to create:

- Drug Name List and Indices,
- Drug Price List,
- Diagnoses List and Indices,
- Age List and Indices.

The lists are created in the appearance order. For example '*FLIXONASE AQUEOUS NASAL SPREY 120 DOSE*' being the first medicament in our database, it is the first in the drug name list, so its index is 1. '*SEDERGINE VIT-C EFERVESANT TABLET 20 TB*' is the 17th cell in the excel file, but because of the recurrent drugs on the file, this medicament's Drug list Index is 12. The same approach for listing and indexing is valid for the rest of the lists.

The next step in processing is building up the incidence matrices for all the domains:

- Medicament and age: MA,
- Medicament and sex: MS,
- Medicament and diagnosis: MD,
- Medicament and medicament: MM,
- Diagnosis and cost: DC.

In building up these incidence matrices we make use of the above-mentioned listings. Consider the Medicament and Sex incidence matrix MS. This matrix's column labels are sexes: Woman and Man whereas the row labels are the drugs as listed in the Drug Name List. Consider the MS(12,1). This is the count of the number of times '*SEDERGINE VIT-C*

EFERVESANT TABLET 20 TB is prescribed to a woman. The code goes through all the drugs and every time a new transaction of '*SEDERGINE VIT-C EFERVESANT TABLET 20 TB*' appears, the code updates MS by incrementing MS(12,1) by one if the drug is given to a woman and updates MS(12,2) if the drug is prescribed to a man. The same arguments hold for the incidence matrices MD, MA, and MM. For the DC matrix, the row labels are diagnoses and column labels are indices from 1 to 204. These indices represent 5 TL intervals, but the last interval is for the diagnosis costs that are above 2500 TL. Thus, the code goes through all the prescriptions and every time a new prescription is encountered, it looks for the diagnoses that this prescription includes. Then for every diagnosis within, the total costs of the corresponding medicaments are calculated. Let this calculated amount to be 73 TL for diagnosis 6. Then the assigned column index for this amount is 15, since $5 \cdot 14 < 73 < 5 \cdot 15$. Last, the code increments DC(6,15) by one.

Now having all the incidence matrices in hand, the code creates risk matrices. These matrices are:

- Medicament and age risks: MAR,
- Medicament and sex risks: MSR,
- Medicament and diagnosis risks: MDR,
- Medicament and medicament risks: MMR,
- Diagnosis and cost couple's risks: DCR.

These matrices are all initialized to be zero. Then, they are built up by calculating the risks for the corresponding incidences in the corresponding incidence matrices. For example, for calculating the '*SEDERGINE VIT-C EFERVESANT TABLET 20 TB*' drug's sex risk when given to men, we calculate MSR(12,2) by using the corresponding categorical risk formulation for MS(12,2). We need to keep the incidence matrices for offline processing, so we do not directly update the incidence matrices for risk computations. Note that, for MAR and DCR, we need to employ the corresponding ordered feature risk formulations.

Having all the risk matrices in hand, the code goes through all the risks that are greater than the thresholds given by the user. The user can indicate any threshold he wants for any of the risk matrices keeping in mind that more prescriptions would be classified as risky when the threshold is kept small. That is, there is a tradeoff between the true positive rate and the human expert screening time. The user should define the level of tradeoff he is ready to accept.

Given the thresholds, the code outputs the fraudulent prescriptions by indicating which types of fraud are seen within the prescriptions. That way, the human expert has the chance to revise the outputted prescriptions, which saves time and money to audit large databases.

4.2. ONLINE PROCESSING

The online prescription fraud detection tool is an interactive tool coded in Matlab that has a graphical user interface. This interface is designed to enable the user to insert new prescriptions to the database and audit a new prescription without the need to re-run the offline code. The methodology that lies behind the online code makes use of the global variables of the incidence and risk matrices. Thus, new prescription auditing can be done once after the offline code is run on the prescription database in hand.

Below you can find the screenshot of the graphical user interface of the auditing tool:

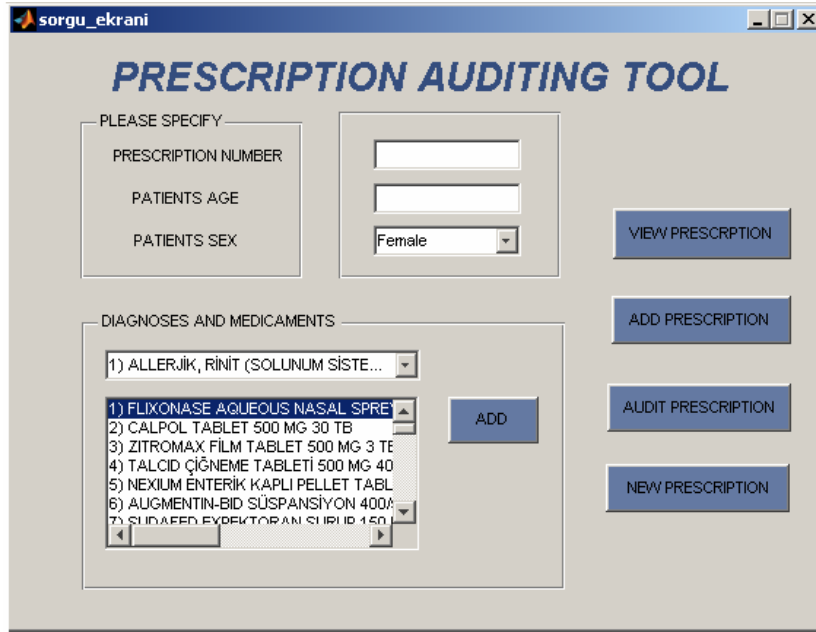


Figure 4: Prescription Auditing Tool User Interface

As seen in the above picture, the user first needs to input the prescription number as well as the age and sex of the patient. Then, in the box below the user puts in the prescribed drug and the corresponding diagnosis by the add button. The drug and diagnosis list boxes are populated by the drug name and diagnosis lists, which are the outputs of the offline fraud detection code. Next step in online fraud detection is checking to see if the input is correct by the show prescription button. If the prescription input is correctly specified, the user might choose to add the prescription directly to the database. That is achieved by fetching the corresponding rows of the incidence and risk matrices and updating those by the online code's input of the incoming prescription specifications. Alternatively, the user might want to audit the prescription directly. That way, input of the prescription is not used to update the incidence and risk matrices permanently. This is preferable since if the incoming prescription is fraudulent, updating the incidence and risk matrices by this input would slightly affect the performance of the code, since increasing the number of outliers in a database would eventually lead the outliers to be the most common transactions. This would hinder the tool to detect those fraudulent transactions. So, the user should add the incoming prescription to the database if the prescription is surely not fraudulent, perhaps after the auditing process.

