Implementation of Data Mining in Medical Fraud Detection

Jacqulin Margret J
Assistant Professor
Department of Computer Science and Engineering
Muthayammal Engineering College,
Rasipuram

Shrijina Sreenivasan
Assistant Professor
Department of Computer Science and Engineering
Malabar College of Engineering and Technology,
Thrissur

ABSTRACT

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, to develop powerful means to analyze and extract interesting knowledge from data. Fraudulent healthcare claims increase the burden on the society. The healthcare fraud detection requires compilation of potentially huge data, involving complex computation and sorting operations. Once such frauds have been detected and classified, data cleaning is applied to it which helps to remove the noise and inconsistencies in the data thereby enhancing its quality. This technique can be used to detect the sale of potentially dangerous medicine by pharmacists thereby preventing such medical fraud.

Keywords

Medical Fraud, Data Mining, Data Cleaning, Classification.

1. INTRODUCTION

Data Mining is a process of discovering trends and patterns in the data obtained from various data sources. It is considered as an increasingly important tool by modern businesses to transform data into business intelligence giving an informational advantage and thus providing newer business tactics that in turn helps to gain more profits. Data mining requires data to be selected, preprocessed and transformed into a form useful to the user, before applying the data mining algorithm. Data warehousing is a repository for storing data that have been refined from multiple data sources. The data so collected is then extracted for the process of data mining.

The Healthcare industry is among the most information intensive industries. Medical information, knowledge and data keep growing on a daily basis. The ability to use these data to extract useful and new information for quality healthcare is crucial. Medical frauds can be anything like providing false and intentionally misleading statements to patients, submitting false bills or claims for service, falsifying medical records or reports, lying about credentials or qualifications, unnecessary medical treatment or drug prescription. In this thesis, the focus is upon detecting false medicine composition. and show that data mining can be applied to the medical databases, which will classify the medicinal data as fake or original, with a reasonable accuracy.

2. RELATED WORKS

When medical institutions apply data mining on their existing data, they can discover new, useful and potentially life-saving knowledge that otherwise would have remained inert in their databases. For instance, an ongoing study on hospitals and safety found that about 87% of hospital deaths in the United States could have been prevented, had hospital staff (including doctors) been more careful in avoiding errors. By mining hospital records, such safety issues could be flagged and addressed by hospital management and government regulators. Arthur D. Chapman1 in [1] highlighted the importance of data quality as it relates to primary species occurrence data facilitating error checking and cleaning. Nevine M. Labib et al in [5] made use of a data mining tool, Clementine, so as to apply Decision Trees technique. they fed it with data extracted from real-life cases taken from specialized Cancer Institutes. Relevant medical cases details such as patient medical history and diagnosis are analyzed, classified, and clustered in order to improve the disease management. Prem Swaroop Dr Bruce Golden in [6] provided a compact idea of the various data mining techniques used with the help of a health care application. Ruben D. Canlas Jr.in [7] provides a survey of current techniques of KDD, using data mining tools for healthcare and public health. It also discusses critical issues and challenges associated with data mining and healthcare in general. Ming Li in [8] described the various predictive modeling techniques with a detailed outlook on the various classification techniques in both data mining and neural networks.

3. DATA CLEANING PROCESS

To ensure the quality of data, it has to undergo a lot of preprocessing steps to maintain the quality of the results that are obtained. Data warehouse also needs a consistent integration of quality data. Data extraction, cleaning, and transformation comprise the majority of the work of building a data warehouse.

Data profiling is the process of examining the data available in an existing data source (e.g. a database or a file) and collecting statistics and information about that data. The purpose of these statistics may be to find out whether existing data can easily be used for other purposes, assess whether metadata accurately describes the actual values in the source database, give metrics on data quality, including whether the data conforms to particular standards or patterns, assess the risk involved in integrating data for new applications, including the challenges of joins, understanding data challenges early in any data intensive project, so that late project surprises are avoided. Finding data problems late in the project can lead to delays and cost overruns.

Data cleaning also plays a pivotal role in the preprocessing of data. Some of the data cleaning tasks include filling in missing values, identifying outliers and smoothing out noisy data, correcting inconsistent data and resolving redundancies caused by data integration. In the proposed technique, the data cleansing process takes two inputs:

1) Data required to be cleaned and
2) Rules for cleaning the data.

This is the area where actual data cleansing processing is done based on certain techniques for the removal of noisy, inconsistent and incomplete data, where the output is an error-free and consistent data that is ready to be loaded into the data warehouse. This output data is standardized, uniform, accurate and complete in accordance with medical correctness. The cleaned data not only enhances data quality but also provides better processing speed and performance.

