Healthcare fraud detection: A survey and a clustering model incorporating Geo-location information

Qi Liu
Rutgers University
Newark, New Jersey, United States
liuqi67@pegasus.rutgers.edu

Miklos Vasarhelyi
Rutgers University
Newark, New Jersey, United States
miklosv@andromeda.rutgers.edu

Abstract: **Health care has become a major expenditure in the US since 1980. Both the size of the health care sector and the enormous volume of money involved make it an attractive fraud target. Therefore, effective fraud detection is important for reducing the cost of health care services. In order to achieve more effective fraud detection, many researchers have attempted to develop sophisticated antifraud approaches incorporating data mining, machine learning or other methods. This introduce some preliminary knowledge of U.S. health care system and its fraudulent behaviors, analyzes the characteristics of health care data, and reviews and compares currently proposed fraud detection approaches using health care data in the literature as well as their corresponding data preprocess methods. Also a novel health care fraud detection method including geo-location information is proposed.**

Key words: healthcare, fraud detection, survey, clustering, geo-location information

I. INTRODUCTION

Health care has become a major expenditure in the US since 1980. Both the size of the health care sector and the enormous volume of money involved make it an attractive fraud target. According to the Office of Management and Budget, in 2010, about 9%, or around $47.9 billion of the US'S Medicare expenditure was lost due to fraud[1]. Therefore, effective fraud detection is important for reducing the cost of health care system.

Detecting health care fraud and abuse, however, needs intensive medical knowledge. Many health insurance systems rely on human experts to manually review insurance claims and identify suspicious ones. This results in both system development and claim reviewing being time-consuming, especially for the large national insurance programs in countries such as US.

In recent years, systems for processing electronic claims have been increasingly implemented to automatically perform audits and reviews of claims data. These systems are designed for identifying areas requiring special attention such as erroneous or incomplete data input, duplicate claims, and medically non-covered services. Although these systems may be used to detect certain types of fraud, their fraud detection capabilities are usually limited since the detection mainly relies on pre-defined simple rules specified by domain experts [1].

Therefore, in order to achieve more effective fraud detection, many researchers have attempted to develop more sophisticated antifraud approaches incorporating data mining, machine learning or other methods. Compared to existing fraud detection system, these new proposed approaches focus on more complicated tasks such as automatic learn of fraud patterns from data, specify "fraud likelihood" of each case to prioritize some suspicious cases, and identify new type of fraud which were not previously documented.

The existing proposed health care fraud detection approaches in the literature can be classified as three categories: supervised approach, such as decision tree and neural network, used when historical fraud data is available and labeled; unsupervised approach, such as clustering, used when there is no labeled historical fraud data; and hybrid approach, which combine supervised and unsupervised approaches and usually use unsupervised approaches to improve the performance of supervised approach.

[1] http://www.politifact.com/truth-o-meter/statements/2011/jan/04/darrell-issa/rep-darrell-issa-claims-government-could-save-125-/

This paper aims to identify health care fraudulent behavior, analyze the characteristics of health care data, and review and compare currently proposed fraud detection approaches using health care data as well as their corresponding data preprocess and discuss the future research directions. Specifically, this paper begins with a background knowledge introduction of US health care system and its fraud behavior. Section 3 analyzes the characteristics of health care data used or can be used in academic research. Then we review and compare the currently proposed fraud detection approaches using health care data in Section 4. In Section 5, we propose a clustering model involving geo-location information. Section 6 discusses future research directions and draws some conclusions.

II. BACKGROUND OF US HEALTH CARE SYSTEM AND ITS FRAUD BEHAVIOR

The health care system in US contains two main programs: Medicare and Medicaid services. Medicare is a social insurance program administered by the United States government, providing health insurance coverage to (1) people age 65 or older, (2) people under 65 with certain disabilities, and (3) people of all ages with End-Stage Renal Disease, i.e., permanent kidney failure requiring dialysis or a kidney transplant. Medicare program provides three types of services: hospital insurance, medical insurance and prescription drug coverage. While Medicaid is a state administered program and each state sets its own guidelines regarding eligibility and services. Medicaid is available only to certain low-income individuals and families who fit into an eligibility group that is recognized by federal and state law.

For both Medicare and Medicaid programs, there are three major parties involve in: (1) service providers, including doctors, hospitals, ambulance companies, and laboratories; (2) insurance subscribers, including patients and patients' employers; (3) 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, health care fraud behaviors can be classified as follows [2][3]:

- Service provider's fraud:

(a) Billing services that are not actually performed;

(b) Unbundling, i.e., billing each stage of a procedure as if it were a separate treatment;

(c) Upcoding, i.e., billing more costly services than the one actually performed;

(d) Perform medically unnecessary services solely for the purpose of generating insurance payments;

(e) Misrepresenting non-covered treatments as medically necessary covered treatments for the purpose of obtaining insurance payments;

(f) Falsifying patients' diagnosis and/or treatment histories to justify tests, surgeries, or other procedures that are not medically necessary.