Pushing the audit button, the user instantly receives a message indicating each levels of fraud riskiness regarding the prescription. Lastly, the new prescription button enables the user to put in a new prescription just after auditing another one.

We can assert that this kind of an online tool is necessary given the nature of the transactions in the health care sector.

4.3. OFFLINE FRAUD DETECTION RESULTS

We have run the offline code on the database of 87,785 prescribed drugs. As stated previously, each run requires the user to specify riskiness thresholds of each kind of confirmation check procedure. The code reveals the prescriptions which possesses higher risks than the thresholds. We have taken several runs in order to refine the preferable threshold for each of the domains.

These runs reveal that the sensitivity levels of each of the criteria are different. The reason for that lies in the fact that the sizes of the incidence matrices are different from each other and thus the sparseness and intensity characteristics of each differ. That is to say, the maximum numbers in a risk matrix's row and the rows themselves change from matrix to matrix for each medicament leading to different sets of risk indicators for the corresponding features. Thus, each threshold needs a separate refinement. We have conducted this refinement in the supervision of a medical doctor who assessed the significance levels of the outputs. The refined model for each auditing task uses the following threshold values:

- Medicament-Diagnosis Domain: 0.85,
- Medicament-Age Domain: 0.90,
- Medicament-Sex Domain: 0.96,
- Medicament-Medicament Domain: 0.95,
- Diagnosis-Cost Domain: 0.85.

For a sample output of the above model, please refer to APPENDIX A. When considering the output, we see that the database has:

- 87,785 lines,
- 26,419 prescriptions,
- 2,659 drugs,
- 332 diagnoses,
- 963 active ingredients,
- Patients of minimum age of 0 and maximum age of 85.

An interesting observation about the audit results is that the fraudulent labeled prescriptions tend to have multiple numbers of riskiness reasons. For example let's consider the prescription 1592467 of which the database features are given below:

P. No	Medicament Name	Age	Sex	Diagnosis	Active Ing.	Price (TL)
1592467	<i>Iliadin</i>	57	M	<i>Glaukoma</i>	<i>Oksimetazoline</i>	4.59
1592467	<i>Cosopt</i>	57	M	<i>Glaukoma</i>	<i>Tymolol Maleate + Dorzolamide</i>	30.80
1592467	<i>Cosopt</i>	57	M	<i>Glaukoma</i>	<i>Tymolol Maleate + Dorzolamide</i>	30.80
1592467	<i>Coraspin</i>	57	M	<i>Glaukoma</i>	<i>Acetylsalicylic acid</i>	2.40

Table 2: Prescription Example-1

The code's output for this prescription is as:

Prescription Number: 1592467

- Incompatibility between Medicament: *Iliadin* Diagnosis: *Glaukoma*, Risk: 0.96
- Incompatibility between Medicament: *Coraspin* Diagnosis: *Glaukoma*, Risk: 0.92
- Incompatibility between Diagnosis: *Glaukoma* Cost(TL): 70, Risk: 0.87

Cosopt, being an ophthalmic suspension, is a legitimate item in the prescription. Nonetheless, *Iliadin* is a nasal spray and *Coraspin* is a kind of aspirin. This might be an indicator that the fraudsters tend to add several fraudulent items in a prescription that could have been legitimate without those.

Let us now consider Medicament and Medicament non-conforming prescriptions. In the first look, it might be surprising to see that there is no prescription with these criteria when the threshold is above 0.90. Nonetheless, if we reconsider the nature of the Medicament*Medicament incidence matrix, we see that this matrix is of size 2,659*2,659. Consider the row i in this matrix, this row consists of the co-occurrence numbers of the i^{th} medicament with any other medicament. Since there are 2658 other medicaments, it is obvious that this medicament i can be seen with a huge number of other drugs in a prescription, given the diagnoses comply. That means, theoretically, the rows of MM do not constitute skewed distributions. Thus, the maximum of each row, which plays an important role in determining the risks regarding any others, is not significant when compared with other elements of the row. This theoretic assumption is validated empirically when the code is employed. There is no significant risk regarding this criterion. We see such risks only if the diagnosis is non-conforming with the medicament also. Please refer to the prescription below for further illustration.

P.No	Medicament Name	Age	Sex	Diagnosis	Active Ing.	Price (TL)
431603	<i>Euthyrox</i>	49	F	<i>Osteoporosis</i>	<i>Levotiroxin</i>	2.63
431603	<i>Fosamax</i>	49	F	<i>Osteoporosis</i>	<i>Alendronate</i>	39.46
431603	<i>Zyrtec</i>	49	F	<i>Osteoporosis</i>	<i>Setirizine hcl</i>	10.77

Table 3: Prescription Example-2

The output of the offline code for this prescription is as:

Prescription No: 431603

Non conformation between Medicament: *Zyrtec* and Diagnosis: *Osteoporosis* Risk: 0.93

Non conformation between Medicament *Fosamax* and Medicament: *Zyrtec*, Risk: 0.61

Zyrtec, being an allergy medicament, does not conform to the diagnosis osteoporosis. Having so rarely been given with this diagnosis, it has a high risk (0.93) and thus, its Medicament vs. Medicament riskiness is high with the osteoporosis medicaments.

In order to enable to check the riskiness of two medicaments, we have coded the Active Ingredient and Active Ingredient conformation check for two medicaments in a prescription. This might overcome the above stated problems with the MM matrix by working on the active ingredients matrix of dimensions 963*963. This scaling down could have worked well for such a problem, nonetheless, we were not able to identify the active ingredients for a portion of the medicaments, and so we were not able to get the results for our database for this kind of detection.