3.1. TYPES OF RULES

Extraction Rules

These are the rules that help data cleansing process in the selection of required data elements that need clean up. The rule is entered into a Rule Repository only once based on data profiling results and user input. These rules mainly describe which sources and their attributes require data cleaning. These rules would not change until sources are changed or added into the system as data warehouse source or the user has experienced any change in existing rules or found new rules that needed to be applied.

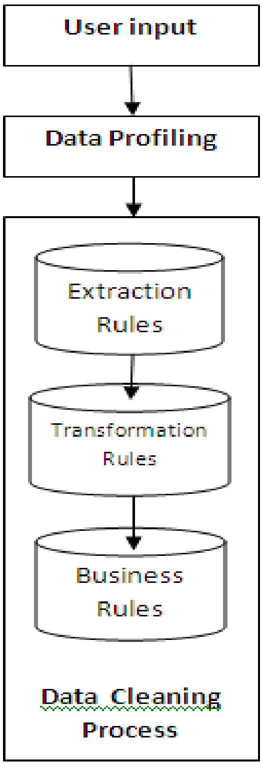


Fig 1: Existing Component-level Framework structure

Transformation Rules

These are the rules that perform core data cleansing operations. The transformation rules are related with the correction and removal of dirty data values in records. This correction could be of different types e.g. dummy values, absence of values correction, contradicting values correction, rules violation correction etc. These rules applied on data one by one by fetching from rules configuration repository, automatically by process.

Business Rules

Business rules are the type of rules that could be required to mark data based on the data modeling perspective. The business rules apply on data to transform data as per the target model of data warehouse. These rules could be of different types according to the industry for which the warehouse is developed. It extracts data based on the requirements and the type of output required by the business industry. The business here is medical industry and the rules that they follow.

4. PROPOSED SYSTEM

In this thesis, shallow knowledge is extracted from the given data using SQL queries. Consider the following assumptions: There is a large data warehouse that contains a database of all medicines that are possibly available namely ORIGINALDB. The fake database that exists is called the FAKEDB. In reality, the fake database is assumed to be the database that is possessed by the pharmacist.

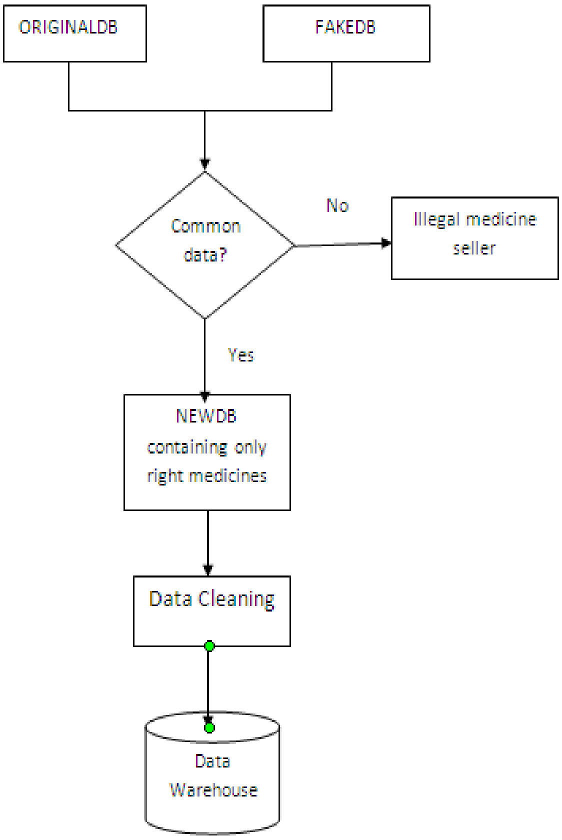


Fig 2: Proposed framework for data Comparison and Cleaning

The fake database is used to detect fraudulent medical composition for use in case of raids or any such purpose. The person who is detecting the fraud should have some hardware device like a barcode reader to detect the composition of all medicines that have to be tested. This detected value is stored into the fake database. The composition is then given as a query to be compared with the original database.

The above mentioned techniques are discussed below with respect to the proposed system.

Data Profiling

Data profiling collects information about that data from various data repositories. Both the original and the fake databases are examined to assess whether the metadata accurately describes the actual values in the source database. The purpose is to clarify at an early stage if the right data is available at the right detail level and any anomalies if exists, can be handled subsequently.