- Insurance subscribers' fraud:

(a) Falsifying records of employment/eligibility for obtaining a lower premium rate;

(b) Filing claims for medical services which are not actually received;

(c) Using other persons' coverage or insurance card to illegally claim the insurance benefits.

- Insurance carriers' fraud:

(a) Falsifying reimbursements;

(b) Falsifying benefit/service statements.

- Conspiracy fraud: the fraud involving more than one party, i.e., a patient colludes with his physician, fabricating medical service and transition records to deceive the insurance company to whom he subscribes.

According to the above classification, we can clearly see that the fraud committed by service providers accounts for the greatest proportion of the total health care fraud among the four types of fraud. And service providers' fraud can cause great damage to the health care system [3]. Hence, it attracts large amount of research effort. In current literature, about 69% of researches have been devoted to detecting service providers' fraud, while the research efforts on the other three types of fraud are limited (31% for insurance subscribers' fraud and 0% for insurance carriers' and conspiracy fraud) [1].

III. HEALTH CARE DATA

Raw data for health care fraud detection come mostly from insurance carriers (this also partly explains why little research exists to detect insurance carriers' fraud),

including governmental health departments and private insurance companies. Major governmental health departments that have been reported in the literature include the Bureau of National Health Insurance (NHI) in Taiwan [2][4][5][6], and the Health Insurance Commission (HIC) in Australia [7][8][9][10][11]. The data from private insurance companies have also been used by several researchers [12][13].

No matter which source the raw data come from, the mostly used raw data in health care fraud detection are insurance claims. An insurance claim involves the participation of an insurance subscriber and a service provider, the layout of the data is shown in Figure 1. The claim data have two characteristics. First, they contain a rich amount of attributes to describe the behaviors of the involved service providers and insurance subscribers, allowing for detection of the types of fraud committed by these two parties. Second, each claim usually contains unique identifiers for the involved service provider and insurance subscriber, respectively. By using the unique identifiers to link different claims, it is possible to obtain a global view of a service provider's behaviors over time and across different insurance subscribers, and also a global view of an insurance subscriber's behaviors over time and across different service providers. The global views help significantly in identifying the fraud committed by service providers and by insurance subscribers.

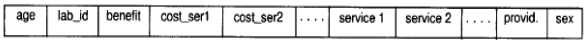

Figure 1: Layout of insurance claim data [7]

Besides the insurance claim data, the other kind of data used in health care fraud detection is general practitioners data [7]. This data is used to provide a general description of service providers in certain time period. The attributes of this data include some personal information of service providers as well as measures of their services such as the cost, usage and quality of the services. The record layout of general practitioners data is shown in Figure 2. General practitioners' data usually used with insurance claim data in supervised fraud detection methods to provide the description of the nature of the practice as well as the identification of the selection and frequency of tests.

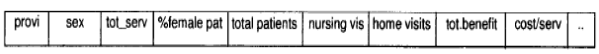

Figure 2: Layout of general practitioners data [7]

In addition, a new kind of data, called clinical-instance data, has been used in some literature using process-mining to detect health care fraud [2][5]. Typically, a clinical instance is a process instance comprising a set of activities, each of which is a logical unit of work performed by medical staffs. For example, a patient treatment flow may involve measuring blood pressure, examining respiration, and medicine treatment. These activities, may execute sequentially, concurrently, or repeatedly. For example, before giving any therapeutic intervention, diagnosis activities are usually executed to verify the condition of a patient. Also, more than one therapeutic intervention may be executed concurrently in order to increase the curative effect in some cases. By mining these activities, fraudulent behaviors can be distinguished from normal activities.

However, from the report above, we can see that the data from U.S. Medicare/Medicaid program has been rarely used in current literature of health care fraud detection, which limit the application of advanced fraud detection techniques to the U.S. health care system. Currently, there are some data from Medicare/Medicaid available online (www.cms.gov). According to this website, the primary data sources for Medicaid statistical data that can be used for fraud detection are the Medicaid Statistical Information System (MSIS) and the Medicaid Analytic eXtract (MAX) files. MSIS is the basic source of state-submitted eligibility and claims data on the Medicaid population, their characteristics, utilization, and payments. While, the Medicaid Analytic eXtract (MAX) data – formerly known as State Medicaid Research Files (SMRFs) – are a set of person-level data files derived from MSIS data on Medicaid eligibility, service utilization and payments. The data are available for all states and the District of Columbia beginning with calendar year 1999 and selected states prior to 1999. Using these data files in academic research projects can explore the appropriated fraud detection techniques for U.S. health care system and improve its fraud detection capacity.