4.4. ONLINE FRAUD DETECTION RESULTS

The most important part of our study is building the online/on time prescription fraud detection tool given the nature of the health care transactions. This tool aims to constitute an example for the application possibilities regarding our proposed fraud detection methodology. As stated above, we have coded the graphical user interface of the tool in Matlab. The tool takes the input prescription and the user might choose to:

- View the prescription,
- Add the prescription to the database,
- Audit the prescription,
- Insert a new prescription.

For illustrating the effectiveness of the online fraud detection tool, let us consider a prescription given to a 55 years old woman. She is diagnosed with the upper respiration tube infection and is given the medicaments Sudafed Syrup, Otrivine Pediatric Spray and Stafine Pomade. The initial user interface is as seen in Figure 4 after inputting the prescription:

The screenshot shows a web application window titled "PRESCRIPTION AUDITING TOOL". The interface is divided into several sections:

- PLEASE SPECIFY:** A section for entering patient details. It includes three input fields: "PRESCRIPTION NUMBER" (containing "64283"), "PATIENT'S AGE" (containing "55"), and "PATIENT'S SEX" (a dropdown menu set to "Female").
- DIAGNOSES AND MEDICAMENTS:** A section for listing medical conditions and drugs. It features a dropdown menu with the selected item "2) BAKTERİYEL, ÜST SOLUNUM YOL...". Below this is a list of medications with checkboxes and an "ADD" button. The list includes:
 - 44) LEVOTIRON 0.1 MG 100 TB
 - 45) MADECASSOL POMAD 40MG/G 40 C
 - 46) ROCALTROL KAPSÜL 0,25 MCG 30
 - 47) TEARS NATURALE FREE SUNİ GÖZ
 - 48) OTRIVINE PEDIYATRİK DOZ AYARLI
 - 49) RINOGEST ŞURUP 30 MG/5ML 100 M
 - 50) DOL MEN PEDIYATRİK 100 ML ŞURUP
- Buttons:** On the right side, there are five blue buttons: "VIEW PRESCRIPTION", "ADD PRESCRIPTION", "AUDIT PRESCRIPTION" (highlighted with a dashed border), and "NEW PRESCRIPTION".

Figure 5: Inserting a Prescription to the Prescription Auditing Tool

If the user chooses to view the prescription a message box appears as:

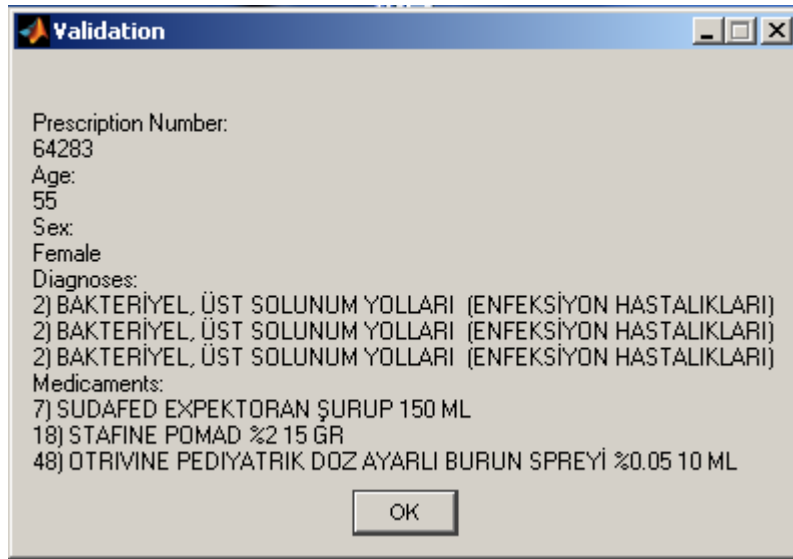


Figure 6: Validation Message Box

After validating the prescription input, the user might choose to add the prescription to the database. If so, the below message box appears:

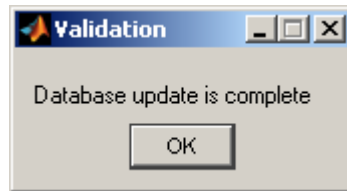


Figure 7: Database Update Notification

If the user chooses to audit the prescription the below message box appears:

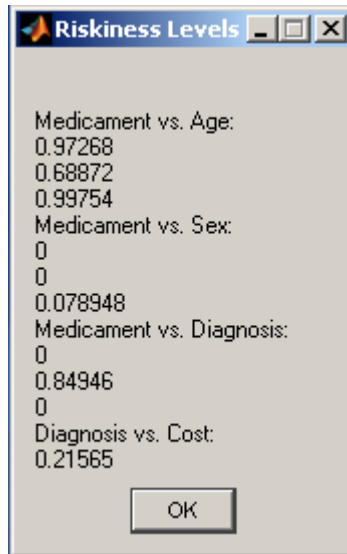


Figure 8: Riskiness Levels Screen

Here, risks regarding Medicament and Age non-conformation are stated in the input order of the medicaments, just as the Medicament and Sex non-conformation. Considering the diagnoses, the Medicament and Diagnosis risks are seen in the screen in the appearance order in the prescription of the medicament and diagnosis couples. Diagnosis and Cost non-conformation risks are seen for each of the diagnosis in the prescription. Here, we see one entry for the Diagnosis and Cost non-conformation risk since there is only one diagnosis in the prescription.

When we consider the prescription, the diagnosis is *upper respiration tube infection*. Since *Sudafed Syrup* and *Otrivine Pediatric Spray* are compatible for this diagnosis, we can conclude that, the tool is effective to calculate 0 risks for the medicament and diagnosis domain for these two medicaments. For *Stafine Pomade*, which is a skin care medicament, we see that the tool calculates a high risk (0.85), which is expected.