Data transformation

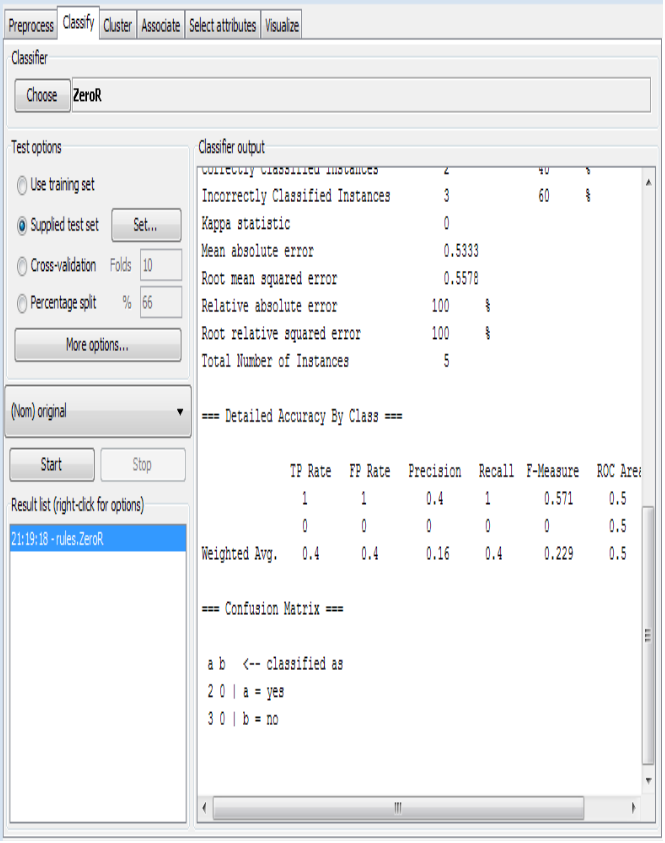
Both the original and the fake databases are compared to detect the mismatches that occur. From the compared data, a new database NEWDB is created which contains only those entries that match in both ORIGINALDB and FAKEDB. The rest of the entries being categorized as that of duplicate medicines with dangerous medical composition that the shopkeeper was trying to sell. This new database thus becomes the registered database of the shopkeeper for further sales thereby preventing him from selling medicines with wrong compositions or even taking an action against him depending upon the seriousness of the crime.

Data Cleaning

After the NEWDB has been formed, the data is cleaned by fixing the duplicate datas, null values and garbage values and is given to the data warehouse, considering it to be the new registered database of that particular shopkeeper.

Data Mining

Classification is a data mining function that assigns items in a collection to target categories or classes[5]. The goal of classification is to accurately predict the label classes. It is a supervised learning technique since it categorizes data based on the class labels that are specified beforehand. It first uses a training set of data that already contain some prescribed class labels for the data in consideration. By applying any kind of a classification technique, some classification rules are obtained. These rules are checked for its correctness based on the test set which is provided. Those rules which satisfy cases in the test data are filtered as a valid rule. These rules are then used for further classification. Here, the name of medicine is considered as label class and the medicine compositions are considered as attributes for classification.



ZeroR

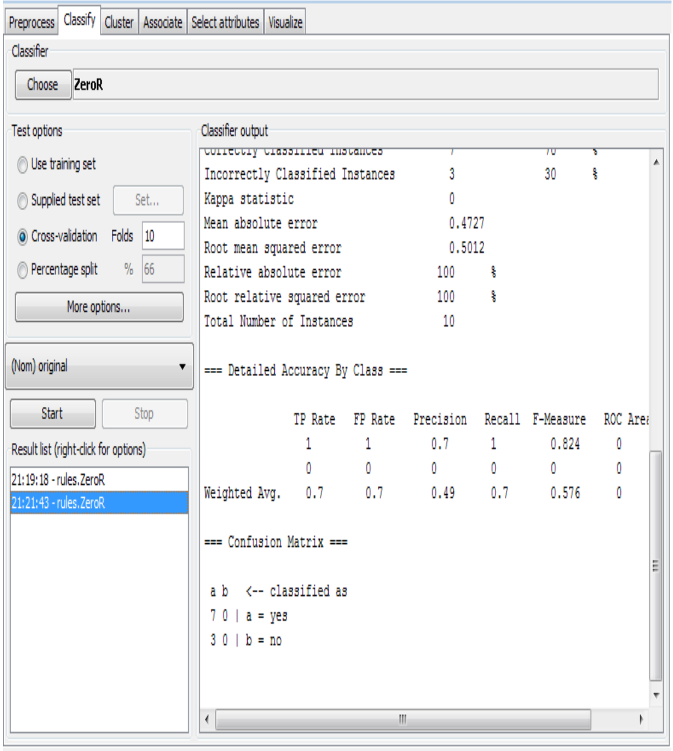
○ Use training set
● Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) original

