IDENTIFING THE FRAUD DETECTION IN HEALTH CARE SYSTEM USING MACHINE LEARNING

Mrs. Arulmozhi. B[*1] , Mrs Arithi. S[*2], Ms. Sivasankari. A[*3]

[*1]Head of the department ,Department of BCA,D.K.M College for women (Autonomous),Vellore.
[*2]Department of Computer science, D.K.M College for women (Autonomous),Vellore.
[*3]Head of the department ,Department of computer science,D.K.M College for women (Autonomous),Vellore.

Abstract- Fraud is the criminal 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 give the impression 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.

Key Words:*Web Usage Mining, Fraud Detection, Prescription Fraud, Data Mining, social security, prescription fraud detection*

I. 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].

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.

II.RELATED WORK

There are various resources relating to fraud detection. Fraud detection being a relatively large field, most of the papers on this subject considers 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.

2.1.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.

2.2 Available Data For Fraud Detection

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 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. Semi-supervised Approaches

SVM (RSVM) is employed to learn a personalized ranking function for rank adaptation of the results according to the user content and location preferences while receiving the user's preferences. From the search results of the document features, a set of content concepts and location concepts can be extracted for a given query. Since each document can be represented by a feature vector, it can be treated as a point in the feature space. SVM aims at finding a linear ranking function which holds many document preference pairs as possible, when preference pairs are used as the input. An adaptive implementation, SVM light available at, is used in our experiments. The two main issues in the SVM training process are discussed below:

1. How to extract the feature vectors for a document,
2. How to combine the content and location weight vectors into one integrated weight vector.

III. PREVIOUS IMPLEMENTATIONS

3.1 Classification of fraudulent behaviors

Three parties may be involved in the commission of health care fraud. They are (a) service providers, including doctors, hospitals, ambulance companies, and laboratories; (b) insurance subscribers, including patients and patients' employers; and (c) insurance carriers, who receive regular premiums from their subscribers and pay health care costs on behalf of their subscribers, including governmental health departments and private insurance companies. According to which party commits the fraud, fraudbehaviors can be classified as follows

3.1.1 Service providers' fraud:

- Billing services that are not actually performed; Unbundling, i.e., billing each stage of a procedureas if it were a separate treatment;
- Upcoding, i.e., billing more costly services than theone actually performed; for example, "DRG creep"is a popular type of upcoding fraud, which classifies patients' illness into the highest possible treatment category in order to claim more reimbursement;
- Performing medically unnecessary services solelyfor the purpose of generating insurance payments;
- Misrepresenting non-covered treatments as medicallynecessary covered treatments for the purpose ofobtaining insurance payments;
- Falsifying patients'diagnosis and/or treatment historiesto justify tests, surgeries, or other procedures that arenot medically necessary

3.1.2 Insurance subscribers' fraud:

- Falsifying records of employment/eligibility forobtaining a lower premium rate;
- Filing claims for medical services which are not actually received;
- Using other persons' coverage or insurance card toillegally claim the insurance benefits.

3.1.3 Insurance carriers' fraud:

- Falsifying reimbursements;
- Falsifying benefit/service statements

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.

Outlier detection methods in the literature are:

- Univariate methods
- Multivariate methods

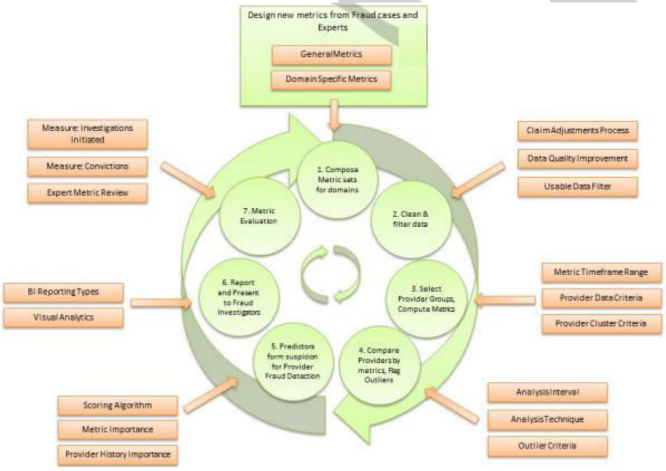

Design new metrics from Fraud cases and Experts
GeneralMetrics
Domain Specific Metrics
1. Compose Metric sets for domains
2. Clean & filter data
3. Select Provider Groups, Compute Metrics
4. Compare Providers by metrics, Flag Outliers
5. Predictors form suspicion for Provider Fraud Detection
6. Report and Present to Fraud Investigators
7. Metric Evaluation
Measure: Investigations Initiated
Measure: Convictions
Expert Metric Review
Claim Adjustments Process
Data Quality Improvement
Usable Data Filter
Metric Timeframe Range
Provider Data Criteria
Provider Cluster Criteria
BI Reporting Types
Visual Analytics
Scoring Algorithm
Metric Importance
Provider History Importance
AnalysisInterval
AnalysisTechnique
Outlier Criteria

Fig1: Method for Applying Outlier Detection to Healthcare Fraud

The set of metrics does not necessarily have to be large; on the contrary, often 25 to 30 features are sufficient. If hundreds of metrics have to be designed, the absolute amount of outliers is increased as well, which eventually will result in all providers displaying outlying behavior for some metrics. Metric identification is dependent on fraud experts and is an iterative process to find a set of metrics that works effectively. For our case study, we initially developed over 100 behavioral metrics. This list wassubsequently refined to fifteen that could be appliedto a relatively homogenous provider pool in the dental domain, feasible for implementation within our research case constraints.

IV. SYSTEM IMPLEMETNATION

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.

4.1 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

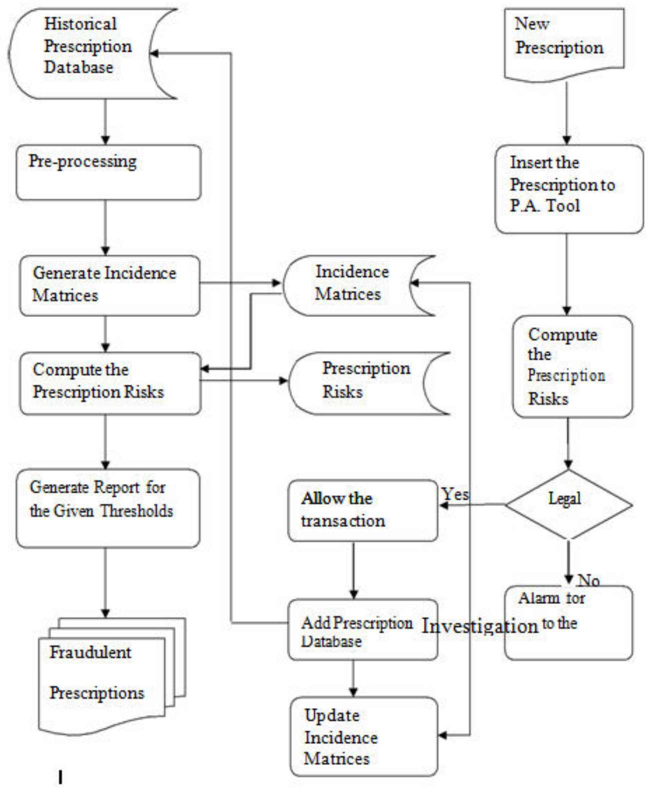

Historical Prescription Database
Pre-processing
Generate Incidence Matrices
Compute the Prescription Risks
Generate Report for the Given Thresholds
Fraudulent Prescriptions
Incidence Matrices
Prescription Risks
Allow the transaction
Add Prescription Database
Update Incidence Matrices
New Prescription
Insert the Prescription to P.A. Tool
Compute the Prescription Risks
Legal
Yes
No
Alarm for Investigation to the

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

4.2 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.