IV. HEALTH CARE FRAUD DETECTION TECHNIQUES

As we have discussed in Section 1, the existing research approaches for health care fraud detection can be divided

into three classes: supervised methods, unsupervised methods, and hybrid methods. And the choice of these methods depends on the availability of historical labeled fraud data. In this section, we review various proposed methods in details and compare their advantages and disadvantages for health care data.

A. Supervised methods

Multilayer perceptron (MLP) neural network is a widely used supervised technique in health care fraud detection because it has many advantages such as it can handle complex data structure especially non-linear relationship and it has high tolerance to noisy data. [8] was the first paper to apply neural network to health care fraud detection. They used a MLP neural network to classify the practice profiles of general practitioners in order to reduce the inconsistencies of experts' classifications due to subjective. After each training of the MLP neural network, the probabilistic interpretation was used to filter the classified profiles. The low-probability filtered profiles were identified and then reassessed by the expert consultants. After a few iterations, many of the incorrectly classified profiles were identified and changed by this procedure. [13] also utilized a committee of MLP multilayer to detect health care fraud in Chile. They implement a committee of ten independently trained neural networks for each one of the entities involved in the fraud/abuse problem: medical claims, affiliates, medical professionals and employers. Upon the medical claim they got, the different entities are analyzed separately using historical data with cross-references among them. This divide-and-conquer strategy allows to feedback information over time, combining affiliates', doctors' and employers' behavior.

Another commonly used supervised technique in health care fraud detection is decision tree. It also has many unique advantages such as its results are easy to interpret, it can generate rules from tree, and it can handle missing values. Among various decision tree algorithms, C5.0 is the one that used mostly due to its advanced mechanisms such pruning level, which allows tuning the severity of tree pruning algorithm; adaptive boosting, which builds a sequence of classifiers and uses a voting strategy to reach the final classification, and misclassification weights, which allows defining different costs for different errors in classification. [18] used C5.0 algorithm to support the task of planning audit strategies in fraud detection. Specifically, the authors want to identify the cases, for which the audit cost can be recovered by audit fee, and assign the majority of audit efforts to this kind of cases. In order to achieve this goal, the authors build a sequence of classifiers, where later classifier is built starting from the errors of the former classifier. The decision trees are trained to distinguish between positive class of actual recovery (car), which are fruitful audits, and negative car, which are unfruitful audits. Five approaches are used to construct the classifier: (1) minimizing the false positive (FP), (2) minimizing FP with misclassification weights, (3) minimizing the false negative (FN), (4) minimizing the false positive with balanced class in conjunction with misclassification weights, and (5) combining diverse classifiers together. And each classifier is evaluated by six metrics: confusion matrix, misclassification rate, actual recovery, audit costs, profitability, and relevance. The results show that this model can provide a viable solution to health care fraud detection problem.

Besides these two extensively adopted supervised methods, many researchers combined several supervised methods in their researches. For example, [9] combined genetic algorithm and k-nearest neighbor (KNN) method in medical fraud detection. They used genetic algorithm to determine the optimal weighting of the features used to classify general practitioners' practice profiles. The weights were used in the KNN algorithm to identify the nearest neighbor practice profiles. Then majority rule and Bayesian rule were applied to determine the classifications of the practice profiles. The results indicated that genetic algorithm is very effective in finding a near optimal set of weights for the KNN classifier and with the utilizing of genetic algorithm the performance of KNN achieved good generalization. In [19] the authors proposed to use Bayesian network (BN) to detect insurance fraud. And its weights were refined by a rule generator called Suspicion Building Tool (SBT).

Moreover, [6] has compared the fraudulent service providers' identification accuracy of three supervised methods: logistic regression, neural network, and decision trees using invoices for diabetic outpatient services. The results imply that all three approaches can detect the

fraudulent and abusive medical care institutions accurately. The classification tree model performs the best with an overall correct identification rate of 99%. It is followed by the neural network (96%) and the logistic regression model (92%).