21:19:18 - rules.ZeroR

```
Correctly Classified Instances          2               40      %
Incorrectly Classified Instances        3               60      %
Kappa statistic                         0
Mean absolute error                     0.5333
Root mean squared error                 0.5578
Relative absolute error               100      %
Root relative squared error           100      %
Total Number of Instances               5

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area
                 1         1          0.4      1         0.571      0.5
                 0         0          0        0         0          0.5
Weighted Avg.    0.4       0.4        0.16     0.4       0.229      0.5

=== Confusion Matrix ===

 a b   <-- classified as
 2 0 | a = yes
 3 0 | b = no
```

Fig 3: Classification using test set



ZeroR

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) original

21:19:18 - rules.ZeroR
21:21:43 - rules.ZeroR

```
Correctly Classified Instances          7               70      %
Incorrectly Classified Instances        3               30      %
Kappa statistic                         0
Mean absolute error                     0.4727
Root mean squared error                 0.5012
Relative absolute error               100      %
Root relative squared error           100      %
Total Number of Instances              10

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area
                 1         1          0.7      1         0.824      0
                 0         0          0        0         0          0
Weighted Avg.    0.7       0.7        0.49     0.7       0.576      0

=== Confusion Matrix ===

 a b   <-- classified as
 7 0 | a = yes
 3 0 | b = no
```

Fig 4: Classification cross-validation

In this, the classification is done with the help of Weka tool. Two methods were carried out. One, using test set and training set and other method is using cross validation. The outputs for the above methods are shown below. It shows the confusion matrix generated for training using test set and cross validation. The confusion matrix[4] is a useful tool for analyzing how well your classifier can recognize tuples of different classes[3].

The following are the database snapshots which have been used in this thesis:

Original Database:

CODE	NAME	C1	C2	C3
1111.11	Oxyteracin	Codine 8mg	Lidocaine 25mg	Sodium 50mg
2924.29	Tylenol1	Codeine 8mg	Paracetamol 325mg	
3024.29	Tylenol2	Codeine 15mg	Paracetamol 300mg	
5024.29	Amplus	Ampicil in 250mg	Lactic Acid 60ms	Sodium 25mg
6014.0	C_Oral	Sulphar 200mg	Trimethopri m 40mg	

Fake Database:

CODE	NAME	C1	C2	C3
1111.11	Oxyteracin	Codine 8mg	Lidocaine 25mg	Sodium 50mg
2924.29	Tylenol1	Codeine 8mg	Paracetamol 305mg	
3024.29	Tylenol2	Codeine 15mg	Paracetamol 300mg	
5024.29	Amplus	Ampicilin 250mg	Lactic Acid 60ms	Sulphar 25mg
6014.0	C_Oral	Sodium 200mg	Trimethoprim 40mg	

New Database:

CODE	NAME	C1	C2	C3
1111.11	Oxytera cin	Codine 8mg	Lidocaine 25mg	Sodium 50mg
3024.29	Tylenol 2	Codeine 15mg	Paracetamo l 300mg	

The new database now contains the common and the matched data after comparison from both the databases. The purpose of the NEWDB is to maintain discrepancy-free cleaned medicinal data of the shopkeeper.

5. CONCLUSION AND FUTURE WORK

This paper describes the application of data mining in medical fraud detection systems. Such a system would help find a pattern that detects discrepancies in medical data. The mined data is then cleaned to remove unwanted, inconsistent and corrupt data. The overall outcome of such an implementation is high quality data that is potentially more accurate and which prevents occurrences of fraud in case of false medical sale and billing. While this procedure has been currently worked out on relational databases, our future work may support multidimensional data and the tuning of queries made to a data warehouse to reduce the execution time of queries.

6. REFERENCES

[1] Principles And Methods Of Data Cleaning, Arthur D. Chapman1

[2] http://en.wikipedia.org/wiki/Data_profiling

[3] Data Mining: Concepts and Techniques Jiawei Han and Micheline Kamber, Morgan Kaufmann, 2001.

[4] http://en.wikipedia.org/wiki/Confusion_matrix

[5] Data Mining for Cancer Management in Egypt Case Study: Childhood Acute Lymphoblastic Leukemia , Nevine M. Labib, and Michael N. Malek

[6] Data Mining: Introduction and a Health Care Application, Prem Swaroop Dr Bruce Golden

[7] Data Mining In Healthcare: Current Applications And Issues, Ruben D. Canlas Jr.

[8] Data Mining ,Ming Li, Department of Computer Science and Technology, Nanjing University Fall 2011 Chapter 10: Predictive Modeling