The patient is a 55-year-old woman. Even though there is no risk associated with the sex of the patient and the medicaments, Both *Sudafed Syrup* and *Otrivine Pediatric Spray* are

for children. So, the tool identifies the high risks regarding the age of the patient as to be 0.97 for *Sudafed Syrup* and 0.99 for the *Otrivine Pediatric Spray*.

4.5. PERFORMANCE EVALUATION

We consider false positive, false negative, and true positive rates as well as the agreement rate as performance indicators for our system. We have gone through 1033 medicaments in 249 prescriptions with M.D. Çağdaş Baran who is a cardio-vascular surgeon in Ankara University Cardio Center. Mr. Baran had labeled the fraudulent prescriptions in this random sample of 249 prescriptions taken from our database. We evaluate the results obtained from the offline prescription fraud detection tool with the thresholds:

- 0.80 for Medicament and Diagnosis,
- 0.90 for Medicament and Sex,
- 0.96 for Medicament and Age,
- 0.80 for Medicament and Medicament,
- 0.85 for Diagnosis and Cost.

In this performance measurement study, false positives are the prescriptions that the system reveals to be fraudulent even though they are not considered so by human experts. False negatives are the prescriptions that the system does not reveal to be fraudulent when those are in fact considered so by human experts to be fraudulent transactions. The true positive rate is the proportion of correctly detected fraudulent prescriptions to the number of actual fraudulent cases. Agreement rate is simply the ratio of the system output which is compatible with the human expert auditing over the total number of instances.

We have identified 17 false positives, 19 false negatives, 72 true positives, and 141 true negatives in this test sample of prescriptions.

The comparison between the human expert labeling and our system has led to the following results:

- False Positive Rate = Number of False Positives/Total Number of Instances
= 6.09%
- False Negative Rate = Number of False Negatives/Total Number of Instances
= 7.63%
- True Positive Rate = Number of True Positives/ Number of Real Positives
= 77.4%
- Agreement Rate = (Number of True Positives + Number of True Negatives)
/ Total Number of Instances
= 85.54%

We have also conducted a benchmarking study to compare our results for the 6% false positive rate. Even though we have selected health care fraud detection tools' performance measures for this comparative study, note that we cannot impose a direct comparison between the systems since none of the data or algorithms are available for applying to our algorithm or database, respectively. Thus, this comparison is conducted in order to give an illustrative performance benchmarking. The studies that we have compared our results are:

- Ortega *et al.* (2007): *A Medical Claim Fraud/Abuse Detection System based on Data Mining: A Case Study in Chile*
- Major, J., *et al.* (2002): *EFD: A Hybrid Knowledge/Statistical-Based System for the Detection of Fraud*
- He, H., *et al.* (1999): *Application of Genetic Algorithms and k-Nearest Neighbor Method in Medical Fraud Detection*
- Yang, W., Hwang S., (2006): *A Process-mining Framework for the Detection of Health Care Fraud and Abuse*

Among those, EFD performs worse with a true positive rate of 26.4% when the false positive rate is 5.9% [23]. This gives a true positive over false positive ratio of 4.47. On the other hand, this ratio is 12.7 for our system.

The Medical Claim Fraud/Abuse Detection System proposed by Ortega *et al.* (2007), gives 71% true positive rate for the 6% false positive rate level. This indicates a true positive over false positive ratio of 11.8 [58]. Eventhough higher than EFD, this rate still can't catch up with our systems performance level of 12.7.

When we consider the next study; by He *et al.* (1999); we see that authors have cooperated a performance measure as the agreement rate. Their system reveals different agreement rates for different trials in between 73% to 79% [24]. The best agreement rate attained is 78.8%. When compared to our agreement rate of 85.5%, this agreement rate is considerably low.

The Process-mining Framework by Yang and Hwang (2006) gives a true positive rate of 69% [59]. When compared to 77.4%, we see that our proposed methodology out performs this proposed framework as well.

Having considerably better performance on health care fraud detection than all the above earlier works in the literature, we can state that our system has revealed satisfactory results for this specific domain of fraud detection.

Chapter 5

CONCLUDING REMARKS AND FURTHER RESEARCH DIRECTION

In this thesis, we studied the prescription fraud detection problem. Our novel methodology proposes dividing down the 6 dimensional features' domain into several sub-domains considering the interaction levels between the features. The studied domains are:

- Medicament and Diagnosis,
- Medicament and Age,
- Medicament and Sex,
- Medicament and Medicament, and
- Diagnosis and Cost.

The methodology consists of populating incidence matrices for each of the above domains and then incorporating a novel data-mining approach for each of the categorical and ordered domain. This approach is modeled to fulfill the requirements imposed by the highly specialized characteristics of the prescription data. The risk formulations employing this data-mining approach return riskiness measures for each of the prescriptions and for each of the above-mentioned domains. This riskiness measure is scaled to be between 0 and 1, in order to

give a straightforward definition of the riskiness level. For each of the domains, the user can specify thresholds. That way, the code returns only those prescriptions with higher riskiness levels than the thresholds.

We have built up a Matlab code for batch auditing the database in hand. The automated fraud detection methodology gives considerably compatible results with the human expert auditing.

We have built up a user-friendly graphical user interface for enabling on time fraud detection for the new prescriptions. We can state that online fraud detection tools such as this one are needed given the nature of the health care transactions.

The superiority of our proposed system to the other possible outlier detection methodologies for prescription fraud detection is that, it is user friendly since it has been flexible enough for an integrated online/on time user interface; it is well tailored for prescription fraud detection, it presents a novel and easy way to keep track of health care transactions in incidence matrices for auditing, other new detection systems can be built on these incidence matrices if needed. Last but not the least, its core methodology is adoptable to many other areas in health care and possibly in other industries.

Given the performance measurements with a true positive rate of 77.4% and a false positive rate of 6%, we can conclude that our system works considerably well for the prescription fraud detection problem. Nonetheless, future research directions can be stated for a superior fraud detection tool.

These future research directions can include refining the offline model in order for it to scale well across all domains, meaning that incorporating different parameters for different domains so that the same risk measurements mean the same level of riskiness across all domains. Also, a tool can be built up where the user can specify the domains he wants work

on. Refining the model to be more modular for this purpose can simplify creating similar models for other health care areas like: blood tests, x-rays and tomographies.

