 ISSN: 2454-4248
45 – 48

A Model System to Identify Health Care Fraud Using Machine Learning Algorithm

Snehal Wandre
Computer Science & Information Technology
G H Raisoni Institute of Engineering and Technology
(Savitribai Phule Pune University)
Pune, India
Snehalwandre9@gmail.com

Akansha Deshmukh
Computer Science & Information Technology
G H Raisoni Institute of Engineering and Technology
(Savitribai Phule Pune University)
Pune, India
gunjandeshmukh59@gmail.com

Sadhana Honkande
Computer Science & Information Technology
G H Raisoni Institute of Engineering and Technology
(Savitribai Phule Pune University)
Pune, India
Sadhanass10@gmail.com

Tejaswi Patil
Computer Science & Information Technology
G H Raisoni Institute of Engineering and Technology
(Savitribai Phule Pune
Pune,India
tejaswipatil60@gmail.com

Under the Guidance of,
Mr. Suryakant Bhalke
Computer science & Information Technology, GHRIET
(Savitribai Phule Pune University)
Pune ,India
suryakant.bhalke@raisoni.net

Abstract—Fraud detection is interesting research topic and it not only needs data mining techniques but also needs a lot of inputs from domain experts. In health care claims, relationships between physicians and patients form complex communities structures and these communities could lead to potential fraud discoveries. Traditionally, researchers have focused on clustering physicians and patients and tried to find the suspicious communities. In this paper, we studied and discussed different types of relationships and focus on small but exclusive relationships that are suspicious and may indicate potential health care frauds. We developed two algorithms to detect these small and exclusive communities. These algorithms can be applied to larger dataset and are highly scalable. We tested these algorithms with a set of synthesized datasets. These synthesized datasets were created to resemble the real health care claims datasets and used to test the fraud detection algorithms. The test results show the these algorithms are very efficient and can evaluate the communities structures of 50,000 providers in about 1 minute.

*****

I. INTRODUCTION

Frauds exist wherever when it involves money transactions. Health care is especially a tempting target for thieves. In the United States, total health spending in America is a massive $2.7 trillion, or 17% of GDP. No one knows for sure how much of that is embezzled, but in 2012 Donald Berwick, a former head of the Centres for Medicare and Medicaid Services (CMS), and Andrew Hackbarth of the RAND Corporation, that fraud (and the extra rules and inspectionsrequired to fight it) added as much as $98 billion, or roughly 10%, to annual Medicare and Medicaid spending, and up to $272 billion across the entire health system [1]. There are different types of frauds in health care systems, such as drug abuses, counterfeit drugs, off-label marketing issues. In this paper, we will focus on the health insurance claims. When health services are provided, a set of claims is submitted to one or more insurers for reimbursements. Health insurance is like any other types of insurance that there is a claim processing system to adjudicate these claims to determine if a claim should be paid or by how much a claim should be paid.

To prevent the possible frauds, there are multiple levels of edits within the claim processing systems. Some edits are implemented to prevent the incorrect payments and are part of pre-payment system adjudication. Some edits are implemented after the payments have been made to the health providers and they are the post-payment edits

This paper will discuss the claims data that are processed and paid to the providers. Using post- payment claims data, we can perform many types of data analytics and data mining techniques to discover potential frauds. There are many types of insurance frauds, the following is a list of 10 types of frauds in health insurance that are most commonly mentioned:.

 ISSN: 2454-4248
45 – 48

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

1) Billing for services not rendered.
2) Billing for a non-covered service as a covered service.
3) Misrepresenting dates of service.
4) Misrepresenting locations of service.
5) Misrepresenting providers of service.
6) Waiving of deductibles and/or co-payments.
7) Incorrect reporting of diagnoses or procedures (including unbundling).
8) Overutilization of services.
9) Corruption (kickbacks and briberies).

In this paper, we developed algorithms that target at one type of frauds. That is the suspicious provider communities that either share patients between or refer patients to each other. These communities are usually small and have exclusive relationships within the communities and no outside connections. The relationships between these communities are suspicious, however we couldn't be 100% confident that these communities are conducting fraudulent activities. There are other factors we need to consider, for example, incomplete data etc. These communities can be put on a watch list for further investigations and review. The additional review will help prevent incorrect payments to go out to these groups of providers or patients.

