Application of Clustering Methods to Health Insurance Fraud Detection

Yi Peng[1], Gang Kou[1,*], Alan Sabatka[2], Zhengxin Chen[1], Deepak Khazanchi[1], Yong Shi[3]

[1]Peter Kiewit Institute of Information Science, Technology & Engineering
University of Nebraska at Omaha
Omaha, NE 68182, USA
[2]Mutual of Omaha
Omaha, NE, USA
[3]Chinese Academy of Sciences Research Center on Data Technology & Knowledge Economy
Graduate University of the Chinese Academy of Sciences
Beijing 100080, China
*The corresponding author.
*Email: gkou@mail.unomaha.edu
*Tel: ++1 402 5543429.

ABSTRACT

Health insurance fraud detection is an important and challenging task. Traditionally, insurance companies use human inspections and heuristic rules to detect fraud. As the size of databases increases, the traditional approaches may miss a great portion of fraud for two main reasons. First, it is impossible to detect all health care fraud by manual inspection over large databases. Second, new types of health care fraud emerge constantly. SQL operations based on heuristic rules cannot identify those new emerging fraud schemes. Such a situation demands more sophisticated analytical methods and techniques that are capable of detecting fraud activities from large databases. The goal of this paper is to understand and detect suspicious health care frauds from large databases using clustering technique. Specifically, this paper applies two clustering methods, SAS EM and CLUTO, to a large real-life health insurance dataset and compares the performances of these two methods.

Keyword: Clustering, Insurance Fraud Detection, Database

1 INTRODUCTION

Health care fraud is "the deliberate submittal of false claims to private health insurance plans and/or tax-funded public health insurance programs such as Medicare and Medicaid [1]". According to the National Health Care Anti-Fraud Association's estimation, at least $51 billion is lost to health care fraud in calendar year 2003. Traditionally, heuristic rules are used by health insurance companies to detect frauds. These rules were summarized from previous fraud cases and are used to detect fraud either through human inspection or interaction with an external entity. As the size of databases increases, the traditional fraud detection approach may miss a great portion of fraud for two main reasons. First, it is impossible to detect all health care fraud by manual inspection over large databases. Second, new types of health care fraud emerge constantly. SQL operations based on historical fraud cases cannot identify those new emerging fraud schemes. Such a situation demands more sophisticated analytical methods and techniques that are capable of detecting fraud activities from large databases. Data mining, defined by Bob Klevecz as "Nontrivial extraction of implicit, previously unknown and potentially useful information from data, or the search for relationships and global patterns that exist in databases [2]", provides methods and techniques to solve this problem.

The goal of this paper is to detect suspicious health care frauds from large databases. In order to achieve this goal, this paper applies two clustering methods, SAS EM [3] and CLUTO [4], to a large real-life health care database and compares the performances of these two methods. Specifically, this paper is organized as follows. The next section describes data and preprocessing steps, including feature selection and data transformation. The third section gives an overview of clustering and softwares. The fourth

1-4244-0451-7/06/$20.00 ©2006 IEEE.

section presents and empirically compares the clustering results. The application of the clustering results is also discussed. The last section concludes the paper with future research direction.

2 DATA UNDERSTANDING AND PREPROCESSING

The dataset used in this paper is provided by a US insurance company and contains health claims data from one state of USA. The storage and processing of this data was conducted within the insurance company's infrastructure in compliance with privacy regulations. The dataset has 1,924,426 rows of data and 53 variables. Each data record describes detailed information about claimants, provider, benefit amounts, and services that claimants received. These variables include numeric, character, and date types. After discussing with business experts, we exclude twenty-seven variables that are irrelevant, redundant, or correlated. Irrelevant variables, such as identification numbers, are excluded because they have no influence on fraud detection. Redundant variables are excluded because the similar information can be found in other variables. A typical example is City and State addresses. They can be removed because zip code variable provides city and state information. Inclusion of redundant variables will increase computational complexity and slow down the clustering process. Related variables will affect the final clustering results and are deleted. As a result of variable selection, twenty-six variables are selected for clustering analysis. Since SAS EM and CLUTO require input variables to be numerical, non-numerical types are transformed into numerical values to satisfy SAS EM and CLUTO's requirements.

3 CLUSTERING SOFTWARES

Clustering is an unsupervised classification which groups similar data objects into clusters [5] and has wide applications. According to cluster criterion, clustering algorithms can be categorized as partitioning methods, hierarchical methods, density-based methods, grid-based methods, model-based clustering methods [5]. Different clustering methods are suitable for different datasets and may result in different clusters for a same dataset. Clustering methods help us get insight into the data distribution and are particularly useful when there are no class labels. Since we do not understand the target data clearly and the targeted data do not have class labels, clustering is chose as the modeling technique.