BIBLIOGRAPHY

1. Phua C., C., Lee V., Smith K. "Comprehensive Survey of Data Mining-based Fraud Detection Research", *Artificial Intelligence Review*, (2005)
2. Elkan, C. "Magical Thinking in Data Mining: Lessons from CoIL Challenge 2000", *Proc. of SIGKDD01*, (2001): 426-431.
3. Fawcett, T. "'In Vivo' Spam Filtering: A Challenge Problem for KDD", *SIGKDD Explorations* **5**(2) (2003): 140-148
4. Lavrac, N., Motoda, H., Fawcett, T., Holte, R., Langley, P. & Adriaans, P. Introduction: "Lessons Learned from Data Mining Applications and Collaborative Problem Solving" *Machine Learning* **57** (1-2): 13-34.
5. "Turkish Health Care Syndicate 2008 Health Care Report" 2008. <<http://www.turksagliksen.org.tr/content/view/6271/55/>>
6. "About Basel Committee" 2003, <<http://www.bis.org/bcbs/>>
7. Lin, J., Hwang, M. & Becker, J. "A Fuzzy Neural Network for Assessing the Risk of Fraudulent Financial Reporting" *Managerial Auditing Journal* **18**(8) (2003): 657-665
8. Bell, T. & Carcello, J. "A Decision Aid for Assessing the Likelihood of Fraudulent Financial Reporting" *Auditing: A Journal of Practice and Theory* **10**(1) (2000): 271-309.
9. Fanning, K., Cogger, K. & Srivastava, R. "Detection of Management Fraud: A Neural Network Approach". *Journal of Intelligent Systems in Accounting, Finance and Management* **4** (1995): 113-126.
10. Summers, S. & Sweeney, J. "Fraudulently Misstated Financial Statements and Insider Trading: An Empirical Analysis" *The Accounting Review* January (1998): 131-146.
11. Beneish, M. "Detecting GAAP Violation: Implications for Assessing Earnings Management Among Firms with Extreme Financial Performance" *Journal of Accounting and Public Policy* **16** (1997): 271-309.
12. Green, B. & Choi, J. "Assessing the Risk of Management Fraud through Neural Network Technology". *Auditing* **16**(1) (1997): 14-28.

13. Bentley, P. "Evolutionary, my dear Watson: Investigating Committee-based Evolution of Fuzzy Rules for the Detection of Suspicious Insurance Claims" *Proc. of GECCO2000*
14. Von Altrock, C. "Fuzzy Logic and Neurofuzzy Applications in Business and Finance" Prentice Hall (1997) 286-294
15. Little, B., Johnston, W., Lovell, A., Rejesus, R. & Steed, S. "Collusion in the US Crop Insurance Program Applied Data Mining". *Proc. of SIGKDD02*, (2002): 594-598.
16. Phua, C., Alahakoon, D. & Lee, V. "Minority Report in Fraud Detection: Classification of Skewed Data", *SIGKDD Explorations* 6(1) (2004): 50-59.
17. Viaene, S., Derrig, R. & Dedene, G. "A Case Study of Applying Boosting Naive Bayes to Claim Fraud Diagnosis". *IEEE Transactions on Knowledge and Data Engineering* 16(5) (2004): 612- 620.
18. Brockett, P., Derrig, R., Golden, L., Levine, A. & Alpert, M. "Fraud Classification using Principal Component Analysis of RIDITs" *Journal of Risk and Insurance* 69(3) (2002): 341-371.
19. Stefano, B. & Gisella, F. "Insurance Fraud Evaluation: A Fuzzy Expert System" *Proc. of IEEE International Fuzzy Systems Conference*, (2001): 1491-1494.
20. Belhadji, E., Dionne, G. & Tarkhani, F. "A Model for the Detection of Insurance Fraud" *The Geneva Papers on Risk and Insurance* 25(4) (2000): 517-538.
21. Artis, M., Ayuso M. & Guillen M. "Modelling Different Types of Automobile Insurance Fraud Behavior in the Spanish Market" *Insurance Mathematics and Economics* 24 (1999): 67-81.
22. Yamanishi, K., Takeuchi, J., Williams, G. & Milne, P. "Online Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms" *Data Mining and Knowledge Discovery* 8 (2004): 275-300.
23. Major, J. & Riedinger, D. "EFD: A Hybrid Knowledge/Statistical-based system for the Detection of Fraud" *Journal of Risk and Insurance* 69(3) (2002): 309-324.
24. Williams, G. "Evolutionary Hot Spots Data Mining: An Architecture for Exploring for Interesting Discoveries" *Proc. Of PAKDD99* (1999)
25. He, H., Graco, W. & Yao, X. "Application of Genetic Algorithms and k -Nearest Neighbour Method in Medical Fraud Detection" *Proc. of SEAL1998*, (1999): 74-81.

26. Cox, E. "A Fuzzy System for Detecting Anomalous Behaviors in Healthcare Provider Claims" *Intelligent Systems for Finance and Business*, (1995): 111-134.
27. Wheeler, R. & Aitken, S. "Multiple Algorithms for Fraud Detection. *Knowledge-Based Systems*" 13(3) (2000): 93-99.
28. Fan, W. "Systematic Data Selection to Mine Concept- Drifting Data Streams" *Proc. of SIGKDD04* (2004): 128-137.
29. Chen, R., Chiu, M., Huang, Y. & Chen, L. "Detecting Credit Card Fraud by Using Questionnaire-Responded Transaction Model Based on Support Vector Machines" *Proc. of IDEAL2004* (2004): 800-806.
30. Chiu, C. & Tsai, C. "A Web Services-Based Collaborative Scheme for Credit Card Fraud Detection" *Proc. of 2004 IEEE International Conference on e-Technology, e-Commerce and e-Service*. (2004)
31. Foster, D. & Stine, R. "Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy" *Journal of American Statistical Association* 99 (2004): 303-313.
32. Kim, M. & Kim, T. "A Neural Classifier with Fraud Density Map for Effective Credit Card Fraud Detection" *Proc. Of IDEAL2002* (2002):378-383.
33. Maes, S., Tuyls, K., Vanschoenwinkel, B. & Manderick, B. "Credit Card Fraud Detection using Bayesian and Neural Networks" *Proc. of the 1st International NAISO Congress on Neuro Fuzzy Technologies* (2002)
34. Syeda, M., Zhang, Y. & Pan, Y. "Parallel Granular Neural Networks for Fast Credit Card Fraud Detection" *Proc. of the 2002 IEEE International Conference on Fuzzy Systems* (2002)
35. Bolton, R. & Hand, D. "Unsupervised Profiling Methods for Fraud Detection" *Credit Scoring and Credit Control VII*. (2001)
36. Bentley, P., Kim, J., Jung., G. & Choi, J. "Fuzzy Darwinian Detection of Credit Card Fraud" *Proc. of 14th Annual Fall Symposium of the Korean Information Processing Society*. (2000)
37. Brause, R., Langsdorf, T. & Hepp, M. "Neural Data Mining for Credit Card Fraud Detection" *Proc. of 11th IEEE International Conference on Tools with Artificial Intelligence* (1999)