II. LITERATURE SURVEY

Fraud and abuse have led to significant additional expense in the health care system of the United

States. This paper aims to provide a comprehensive survey of the statistical methods applied to health care fraud detection, with focuses on classifying fraudulent behaviors, identifying the major sources and characteristics of the data based on which fraud detection has been conducted, discussing the key steps in data preprocessing, as well as summarizing, categorizing, and comparing statistical fraud detection methods. Based on this survey, some discussion is provided about what has been lacking or under- addressed in the existing research, with the purpose of pinpointing some future research directions.

Community Detection is a well studied area. There are so many research that have been done in Community Detection algorithm development. In [2], it gives a real good comprehensive overview of community detection algorithms. One of the most popular algorithms is to use Modularity as the objective function to optimize the cluster assignments until it reaches to an optimal structure [3]. These algorithms are effective to parse the whole physician network into smaller communities based on their similarities. The selection of similarities is another research topic. Some algorithms use the distance between nodes as the similarity measurements. Some similarities are based on the existence of connections. In this paper, we select to use the connection-based similarities rather than the distance-based similarities. In another research we conducted, we developed an algorithm based on spectral analysis and it can parse a network Identify applicable funding agency here. If none, delete this text box into similar communities using Fielder vector.

III. TYPES OF COMMUNITIES

I am defining three types of community structures that exist within a health care dataset. As we discussed in , it introduced one type of community connected by referral relationships between primary care physicians and specialty physicians. There are two other connections to establish a community structure between physicians or between physician and patients.

The two types of relationships between physicians are:

1) If two physicians treated the same patients.
2) If one physician refers a patient to a second physician.

Physicians have connections with each other through patients they served. When a physician provides any services to a patient, this physician establishes a connection this patient. This connection is reflected on the health care claim and submitted into claims database for payment processing and analytics.

IV. COMMUNITIES BETWEEN PHYSICIANS

We would like to focus on the patients are shared between physicians, or the number of patients that are treated by two or more physicians. In order to detect the relationships between physicians, we will need to first build a matrix that can reflect the relationships. The matrix values are the number of patients shared between any pairs of physicians

A. Synthesized Data

In order to test and compare the performance of this Algorithm 2, we need to find the health claims datasets. However, the ideal health claims datasets are not easy to obtain due to privacy issues. Some public use files have limited information to test this algorithm. We chose to create a set of synthesized health claims datasets and use them to test this algorithm. The following are a few features of this synthesized dataset.

1) it generates any N number of physicians.
2) its patients distribution follows a Power Law, which is also the distribution of most health claims datasets.
3) it contains more than 5,000 procedure codes that are used in professional health claims data.

ISSN: 2454-4248
45 – 48

4) its payments to each of procedure code are the average amounts according to the public data.
5) it creates referral relationships between physicians.
6) it creates the Date of Services within one year period of time.
7) its claims frequency of each patient visit to one physician follows a Power Law.
8) it contains a few fraudulent features for further fraud algorithm testing. Some of the fraud features include impossible service days, work on holidays, exclusive referrals and impossible code pairs etc. Six test datasets were created for the performance tests. These test datasets have 100, 1K, 5K, 10K,20K and 50K physicians. The largest dataset of 50K physicians has about half an million claims. The algorithm's performance is tested on the communities of size 2,3,5,10, 20 and 100 of physicians.

B. Running Time Comparison

These tests were running using the Algorithm 2. Here is the comparison of the algorithm performance and their running time in seconds in Figure.