However, all the supervised methods have to deal with a problem, which is the choice of training-set and test-set. The correct size of the training set is an important parameter in a classification experiment. With the increase of the size of the training set, the complexity of the model also increases, meanwhile the training error decreases. This does not imply that large training-sets are necessarily better: a complex model, with a low training error, may behave poorly on new instances. This phenomenon is named over-fitting: the classifier is excessively specialized on the training data, and has a high misclassification rate on new data. In order to deal with this problem [13] used a technique called "early stopping" in their model. This technique uses two different datasets in training an NN: one is used to update the weights and biases, and the other is used to stop training when the network begins to over-fit the data. Also [8] added a small weight delay term to their error function to avoid this problem.

B. Unsupervised methods

Compared to supervised health care fraud detection methods, which centralized on MLP neural networks and decision trees, unsupervised health care fraud detection methods various a lot, ranging from self-organizing map, association rules, clustering, to rule-based unsupervised methods.

In [7], a self-organizing map (SOM), a type of unsupervised neural network, was applied to general practitioners database to create an unbiased subdivision of general practitioners' practices for the purpose of more effective monitoring of test ordering. The results indicated that after applying SOM, general practitioners were successfully classified into groups of various sizes, which reflect the nature and style of their practices. The author also used association rules on insurance claims database to identify commonly related test. Therefore, if one or more of the tests in a claim has little association with the other tests, this claim would be identified as suspicious fraudulent claim. Association rules have given a new perspective to health care fraud detection problem, and the output rules can be utilized to facilitate other fraud detection methods.

[12] has developed an expert system, called electronic fraud detection, to detect service providers' fraud. This system based on unsupervised rule-based algorithm to scan health insurance claims in search of likely fraud. EFD has applied rule-based methods on two levels. On the first level, EFD integrates expert knowledge (27 behavior heuristics) with statistical information assessment to induce rules to identify cases of unusual provider behavior. On the second level, these rules were validated by the set of known fraud cases. According to the validation results, fuzzy logic is used to develop new rules and improve the identification process. When operating this system, each provider's behavior was measured first. Then it was compared to that of its peers (providers with the same organizational structure and specialty, and practicing in the same geographical area); if the provider stands out from the mainstream it was recognized as the suspicious fraudulent service provider.

In [11], SmartSifter, a system based on finite mixture model, is designed for on-line unsupervised outlier detection. SmartSifter uses a probabilistic model to represent the underlying data-generating mechanism. In the probabilistic model, a histogram is used to represent the probability distribution of categorical variables; for each bin of the histogram, a finite mixture model is used to represent the probability distribution of continuous variables. When a new case is coming, SmartSifter updates the probabilistic model by employing an SDLE (Sequentially Discounting Laplace Estimation) and an SDEM (Sequentially Discounting Expectation and Maximizing) algorithm to learn the probability distributions of the categorical and continuous variables, respectively. A score is given to this new case, measuring how much the probabilistic model has changed since the last update. A high score indicates that this new case may be an outlier.

[16] proposed two unsupervised models to investigate service providers' fraud. The first model attempted to analyze service providers' fraud using geographical information. First, the author used clustering procedures to group areas of similar socio-demographic zip code regions. Then he associated each zip code region with a random

variable that can help identify fraud, and run regression analysis to group together the clusters that were formed before but are not statistically significant into one large cluster. Finally, for each group, he detected possible outliers in terms of rates of utilization or billing. The second model is implicitly based on a subjective utility model. The author assumed that the utility of a beneficiary is composed of two attributes, the distance between the beneficiary and provider address, and the expected quality of service, and that utility of the beneficiary decreases with the increase in mileage between beneficiary and provider address. So based on these assumptions, the case, in which beneficiaries travel relatively long distances to get a health care service, may be a fraud or abuse. Therefore, the second model was built based on distances that beneficiaries travel in a given day from the centroid of their zip code region to the centroid of the provider's zip code. An impractical traveled distance was defined based on this data. The claims that have respective distances that are at or greater than this distance were recognized as suspicious fraud cases.

C. Hybrid methods

A hybrid model that combines unsupervised SOM and supervised MLP neural network was proposed by [8] to classify service providers' profile. In this paper, the training data were originally divided into four classes indicating different likelihoods of fraud. When the authors only apply the MLP neural network to the data, the classification results were not satisfactory. Therefore, the SOM was employed to refine the training data. The SOM indicate that only two classes were well defined by the classification given by human experts. Hence, the two classifications were used to retrain the MLP neural network and this led to better classification results.

[10] combined a clustering tool and a decision tree induction tool to detect insurance subscribers' fraud. Their method contains three steps: (1) develop a raw and unsupervised clustering of insurance subscribers' profiles; (2) a decision tree was built for each group and then converted into a set of rules; (3) each rule was evaluated by establishing a mapping from the rule to a measure of its significance using simple summary statistics; after that the extremes could be identified for further investigation. This method can significantly reduce the rules generated by decision tree to make the results easier to interpreter.