4.3.1Medicament – 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 features have two entities: female and male. Thus the size of this matrix is 2*(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:

4.4 ONLINE DECISION

The profile-based personalization contributes little, even reduces the search quality while exposing the profile to a server would for sure risk the user's privacy. To address this problem, we develop an online mechanism to decide whether to personalize a query. The basic idea is if a distinct query is identified during generalization, the entire runtime profiling will be aborted and the query will be sent to the server without a user profile.

$$risk_{MS}(i,j)=\frac{\exp(-MS(i,j\ Max_{MS}(i))-\exp(-1))}{1-\exp(-1)} \qquad \text{E.q 4.1}$$

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*

4.3.2 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*.

$$risk_{MD}(i,j) = \frac{\exp(-MD(i,j)\ Max_{MD}(i)) - \exp(-1)}{1-\exp(-1)}$$

The risk matrix of the Medicament and Diagnosis domain is defined as

$$MDR(i,j) = risk_{MD}(i,j)$$

4.3.2 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 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 *iis* $r_i = Max_j(i) - Min_j(i)$. Thus, the modified risk formulation would be:

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

$$V_i = \sum_k ((k * MA(i,k)) \quad \sum_k MA(i,k)) \quad \text{and} \quad d_i(j) = |j - V_i$$

and the risk matrix of the Medicament and Age domain is defined as

$$MAR(i,j) = risk_{MA}(i,j)$$

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.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.

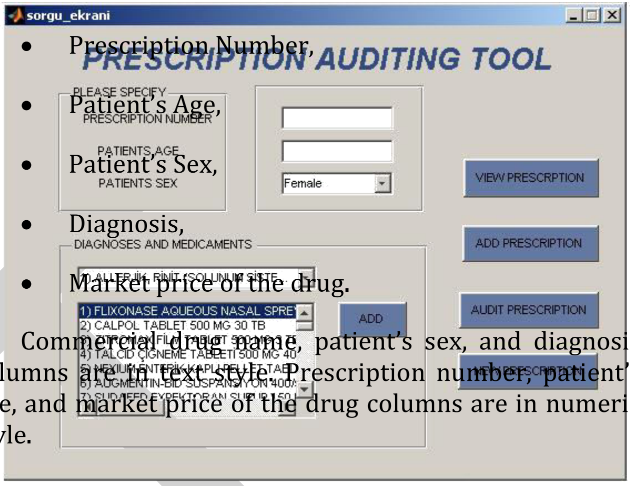
V. COMPUTATIONAL RESULT:

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.

5.1 Online Processing



PATIENTS SEX: Female

1) ALLERJİK RİNİT (SOLUNUM SİSTE...

1) FLIXONASE AQUEOUS NASAL SPREY
2) CALPOL TABLET 500 MG 30 TB
3) ZITROMAX FİLM TABLET 500 MG 3 TB
4) TALCID ÇİGNEME TABLETİ 500 MG 40
5) NEXIUM ENTERİK KAPLI PELLET TAB
6) AUGMENTIN-BID SUSPANSIYON 400/
7) SUDAFED EXPEKTORAN ŞURUP 150

Fig 3: 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.

5.2 OFFLINE FRAUD DETECTION

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.

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

Table 1: Prescription Example-1

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 ith 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.

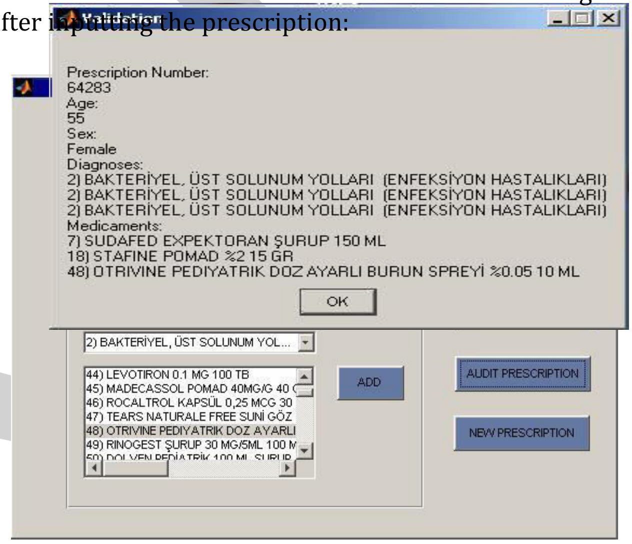
P.No	Medicament Name	Age	Sex	Diagnosis	Active Ing.	Price (TL)
431603	*Euthyrox*	49	F	*Osteoporosis*	*Levotiroxin*	2.63
431603	*Fosamax*	49	F	*Osteoporosis*	*Alendronate*	39.46
431603	*Zyrtec*	49	F	*Osteoporosis*	*Setirizine hcl*	10.77

Table 2: Prescription Example-2

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.