Real Time in Seconds		Community Size				
Dataset Names	Number of Claims	2	3	5	10	20
100	2,074	0.58	0.61	0.65	0.70	1.14
1K	9,649	0.64	0.67	0.68	0.62	0.59
5K	43,217	0.98	0.89	1.00	1.06	0.90
10K	86,364	1.74	1.78	2.38	1.86	2.73
20K	172,659	4.41	4.80	6.23	8.47	5.73
50K	427,538	24.39	23.08	28.84	1'59"	2'36"

Fig. . Running Time (in seconds) of different claims datasets and community

From this results information, this algorithm is implemented really efficiently. It can evaluate 50,000 physicians in about 1 minute to find the communities of 100 physicians. It is even quicker if we were to find the smaller size of communities, which are probably more
suspicious in terms of fraud potentials.
In general, the larger communities we were trying to evaluate, the longer it may take. However, we also noticed that the running time is reduced for some datasets when the community size increases. Theoretically, there should be more calculations and database operations when we want to find larger communities. The reason is due to the fact that when the community size we are looking for becomes bigger, there are less communities found in the claim datasets and thus reduced the running time in the following process. There might not any be any
communities we can find. There are more communities with smaller number of physicians and less communities with larger number of physicians.

C. Community Sizes Comparison

In Figure 7, it summarizes number of physicians that are identified in each of these tests. We can see that the communities of 100 providers can only be found in the larger datasets of 20K and 50K physicians. it is reasonable that in the smaller dataset, it is harder to form a large and exclusive communities.
The probability of N have not been applied yet to filter to the set of physicians that are most interested to us. Once we have a desired probability, for example, we can filter these providers to exclusive communities or highly exclusive communities.

Number of Physicians		Community Size					
Dataset Names	Number of Claims	2	3	5	10	20	100
100	2,074	24	1	0	0	0	0
1K	9,649	197	107	27	2	0	0
5K	43,217	1,010	549	249	127	23	0
10K	86,364	1,519	1,219	619	192	132	0
20K	172,659	1,600	1,581	1,485	487	211	7
50K	427,538	1,579	1,565	1,697	1,648	734	92

Fig. . Number of Physicians idenfied in different claims datasets and

V. RESULTS

section will discuss the results from this algorithm as run on the synthesized datasets and how well lgorithm performs on finding the communities that fined. We will compare this algorithm's mance on different sizes of the synthesized h claims data. This algorithm could be mented in different ways. We chose to implement lgorithm in database operations, rather than to execute loops to iterate through all physician pairs, which is probably the least efficient implementation. The algorithm is written in SAS programming language and tested on a Windows 10 64-bit operation system, with a Interl Core(TM)2 Duo CPU and 4.00 GB memory on the board.

VI. DISCUSSIONS

As we discussed previously, there are two ways to connect two physicians in a claim dataset.
1) Physicians treating the same patients
2) physicians referring to other physicians
In this community detection algorithm, I only examined the physicians relationships when they treated the same patients. When two physicians treated the same patients, they are considered having a connection. For the referral relationship when one physician refers patients to another physician, we can apply this algorithm similarly by starting building the relationship matrices discussed in Algorithm 1. Community Detection is an interesting topic in fraud detection. We tested this algorithm with health care data, but it can also be applied to other insurance data for fraud detections. Our algorithm solved the big data issue by only looking for the suspicious communities, which are those communities with exclusive relationships and contain fewer physicians. By

ISSN: 2454-4248
45 – 48

introducing a probability of P, this algorithms becomes more general and can detect more types of communities. We did not test for accuracies of the test results, because we target at 100% detection rate in this algorithm. We will find all the communities that exhibit exclusive relationships. When we don't know what the size of the fraudulent communities, we will need to run this algorithm for all possible sizes. Usually you may want to start with the small size of communities because most of the collusion frauds involve less physicians. For example, we can first examine the communities of size 1 through 10 to see if any of the small communities exist before checking larger ones.

VII. REFERENCES

[1]. "The $272 billion swindle." [Online]. Available: http://www.economist.com/news/united- states/21603078-whythieves- love-americas-health- care-system-272-billion-swindle

[2]. S. Fortunato, "Community detection in graphs," Physics reports, vol. 486, no. 3, pp. 75–174, 2010.

[3]. M. E. Newman, "Modularity and community structure in networks, "Proceedings of the national academy of sciences, vol. 103, no. 23, pp.8577–8582, 2006.

[4]. S. Chen and A. Gangopadhyay, "A novel approach to uncover health care frauds through spectral analysis," in Healthcare Informatics (ICHI), 2013 IEEE International Conference on. IEEE, 2013, pp. 499–504.

[5]. A. Gangopadhyay, S. Chen, and Y. Yesha, "Detecting healthcare fraud