38. Chan, P., Fan, W., Prodromidis, A. & Stolfo, S. "Distributed Data Mining in Credit Card Fraud Detection" *IEEE Intelligent Systems* 14 (1999): 67-74
39. Cortes, C., Pregibon, D. & Volinsky, C. "Computational Methods for Dynamic Graphs" *Journal of Computational and Graphical Statistics* 12 (2003): 950-970.
40. Cahill, M., Chen, F., Lambert, D., Pinheiro, J. & Sun, D. "Detecting Fraud in the Real World" *Handbook of Massive Datasets* (2002): 911-930.
41. Rosset, S., Murad, U., Neumann, E., Idan, Y. & Pinkas, G. "Discovery of Fraud Rules for Telecommunications - Challenges and Solutions" *Proc. of SIGKDD99*, (1999): 409-413.
42. Kim, H., Pang, S., Je, H., Kim, D. & Bang, S. "Constructing Support Vector Machine Ensemble" *Pattern Recognition* 36 (2003): 2757-2767.
43. Burge, P. & Shawe-Taylor, J. "An Unsupervised Neural Network Approach to Profiling the Behavior of Mobile Phone Users for Use in Fraud Detection" *Journal of Parallel and Distributed Computing* 61 (2001): 915-925.
44. Moreau, Y., Lerouge, E., Verrelst, H., Vandewalle, J., Stormann, C. & Burge, P. BRUTUS: A Hybrid System for Fraud Detection in Mobile Communications" *Proc. of European Symposium on Artificial Neural Networks*, (1999): 447-454.
45. Murad, U. & Pinkas, G. "Unsupervised Profiling for Identifying Superimposed Fraud" *Proc. of PKDD99* (1999)
46. Barse, E., Kvarnstrom, H. & Jonsson, E. "Synthesizing Test Data for Fraud Detection Systems" *Proc. of the 19th Annual Computer Security Applications Conference*, (2003): 384-395.
47. McGibney, J. & Hearne, S. "An Approach to Rules-based Fraud Management in Emerging Converged Networks" *Proc. Of IEI/IEEE ITSRS 2003* (2003)
48. Bhargava, B., Zhong, Y., & Lu, Y. "Fraud Formalisation and Detection." *Proc. of DaWaK2003*, (2003): 330-339.
49. Sherman, E. "Fighting Web Fraud" *Newsweek* June 10 2002.
50. Bonchi, F., Giannotti, F., Mainetto, G., Pedreschi, D. "A Classification-based Methodology for Planning Auditing Strategies in Fraud Detection" *Proc. of SIGKDD99*, (1999): 175-184.

51. Shao, H., Zhao, H. & Chang, G. "Applying Data Mining to Detect Fraud Behavior in Customs Declaration" *Proc. of 1st International Conference on Machine Learning and Cybernetics*, (2002): 1241-1244.
52. Ormerod, T., Morley, N., Ball, L., Langley, C., Spenser, C. "Using Ethnography to Design a Mass Detection Tool (MDT) for the Early Discovery of Insurance Fraud" CHI '03 extended abstracts on Human factors in computing systems, (2003): 650 - 651
53. Chan CL, Lan CH "A data mining technique combining fuzzy sets theory and Bayesian classifier—an application of auditing the health insurance fee" (2001)
54. Viveros MS, Nearhos JP, Rothman MJ "Applying data mining techniques to a health insurance information system" *Proc. of 22th International Conference on Very Large Data Bases*, (1996): 286 - 294
55. Williams, G., Huang, Z. "Mining the Knowledge Mine: The Hot Spots Methodology for Mining Large Real World Databases" *Lecture Notes in Computer Science* (1997): 340-348
56. Lee J., Huang K., Jin j. "A survey on statistical methods for health care fraud detection, *Journal of Health Care Manage Science*" (2007): 275-287
57. "USA's National Health Care Anti-Fraud Association web page" 2009 <<http://www.nhcaa.org/eweb/StartPage.aspx>>
58. Ortega, P., Figueroa, C., Ru, G. "A Medical Claim Fraud/Abuse Detection System based on Data Mining: A Case Study in Chile" *Proc. of DMIN'06* (2007) : 224-231
59. Yang, W. and Hwang, S. "A Process-Mining Framework for the Detection of Healthcare Fraud and Abuse" *Expert Systems with Applications*, 31 (2006) : 56-68
60. Ben-Gal, I. "Outlier Detection. *Data Mining and Knowledge Discovery Handbook*" (2005): 131-146
61. Papadimitriou S., Kitawaga H., Gibbons P.G., Faloutsos C., "LOCI: Fast Outlier Detection Using the Local Correlation Integral" Intel research Laboratory Technical report no. IRP-TR-02-09 (2002)
62. Knorr E., Ng R., "A unified approach for mining outliers" *Proc. of Knowledge Discovery KDD*, (1997): 219-222