5.3 ONLINE FRAUD DETECTION

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:



Prescription Number:
64283
Age:
55
Sex:
Female
Diagnoses:
2) BAKTERİYEL, ÜST SOLUNUM YOLLARI (ENFEKSİYON HASTALIKLARI)
2) BAKTERİYEL, ÜST SOLUNUM YOLLARI (ENFEKSİYON HASTALIKLARI)
2) BAKTERİYEL, ÜST SOLUNUM YOLLARI (ENFEKSİYON HASTALIKLARI)
Medicaments:
7) SUDAFED EXPEKTORAN ŞURUP 150 ML
18) STAFINE POMAD %2 15 GR
48) OTRIVINE PEDIYATRIK DOZ AYARLI BURUN SPREYİ %0.05 10 ML

OK
2) BAKTERİYEL, ÜST SOLUNUM YOL...
44) LEVOTIRON 0.1 MG 100 TB
45) MADECASSOL POMAD 40MG/G 40
46) ROCALTROL KAPSÜL 0,25 MCG 30
47) TEARS NATURALE FREE SUNİ GÖZ
48) OTRIVINE PEDIYATRIK DOZ AYARLI
49) RINOGEST ŞURUP 30 MG/5ML 100 M
ADD
AUDIT PRESCRIPTION
NEW PRESCRIPTION

Fig 4: Inserting a Prescription to the Prescription Auditing Tool

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

Fig 5: Validation Message Box

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 arefor 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

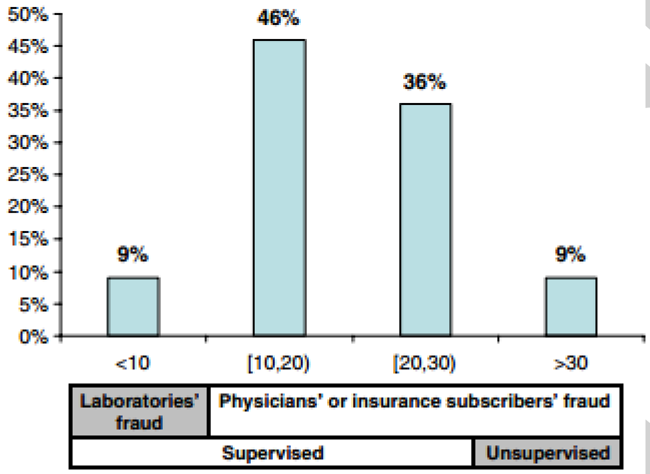


Fig 6: Percentages of papers in which certain numbers of features are used

These features include the percentage distributions over several test categories (chemical, microbiology, and immunology), the number of different patients dealt with, and the frequency of tests performed. In addition, the paper by Major and Riedinger [27] classified potentially useful features into five categories, capturing the major aspects of service providers' profiles. These categories of features are shown in Table 1.Automatic computer algorithms for feature selection were developed by a group of researchers focusing on the detection of service providers' fraud from the claim data in the Bureau of National Health Insurance (NHI) in Taiwan. Their algorithms can only be applied to the service providers whose practices, if legitimate, are supposed to follow well-defined standard clinical pathways. The concept of clinical pathways was initially developed in the early 1990s. A clinical pathway is a flow chart that sequences the necessary medical care activities (e.g. activities involved in diagnosis and treatment) given to a patient or a patient group with a certain disease.

CONCLUSION

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 toWe 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.

REFERENCES

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 Dat”a, 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 Engineering16(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.