D. Summary

Table 1 summarizes different fraud detection models that have been studied in literature and the fraudulent behaviors that they attempt to detect. According to it, we can see that except neural network and decision tree, all the other proposed methods have been applied to only one kind of fraudulent behavior. Since how to extend these methods to detect other kinds of fraudulent behaviors especially conspiracy fraud should be the research areas that deserved more attention.

Table 1: Summary of health care fraud detection methods and their corresponding tasks

Type	Methods	Fraudulent Behavior
Supervised methods	Neural Network	• Service Providers' Fraud • Insurance Subscribers' Fraud
	Decision Tree	• Service Providers' Fraud • Insurance Subscribers' Fraud
	Genetic Algorithm & KNN	Service Providers' Fraud
	Rule-based Classifier & BN	Insurance Subscribers' Fraud
Unsupervised methods	SOM	Service Providers' Fraud
	Association Rules	Insurance Subscribers' Fraud
	Rule-based Method	Service Providers' Fraud
	Finite Mixture Model	Insurance Subscribers' Fraud
	Clustering	Service Providers' Fraud
	Subjective Utility model	Insurance Subscribers' Fraud
Hybrid methods	SOM & Neural Network	Service Providers' Fraud
	Clustering & Decision Tree	Insurance Subscribers' Fraud

Based on the above review and discussion, supervised methods have attracted the most research effort. However, supervised methods can only be used when labeled fraud cases are available. While this kind of labeled data are not always easy to obtain. Moreover, because of the subjectivity, labeled data is not very accurate sometimes. Hence, considering data availability and accuracy, unsupervised methods will be more applicable. Accordingly, more research effort should be devoted to unsupervised fraud detection methods for health care data.

V. GEO-LOCATION CLUSTERING MODEL

In this section, we propose a clustering model considering geo-location information of both Medicare/Medicaid beneficiaries and providers to flag suspicious claims. According to the survey in the previous sections, only one existing model [16] considered Geo-location information in healthcare data, which is an important indicator of fraudulent behavior.

Intuitively, Medicare / Medicaid beneficiaries, who are senior, disabled or poor, prefer to choose the health service providers locating in a relatively short distance. Therefore,

if a Medicare/Medicaid beneficiary traveled a long distance for a service, this may imply a fraud. Several reasons are possible for the beneficiaries to choose long-distance service providers. For example, compare with the services provides by short-distance services providers, the quality of the services provided by long-distance service providers are better. Although it is impossible to observe the quality-of-service of each service provider, we assume that the quality-of-service of certain service provider is comparable for each individual. Another explanation for selecting long-distance service providers is that none of the service providers locating in short distance can treat the beneficiaries' diseases. The model we construct accounts for these cases.

In addition, in this model, we don't focus on one specific type of fraud; instead, the model can hope to identify any types of healthcare fraud such as service providers' fraud, insurance subscribers' fraud, and collusive fraud. For instance, if a Medicare/Medicaid beneficiary's identification is stole, the thief can use this identification in a service provider close to himself but far from the beneficiary. This is a kind of insurance subscribers' fraud. If the thief is a physician or staff in certain service provider and they bill forged services using the beneficiary's identification, it becomes to a type of providers' fraud. It is also possible that a service provider and a beneficiary make an agreement that the beneficiary travels long distance to get a kickback, and the service provider can bill unnecessary services to cover the expense and earn extra money. In this case, it is a collusive fraud.

A. Methodology

1) Research Design: Cluster analysis

Cluster analysis groups the objects on information found in the data that describes the objects and their relationships [17]. Each object is very close or similar to other objects in the same group, but different from objects in the other groups. It begins with a single group, follows by attempt to form subgroups which are different on selected variables [20]. In this study, we use cluster analysis to group Medicare claims according to claim payment amount and the distance between beneficiary and service provider. During the clustering process, abnormal claims will be differentiated from normal claims and form separated cluster due to their difference. Therefore, using cluster analysis, we are able to not only identify the claims with extreme payment amounts and distances but also recognize some potential outliers that cannot be easily detected when analyzing only claim payment amount or distance.