63. Knorr E., Ng R., "Algorithms for mining distance-based outliers in large datasets" *Proc. of 24th Int. Conf. Very Large Data Bases (VLDB)*, (1998): 392-403
64. Knorr, E., Ng R., Tucakov V. "Distance-based outliers: Algorithms and applications" *VLDB Journal: Very Large Data Bases*, 8(3-4) (2000): 237-253
65. Knorr E. M., Ng R. T., Zamar R. H., "Robust space transformations for distance based operations" *Proc. of the 7th International Conference on Knowledge Discovery and Data-mining (KDD01)*, (2001): 126-135
66. Fawcett T., Provost F., "Adaptive fraud detection," *Data-mining and Knowledge Discovery*, 1(3), (1997): 291–316,
67. Rousseeuw P., Leory A., "Robust Regression and Outlier Detection" *Wiley Series in Probability and Statistics* (1987)
68. Acuna E., Rodriguez C. A., "Meta analysis study of outlier detection methods in classification," *Proc. of IPSI* (2004)
69. Ramaswamy S., Rastogi R., Shim K., "Efficient algorithms for mining outliers from large data sets," *Proc. of the ACM SIGMOD International Conference on Management of Data* (2000)
70. Kaufman L., Rousseeuw P.J., "Finding Groups in Data: An Introduction to Cluster Analysis" *Wiley*, New York, (1990)
71. Barbara D., Chen P., "Using the fractal dimension to cluster datasets," *Proc. of ACM KDD* (2000): 260-264
72. Lu C., Chen D., Kou Y., "Algorithms for spatial outlier detection," *Proc. of the 3rd IEEE International Conference on Data-mining (ICDM'03)* (2003)
73. Haining R., "Spatial Data Analysis in the Social and Environmental Sciences" *Cambridge University Press* (1993)
74. Haslett J., Brandley R., Craig P., Unwin A., Wills G., "Dynamic Graphics for Exploring Spatial Data With Application to Locating Global and Local Anomalies," *The American Statistician*, 45 (1991): 234–242
75. Panatier Y., Variowin. "Software for Spatial Data Analysis in 2D"., *Springer- Verlag*, New York, (1996)

76. Schiffman S. S., Reynolds M. L., Young F. W., "Introduction to Multidimensional Scaling: Theory, Methods and Applications" New York: Academic Press, (1981)
77. Penny K. I., Jolliffe I. T., "A comparison of multivariate outlier detection methods for clinical laboratory safety data," *The Statistician* 50(3) (2001): 295-308,
78. Shekhar S., Lu C. T., Zhang P., "Detecting Graph-Based Spatial Outlier: Algorithms and Applications (A Summary of Results)" *Proc. of the Seventh ACM-SIGKDD Conference on Knowledge Discovery and Data Mining*, SF, CA, (2001)
79. Shekhar S., Lu C. T., Zhang P., "Detecting Graph-Based Spatial Outlier," *Intelligent Data Analysis: An International Journal*, 6(5) (2001): 451–468
80. Shekhar S., Lu C. T., Zhang P., "A Unified Approach to Spatial Outliers Detection," *GeoInformatica, an International Journal on Advances of Computer Science for Geographic Information System*, 7(2) (2003)
81. Haglin, D., Manning, A. "On Minimal Infrequent Itemset Mining" *Proc. of DMIN'2007*, (2007): 141-147
82. Kim, J., Ong, A., Overill, R.E. "Design of an artificial immune system as a novel anomaly detector for combating financial fraud in the retail sector" *Proc. of Evolutionary Computation Congress 2003*, 1(2003): 405- 412
83. Ghosh, R., Reilly, D. "Credit Card Fraud Detection with a Neural-Network" *Proc. of the Twenty-Seventh Annual Hawaii International Conference on System Sciences* (1994)
84. Ezawa, K. and Norton, S. "Constructing Bayesian networks to predict uncollectible telecommunications accounts" *IEEE Expert* 11-5 (1996): 45–51
85. Viaene, S., Derrig, R., Dedene, G. "A case study of applying boosting naive Bayes to claim fraud diagnosis" *Knowledge and Data Engineering, IEEE Transactions* 16-5 (2004): 612-620.
86. Fan, W., "Systematic data selection to mine concept-drifting data streams" *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (2004): 128 – 137
87. Wang, H., Fan, W., Yu, P. & Han, J "Mining Concept- Drifting Data Streams Using Ensemble Classifiers" *Proc. Of SIGKDD03*, (2003): 226-235.

88. Bonchi, F., Giannotti, F., Mainetto, G., Pedreschi, D. "A Classification-based Methodology for Planning Auditing Strategies in Fraud Detection" *Proc. of SIGKDD99*, (1999): 175-184.
89. Shao, H., Zhao, H. & Chang, G. "Applying Data Mining to Detect Fraud Behaviour in Customs Declaration" *Proc. of 1st International Conference on Machine Learning and Cybernetics*, (2002): 1241-1244.
90. Pathak, J., Vidyarthi, N. & Summers, S. "A Fuzzy-based Algorithm for Auditors to Detect Element of Fraud in Settled Insurance Claims" Odette School of Business Administration, (2003)
91. Cortes, C. & Pregibon, D. "Signature-Based Methods for Data Streams" *Data Mining and Knowledge Discovery* 5 (2001): 167-182.
92. Brockett, P., Xia, X. & Derrig, R. "Using Kohonen's Self Organising Feature Map to Uncover Automobile Bodily Injury Claims Fraud" *Journal of Risk and Insurance* 65(2) (1998): 245-274.

APPENDIX A

SAMPLE MODEL OUTPUT

Input File: prescriptions.xlsx contains:

87785 lines,

26419 prescriptions,

2659 drugs,

332 diagnoses.