In this study, SAS Enterprise Miner 4.2 and CLUTO 2.1.1, are used for clustering analysis. For both tools, we use their default settings. CLUTO is a software package developed by the Department of Computer Science & Engineering, University of Minnesota for clustering datasets and for analyzing the characteristics of the various clusters [4]. CLUTO has been proved to be able to cluster various datasets in diverse application areas and is capable of handling high-dimensional datasets. SAS EM is a commercial software package provided by SAS to support a collection of DM functions. SAS has years of experiences in data analysis and data mining and SAS EM is used by many industries. Because CLUTO and SAS EM are designed for different purposes and have different features, it is interesting to compare their performances in terms of cluster results, computation time, and usefulness.

4 EMPIRICAL CLUSTERING RESULTS CLUSTERING RESULTS

Most clustering algorithms require the users to input the number of clusters they desires. As mentioned in the previous section, we do not have a clear understanding about the dataset and can not determine the number of clusters. Therefore, we conduct a series of clustering experiments and discuss the results with business experts to decide which numbers of clusters are desirable. We design thirteen different numbers of clusters: 2~10, 20, 30, 50, and 100 using CLUTO and SAS EM. Among these results, we chose 10-way and 20-way to report (Table 1 and 2).

Table 1 10-way Clustering Results

Clusters	SAS EM	CLUTO
1	5915	1885
2	9719	2305
3	24876	4141
4	27371	4297
5	34203	5970
6	34750	6516
7	35089	9917
8	56514	10610
9	111612	14949
10	1584377	1863836

The column clusters indicates individual clusters. CLUTO and SAS EM columns indicate the number of data objects within each cluster and these clusters are listed in an ascending order. After examining the

clustering results, we have the following observations:

- According to our experiments, CLUTO takes less computation time than SAS EM. For example, it takes CLUTO 6.5 minutes and SAS EM several hours to complete a 20-way clustering.

Table 2 20-way Clustering Results

Clusters	CLUTO	SAS EM	Clusters	CLUTO	SAS EM
1	744	2718	11	2171	18630
2	851	3153	12	3929	20063
3	1019	4814	13	4115	21311
4	1126	4839	14	4691	24876
5	1141	6404	15	5802	30335
6	1275	7646	16	6516	53531
7	1279	9549	17	6596	88879
8	1286	9596	18	6681	458784
9	1495	13168	19	8353	530887
10	1520	15828	20	1863836	599415

- Records in CLUTO are concentrated in one cluster, while the results of SAS EM are distributed more evenly. For instance, the largest cluster in 10-way and 20-way SAS EM contain 82.33% and 31.15% data records, while the largest cluster in 10-way and 20-way CLUTO contains 96.85% data records. Use Table 1 and 2 as examples, the smallest cluster in SAS EM 10-way clustering contains 0.3% data records and the smallest cluster in CLUTO contains 0.1% data records; in 20-way clustering, the smallest cluster in SAS EM contains 0.14% data records and the smallest cluster in CLUTO contains 0.04% data records.
- Some of the abnormal clusters that have the smallest number of records may not have distinct values in one or two attributes, instead, they may have slight deviations in many attributes and the sum of these deviations made them different from other clusters.

4.1 Understand clustering results

Assessing cluster quality is the most challenging step in clustering analysis. Different clustering methods may generate very different clusters for the same dataset. Researchers have proposed various quality measures for clustering and most of them can be grouped into two categories: external quality measure and internal quality measures. *External quality measures* require prior knowledge of the structure of dataset to assess the quality of clustering techniques. The quality of clustering technique is judged by the agreement between the discovered clusters and the known information [6]. *Internal quality measures* normally evaluate the clustering solution by checking "how similar the objects are within each cluster and how well the objects of different clusters are separate" [p. 13, 6]. Since the same criteria are used both in discovering and in evaluating the clusters, internal quality measures are criticized as providing no independent judgment. In this study, we select external quality measure to evaluate cluster quality because we have some previously known suspicious IDs. A meaningful clustering solution should generate clusters within which the ratio of suspicious records to total records in that cluster is high.

4.2 Assess cluster quality using suspicious IDs

Because business users are particularly interested in a subset of the data, which has more than 50% of the records, we will concentrate on this subset in the following analysis.

After removing the suspicious ID variable, we rerun 100-way clustering using CLUTO and SAS EM. We observe the difference between CLUTO and SAS EM. Records in CLUTO are still centered in two clusters, while the results of SAS EM are distributed more evenly. Highly concentrated clusters are less useful than evenly distributed clusters because the suspicious records are only small parts in large clusters and hence business analysts have to examine a large number of records. From this point of view, SAS EM is more suitable than CLUTO in this application.