2) Data

The data we use in the experiments is purchased from the center for Medicare and Medicaid services (http://www.cms.gov/). It includes all the Medicare inpatient claims in 2010. There are in total 12,453,186 records and 1627 fields in the dataset. The information in the dataset include insurance subscribers' information (age, sex, medical status code…), insurance providers' (institutional) information (Provider number, providers' state..), physicians' information (claim operating physician number, claim attending physician number), diagnosis information (diagnosis code count…), payment/payer information (claim payment amount, payer code…), claim information (claim total charge amount, claim diagnosis code count, claim admission date, claim pass through per Diem amount, claim total capital amount…). Considering privacy issues, the data is anonymized; all the information that can be used to identify beneficiaries has been deleted. So the data doesn't contain the exact address information of beneficiaries, we only know beneficiaries' living counties and service providers' locating states.

3) Experiment Process

The general experiment process is shown in Figure 5. As most of the previous study, the experiments contain three major phases: preparation, data preprocessing, and analysis. Three groups of comparative trials are conducted in the third phases.

Since in the original Medicare dataset, we only know beneficiaries' living county and service providers' locating state. In order to calculate the distance between beneficiaries and services providers we need to know the latitude and longitude of each county and state. Thus, in the preparation phase, I first collect this information from the US census website. Then I map it to the Medicare dataset according to the SSA code of each county and state. Therefore, every claim in the Medicare dataset has both beneficiary's and service provider's latitude and longitude information.

In the data preprocessing phase, I first calculate the Euclidean distance between beneficiaries and service

providers using the latitude and longitude information mapped to the Medicare dataset in the data preparation stage. Considering the availability of the service and payment amount vary with disease, we want to use diagnose as the control variable in our experiments. Therefore, in this phase, we classify the claims according to their principal diagnoses. Three most common diagnoses are Pneumonia (486), Rehabilitation procedure (v5789), and Septicemia (0389). There are 442816 claims for Pneumonia (486), 352059 claims for Rehabilitation procedure (v5789) and 338149 claims for Septicemia (0389). We extract these claims to form three datasets; each dataset contain claims with the same principal diagnose. Since payment amount is the most commonly considered variable in existing healthcare fraud detection literature we include it together with distance in our clustering model.

Finally, in the analysis phase, we first separately analyze the payment amount and distance in the three datasets, respectively. Then use our model to cluster the three datasets, respectively. At last, we compare the analytics results of all the experiments.

B. Preliminary results and discussion

For the dataset containing Pneumonia related claims, the maximum payment amount is 357384.94, the minimum payment amount is -13180.52, average (mean) payment amount is 7275.18, median payment amount is 5909.48, and standard deviation is 7420.16. We can see that the distribution of payment amount in this dataset is not symmetric. Since mean is large than median, the distribution is skewed to the left, which indicate. Actually, large number of relatively small payment amounts and some extreme large payment amount make the average payment amount larger than the median payment amount.

In the same dataset, the maximum distance is 317.00995, the minimum distance is 0.02156, average distance is 37.73775, median distance is 1.75870, and standard deviation is 83.68132. Similar as payment amount, the distribution of distance is also skew to the left to make average distance much larger than median distance.

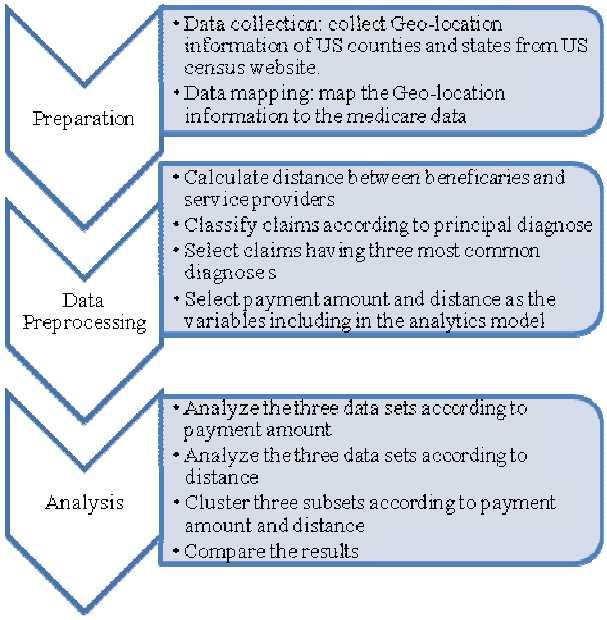

Figure 5: Experiment process

The first few steps of clustering process of this dataset are shown in Figure 6. (Note that the clustering process doesn't stop at the stage showing in the figure, each cluster will be divided until all the records in the cluster are exactly the same. Generally speaking, earlier detected outliers are more informative.) The first separated small cluster (CL2) contains 13 claims. In this cluster, average distance is 61.292367526, and average payment amount is 298240.85. The second separated small cluster (CL6) contains 43 claims. The average distance in this cluster is 83.84936501, and the average payment amount in this cluster is 188645.91. We can see that the two identified abnormal clusters contain the claims with relatively long distance and large amount of payment.