963 active ingredients,

minimum age= 0, maximum age=85

FrTh_MedicamentDiagnosis = 0.80; %Fraud Threshold for Medicament and Diagnosis risks

FrTh_MedicamentSex = 0.90; % FraudThreshold for Medicament and Sex risks

FrTh_MedicamentAge = 0.96; % Fraud Threshold for Medicament and Age risks

FrTh_MedicamentMedicament = 0.80; % Fraud Threshold for Medicament and Medicament risks

FrTh_DiagnosisCost = 0.85; % Fraud Threshold for Diagnosis and Cost risks

PRESCRIPTION AUDITING:

Prescription No: 88261

Prescription No: 124608

Prescription No: 127947

Prescription No: 143484

Prescription No: 143485

Medicament: TILCOTIL 20 MG 10 Diagnosis: ULCER non-compatible, Risk: 0.93

Medicament: TONIMER NORMAL SPREY %100 125 ML Diagnosis: ULCER non-compatible, Risk: 0.87

Diagnosis: ULCER Cost: 115 non-compatible, Risk: 0.86

Prescription No: 159113

Medicament: FLIXONASE AQUEOUS NASAL SPREY 120 DOZ Diagnosis: BACTERIAL OTITIS MEDIA non-compatible, Risk: 0.94

Medicament: KONGEST PILL 30 Diagnosis: BACTERIAL OTITIS MEDIA non-compatible, Risk: 0.87

Medicament: SEDERGINE VIT-C EFERVESANT PILL 20 Diagnosis: BACTERIAL OTITIS MEDIA non-compatible, Risk: 0.95

Medicament: KONGEST PILL 30 Diagnosis: BACTERIAL OTITIS MEDIA non-compatible, Risk: 0.87

Medicament: KONGEST PILL 30 Diagnosis: BACTERIAL OTITIS MEDIA non-compatible, Risk: 0.87

Medicament: SEDERGINE VIT-C EFERVESANT PILL 20 Diagnosis: BACTERIAL OTITIS MEDIA non-compatible, Risk: 0.95

Prescription No: 159116

Medicament: KONGEST PILL 30 Diagnosis: ALLERGIC RINIT non-compatible, Risk: 0.81

Medicament: AUGMENTIN-BID FILM PILL 1000 MG 10 Diagnosis: ALLERGIC RINIT non-compatible, Risk: 0.88

Medicament: KONGEST PILL 30 Diagnosis: ALLERGIC RINIT non-compatible, Risk:
0.81

Medicament: KONGEST PILL 30 Diagnosis: ALLERGIC RINIT non-compatible, Risk:
0.81

Prescription No: 159221

Prescription No: 159358

Medicament: CEFATIN 500 LAKPILL 500 MG 10 Diagnosis: COMPLEXION DISEASE
non-compatible, Risk: 0.81

Diagnosis: COMPLEXION DISEASE Cost: 130 non-compatible, Risk: 0.99

Prescription No: 159368

Prescription No: 159387

Medicament: CALCIUM SANDOZ+VITAMIN C EFERVESANT PILL 10 (NOVARTIS)
Diagnosis: BRONCHIECTASIS non-compatible, Risk: 0.86

Medicament: CATAFLAM PILL 50 MG 20 DR Diagnosis: BRONCHIECTASIS non-
compatible, Risk: 0.91

Prescription No: 159510

Prescription No: 159544

Prescription No: 159553

Prescription No: 159696

Prescription No: 159706

Medicament: PHARMATON CAPSULE 30 CAP Diagnosis:
HYPERCHOLESTEROLEMIA non-compatible, Risk: 0.87

Medicament: PHARMATON CAPSULE 30 CAP Diagnosis: HYPERTHYROID non-compatible, Risk: 0.98

Prescription No: 159708

Medicament: METPAMID PILL 10 MG 30 Diagnosis: BACTERIAL COLON DISEASES non-compatible, Risk: 0.90

Medicament: REFLOR CAPSULE 250 MG 10 CAP. Diagnosis: BACTERIAL COLON DISEASES non-compatible, Risk: 0.86

Prescription No: 159711

Medicament: TEARS NATURALE FREE 0,8 ML 32 TUBE Diagnosis: DIABETES MELLITUS , Risk: 0.96

Medicament: CALCIUM SANDOZ FORTE EFERVESANT PILL 10 Diagnosis: DIABETES MELLITUS , Risk: 0.88

Medicament: CALCIUM SANDOZ FORTE EFERVESANT PILL 10 Diagnosis: DIABETES MELLITUS , Risk: 0.88

Medicament: TEARS NATURALE FREE 0,8 ML 32 TUBE Diagnosis: OSTEOPOROSIS non-compatible, Risk: 0.92

Prescription No: 159938

Prescription No: 159959

Medicament: DOLVEN PEDIATRIC 100 ML SYRUP Diagnosis: KAS-ISKELET SISTEMI HASTALIKLARI non-compatible, Risk: 0.99

Prescription No: 160000

Prescription No: 160054

Prescription No: 160080

Prescription No: 160217

Prescription No: 160441

Prescription No: 160459

Prescription No: 160503

Prescription No: 160507

Medicament: ANDOREX 30 ML SPREY Diagnosis: BRONCHIECTASIS non-compatible,

Risk: 0.83

non-compatible, Risk: 0.90

Medicament: UMCA 50 ML SOLUTION Diagnosis: BRONCHIECTASIS non-compatible,

Risk: 0.83

Prescription No: 160512

Medicament: ANDOREX 30 ML SPREY Diagnosis: BRONCHIECTASIS non-compatible,

Risk: 0.90

Medicament: OTRIVINE PEDIATRIC NASAL SPREY %0.05 10 ML Diagnosis:
BRONCHIECTASIS non-compatible, Risk: 0.89

Medicament: OTRIVINE PEDIATRIC NASAL SPREY %0.05 10 ML Age: 15 non-
compatible, Risk: 0.99

Prescription No: 160521

Prescription No: 160551

Prescription No: 160787

Medicament: SERETIDE DISKUS 500 MCG 60 DOSE Diagnosis: BACTERIAL UPPER RESPIRATION TUBE INFECTION non-compatible, Risk: 0.89

Medicament: SERETIDE DISKUS 500 MCG 60 DOSE Diagnosis: BACTERIAL UPPER RESPIRATION TUBE INFECTION non-compatible, Risk: 0.89

Diagnosis: BACTERIAL UPPER RESPIRATION TUBE INFECTION Cost: 320 non-compatible, Risk: 1.00