Let's take a closer look at SAS EM results. Table 3 summarizes some statistics of 100-way SAS EM clustering against the suspicious IDs.

Table 3 Subset 100-way SAS EM Clustering Results

Count	0%	<5%	5%-15%	>15%
Sus Records	0	5373	19657	63589
All Records	200033	475458	264109	300127
Sus/All	0.00%	1.13%	7.44%	21.19%
Sus/Sus Total	0.00%	6.06%	22.18%	71.76%
% in Total	16.14%	38.35%	21.30%	24.21%

This first row in Table 3 is the ratio of suspicious records to the number of records in a cluster. The subset data are separated into four groups using this ratio. 0% refers to those clusters that have no

suspicious records. >5% refers to those clusters that have less than 5% suspicious records. 5%-15% refers to those clusters that have more than 5% and less than 15% suspicious records. >15% refers to those clusters that have more than 15% suspicious records. The second row "Sus Records" refers to the number of suspicious records within each group. For instance, the number "5373" in this cell means that there are 5373 suspicious records located in those clusters that have less than 5% suspicious records. The third row "All Records" refers to the number of records within each group. For example, the number "200033" in this cell means that there are 200033 records located in those clusters that have no suspicious records. In other words, these records can be excluded from investigation. The fourth row "Sus/All" is the ratio of suspicious records to the number of total records in that group. The fifth row "Sus/SusTotal" is the ratio of suspicious records to the total number of suspicious records. The number "71.76%" indicates that 71.76% suspicious records are located in >15% group. The sixth row "% in Total" is the ratio of records within that group to the total number of records in the subset data. We see that only 24.21% records are located in clusters which have more than 15% suspicious records.

From business perspective, a meaningful clustering solution should at least provide two advantages: first, it should reduce the number of records need to investigated; second, it should recommend suspicious data to business analysts. Table 3 helps us to evaluate the quality of SAS EM clusters from these two directions.

- If we do not consider clusters that have less than 5% suspicious records, we can exclude 54.49% of the total records.
- SAS EM points out that 24.21% of the total data are located in clusters that have more than 15% suspicious records.

5 CONCLUSION AND FUTURE RESEARCH DIRECTION

This paper applies SAS EM and CLUTO to a health insurance dataset to understand the data and detect frauds. Experimental results indicate that CLUTO is faster than SAS EM while SAS EM provides more useful clusters than CLUTO.

Clustering has two typical applications. It can be used as a stand-alone tool to get insight into data distribution or as a preprocessing step for other algorithms [5]. This project utilizes clustering as a stand-along tool to understand insurance claim data and group them into clusters. After the completion of this study, we will have some labeled insurance claim records. These labeled data allows us to implement other algorithms.

Since the ultimate goal is to predict insurance claims in a reliable precision, we suggest using classification algorithms in the future. Classification is a data mining function that builds classifiers based on the training set and uses it in classifying new data. Data mining researchers proposed various classification methods. When selecting the appropriate classification methods for insurance claims fraud detection, two issues should be considered. First, the selected classification methods must be able to handle datasets with large datasets with high dimensionality efficiently and produce acceptable classification accuracies. Second, the selected methods should be able to provide understandable results for business experts. There are many performance metrics for classification algorithms. From researchers' viewpoint, an algorithm that has the highest classification accuracy may be the desirable one. However, from insurance companies' position, an algorithm that can generate understandable and practical classification rules may be superior to others.

REFERENCES

[1] The National Health Care Anti-Fraud Association, available at: http://www.nhcaa.org/ (as of June 27, 2005).

[2] Klevecz, B. (1999) The Whole EST Catalog" Scientist 12 (2): 22 Jan 18.

[3] SAS Institute Inc. (2000) Getting Started with SAS Enterprise Miner 4.1 Cary, NC: SAS Institute Inc.

[4] CLUTO (2003) available at: http://www-users.cs.umn.edu/~karypis/cluto/index.html.

[5] Han, J. W. and Kamber, M. (2001) *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publication.

[6] Zhao, Y. and Karypis, G. (2003) Clustering in Life Science, In "Functional Genomics: Methods and Protocols", M. Brownstein, A. Khodursky and D. Conniffe (editors). Humana Press, available at: http://www-users.cs.umn.edu/~karypis/publications/Papers/PDF/biocluster.pdf.

BIO of the first author

Yi Peng is a PhD student in School of Information Science & Technology, University of Nebraska at Omaha, USA. She got her Master degree in Dept of Info. Science & Quality Assurance, University of

Nebraska at Omaha and B.S. degree in Department of Management Information Systems, Sichuan University, Chengdu, China. Yi Peng's research interests are Knowledge Discovery in Database and data mining, mathematical modeling in data mining and knowledge discovery, and foundations of data mining. She published more than twenty papers in various peer-reviewed journals and conferences.