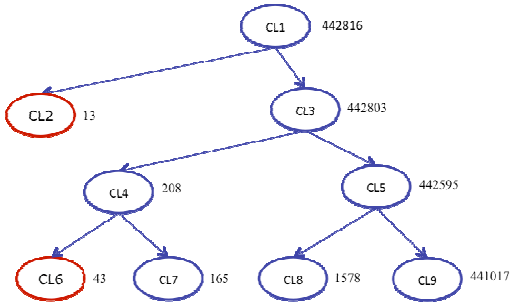

Figure 6: Clustering results of Pneumonia dataset

In the dataset involving Rehabilitation related claims, the maximum payment amount is 841690.39, the minimum payment amount is -15629.01, average payment amount is 16269.64, median payment amount is 15305.02, and standard deviation is 8491.58. Compared with the

Pneumonia dataset, the payment amount in this dataset is larger, and the payment distribution is closer to symmetric.

In the same dataset, the maximum distance is 317.00995, the minimum distance is 0.02156, average distance is 34.71061, median distance is 1.82595, and standard deviation is 79.85406. Like Pneumonia dataset, the distribution of distance in this dataset is also skew to the left to make average distance much larger than median distance.

The first few steps of clustering process of this dataset are shown in Figure 7. The first separated small cluster (CL2) contains only one claim whose distance equals to 1.10706 and payment amount equals to 841690.39, which is the largest payment amount in the Rehabilitation dataset. The second separated small cluster (CL4) also contains only one claim. The distance of this claim is 2.18915; and its payment amount is 269041.80, the second largest payment amount in the dataset. The third separated small cluster (CL6) contains 33 claims. The average distance in this cluster is 87.20961, and the average payment amount in this cluster is 140009.11. This cluster contains the claims with relatively long distance and large amount of payment.

In Septicemia dataset, the maximum payment amount is 823920.20, the minimum payment amount is -8565.37, average payment amount is 15967.17, median payment amount is 11037.08, and standard deviation is 17178.91.

For the same dataset, the maximum distance is 317.00995, the minimum distance is 0.02156, average distance is 46.52659, median distance is 1.80018, and standard deviation is 92.0566907.

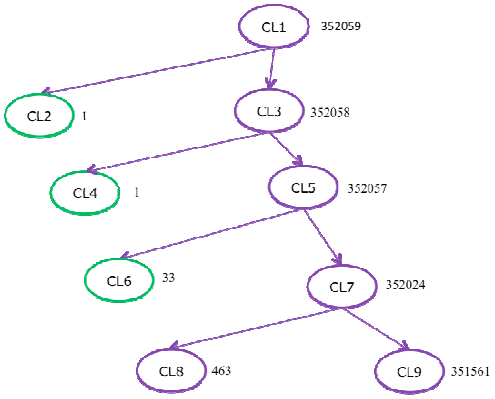

Figure 7: Clustering results of Rehabilitation dataset

The first few steps of clustering process of this dataset are shown in Figure 8. The first separated small cluster (CL2) contains 2 claims. In this cluster, average distance is 241.27532, and average payment amount is 793376.38. The second separated small cluster (CL4) contains 72 claims. The average distance in this cluster is 100.00923, and the average payment amount in this cluster is 325850.74. Similar as the Pneumonia dataset, the two identified abnormal clusters contain the claims with relatively long distance and large amount of payment.

In summary, according to the clustering results of these three datasets, our model can not only detect claims with extreme payment amount or distance but also identify some suspicious claims having relatively long distance and large payment amount. However, we do not claim that every suspicious claim detected by our model is involved in fraud. On the other hand, we argue that our model can identify possible fraudulent cases and this is useful in the preliminary analytic procedure.

In the future, we can incorporate more variables in our clustering model to achieve more accurate results. Also we may use the clustering results as the weight to build a more sophisticated prediction model to forecast fraud.

VI. CONCLUSION AND FUTURE WORK

In conclusion, this paper introduces some preliminary knowledge of U.S. health care system and its fraudulent behaviors, analyzes the characteristics of health care data, reviews and compares currently proposed fraud detection approaches using health care data in the literature as well as their corresponding data preprocess methods and finally proposed a geo-location clustering model.

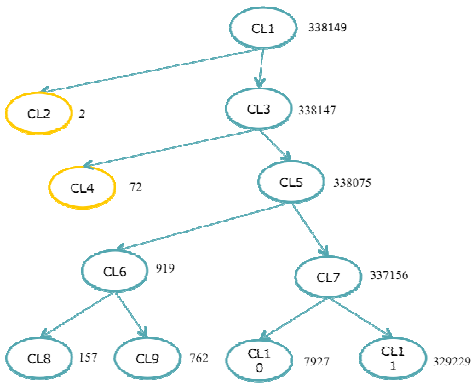

Figure 8: Clustering results of Septicemia dataset

For future research, several directions have been pointed out. First, all of the currently proposed fraud detection approaches has focused on the discrimination of fraudulent and legitimate cases, none of them has considered the causes of fraud [1]. However, to identify and eliminated the causes of fraud is the ultimate goal, so that fraud can be prevented in the future. Therefore, the application of causal model, which is applied to several domains including social science and process control, to health care fraud detection is a research area that deserves more research effort in the future. Second, because both fraudulent and legitimate patterns in health care data may change over time, health care fraud detection method have to be dynamic enough to adapt these changes. Hence, future researches can attempt to develop self-evolving fraud detection methods.

REFERENCES

[1] Jing Li. Kuel-Ying Huang, Jionghua Jin, Jianjun Shi (2007) A Survey on statistical methods for health care fraud detection, Health Care Management Science

[2] Yang WS, Hwang SY (2006) A process-mining framework for the detection of health care fraud and abuse. Expert Syst Appl 31:56–68

[3] NHCAA (2005) The Problem of Health Care Fraud: A serious and costly reality for all Americans, report of National Health Care Anti-Fraud Association (NHCAA)

[4] Chan CL, Lan CH (2001) A data mining technique combining fuzzy sets theory and Bayesian classifier—an application of auditing the health insurance fee. In Proceedings of the International Conference on Artificial Intelligence, 402–408

[5] Yang WS (2003) A Process Pattern Mining Framework for the Detection of Health Care Fraud and Abuse, Ph.D. thesis, National Sun Yat-Sen University, Taiwan

[6] Liou F.M, Tang Y.C, Chen J.Y (2008) Detecting hospital fraud and claim abuse through diabetic outpatient services. Helth Care Manage Sci 11: 353-358

[7] Viveros MS, Nearhos JP, Rothman MJ (1996) Applying data mining techniques to a health insurance information system. In Proceedings of the 22nd VLDB Conference, Mumbai, India, 286– 294

[8] He H, Wang J, Graco W, Hawkins S (1997) Application of neural networks to detection of medical fraud. Expert Syst Appl 13:329–336

[9] He H, Hawkins S, Graco W, Yao X (2000) Application of Genetic Algorithms and k-Nearest Neighbour method in real world medical fraud detection problem. Journal of Advanced Computational Intelligence and Intelligent Informatics 4(2):130–137

[10] Williams G, Huang Z (1997) Mining the knowledge mine: The Hot Spots methodology for mining large real world databases. Lect Notes Comput Sci 1342:340–348

[11] Yamanishi K, Takeuchi J, Williams G, Milne P (2004) On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. Data Mining and Knowledge Discovery 8:275–300

[12] Major JA, Riedinger DR (2002) EFD: A hybrid knowledge/ statistical-based system for the detection of fraud. The Journal of Risk and Insurance 69(3):309–324

[13] Ortega PA, Figueroa CJ, Ruz GA (2006) A medical claim fraud/ abuse detection system based on data mining: a case study in Chile. In Proceedings of International Conference on Data Mining, Las Vegas, Nevada, USA

[14] Lin J-H, Haug PJ (2006) Data preparation framework for preprocessing clinical data in data mining, AMIA Symposium Proceedings 489–493

[15] Sokol L, Garcia B, West M, Rodriguez J, Johnson K (2001) Precursory steps to mining HCFA health care claims. In Proceedings of the 34th Hawaii International Conference on System Sciences

[16] Musal R.M. (2010) Two models to investigate Medicare fraud within unsupervised databases. Expert Systems with Applications 37: 8628-8633

[17] Tan PN, Steinbach Michael, Kumar Vipin (2006) Introduction to Data Mining, Pearson Education

[18] Bonchi F, Giannotti F, Mainetto G, Pedreschi D (1999) A classification-based methodology for planning auditing strategies in fraud detection. In Proceedings of SIGKDD99, 175–184

[19] Ormerod T, Morley N, Ball L, Langley C, Spenser C (2003) Using ethnography to design a Mass Detection Tool (MDT) for the early discovery of insurance fraud. In Proceedings of the ACM CHI Conference

[20] Thiprungsri, S (2012) Cluster analysis for anomaly detection in accounting, Rutgers Doctoral Thesis