Application of Data Mining Technique for Fraud Detection in Health Insurance Scheme Using Knee-Point K-Means Algorithm

Fashoto Stephen G., Owolabi Olumide, Sadiku J. and Gbadeyan Jacob A.

Redeemer’s University Mowe Ogun State Nigeria
University of Abuja Nigeria
University of Ilorin Nigeria

Abstract: Healthcare delivery is one of the most important functions of government. This task is, however, beset with so many difficulties in a developing country like Nigeria. The application of Information Technology (IT) is one way to overcome several of the problems besetting the sector. This work, therefore, focuses on the application of some computer-based techniques that could help to properly target investment in the sector and also drastically reduce fraud in health insurance. The Nigerian Health Insurance Scheme (NHIS) was introduced by the Nigerian government to make healthcare affordable to all citizens, irrespective of economic situation or occupation. This scheme is, however, known to be beset by fraudulent claims from health practitioners within the system. To reduce or possibly eliminate this fraud, we also applied Knee-point K-means Clustering method, which is capable of detecting fraudulent claims from Health providers. Cluster-based outliers were examined. Health providers claims submitted to Health Maintenance Organization (HMO) were grouped into clusters. Claims with similar characteristics were grouped together. Clusters with small populations were flagged for further investigations. For clustering using two (2) attributes, six (6) clusters are formed. 74% of claims are clustered into cluster 2, 16% are in cluster 1, 2% are in cluster 3 and 4, 6% are in cluster 5 and 0% are in cluster 0 which have a membership of less than 1%. It can be difficult to get your clustering model correctly without determining the value of k clusters first; we were able to carve out some interesting information from the results on health insurance claims. The results from the data collected from an HMO in Lagos Nigeria show that the total number of claims identified as possible anomalies from cluster-based outliers is 7 in Nigeria health insurance using probability of 0.6 as the cutoff point.

Key words: Data Mining, Fraud, Health Insurance Scheme and K-Means Algorithm

INTRODUCTION

According to (Offen, 1999) healthcare fraud as defined by National Health Care Anti-fraud Association (NHCAA) is a form of misrepresentation or intentional deception by an individual or entity knowing full-well that misrepresentation could lead to an unauthorized benefit of the individual or entity involved or some other party. In the light of this unauthorized gain, the financial fraud has emerged as the common phenomenon of this age. This monster has soiled the name of personalities of repute, well-renowned companies and countries like Nigeria, Philippines, and Russia to mention a few in international communities (Wymoo 2012).

Fraud includes pretentious claim, highly crafted or drawn-up irregularities with intention to mislead. For example in Nigeria, the rate of fraud is increasing yearly with multi-million naira financial benefit. This trend has found its way into the insurance sector of the economy of which health insurance is vital. The healthcare environment seems attractive to fraudsters because of the vast information available but relatively inadequate knowledge.

Though there is enormous data available within the healthcare system, yet these vast resources are put to limited use especially in the area of fraud detection. A possible means of fraud is when a healthcare practitioner utilizes their recorded cases to exploit the large historical database by submitting false bills to their health insurance companies. Being aware of this tendency in healthcare system and the inherent human nature to cheat, it is expedient to develop a measure that detects fraudsters as quickly as possible. Although this may be considered a bit of pessimism, it is a realistic approach. In addition, entrusting the task of fraud detection solely to technology may not yield a satisfactory or lasting result.

Health Insurance Scheme:

Health insurance scheme as a way of financing health care was first introduced in Germany in 1887, followed by Austria 1897, Norway 1902 and United Kingdom 1910. By 1930, Health insurance scheme had been well established and recognized in all European countries (Okezie, 2001).
According to Minnesota Department of Health state that Health insurance is a medical coverage purchased by individual or group from an insurance company through the payment of premium (a regular price paid in advance for the service).

With over 150 million people by the 2010 population estimation, Nigerian government policy makers, and professionals has no option than to suggest for reform in the health sector. The reform has the following objectives as stated in the goals of the National Health Insurance Scheme, which was backed by the law of Nigeria with a decree enacted in 1999, known as NHIS, decree of 1999. This decree is the main instrument that guides the operation and implementation of the reform, and as well to ensure that every one has access to good healthcare services in the country.

- Ensure social protection of the families from the huge financial bills of medical care
- To curtailed the rise in cost of health care in the country
- To ensure quality in health care services delivery
- To promote efficient healthcare services
- To improve on private participation in healthcare services provision
- To ensure equitable distribution of healthcare facilities
- To ensure the availability of funds to the health sector (NHIS, 2006).

In private health insurance scheme, the individual or group taking up the policy is ready to pay the premium that is set by the insurance company. The task undertaken by the insurance company in private health insurance scheme is to group people on the same policy and ensure that health expenses are covered in case of risks. In addition to insuring the policy-taker, insurance company also set the prices for the premium to be paid which sometime provides benefit or profit to the insurance provider’s institution.

Employer-based health insurance on the other hand is when a company ensures the health insurance coverage for its employees through a self-managed facility by paying a lump sum into employee’s healthcare services. This is a usual practice for employers in both public and private sector. The employer-based health insurance policy sometimes covers health expenditure incurred for outpatient treatment and hospitalization of employees. An example of this policy is National Health Insurance Scheme instituted in Nigeria in 2005.

Community-based health insurance scheme is offered by non-governmental organizations (NGOs). It is such a policy whereby members pay a regular amount in advance (this amount is usually a flat-rate) for specified services. The sole purpose is not for profit-making but to increase members’ accessibility to the services (Dogo, 2012).

The issue of fraud in the health insurance sector has become a major concern because of the stealthy way it is perpetrated as well as the fail-safe nature of fraud detection and prevention system available. In order to figure out the possibilities of fraud, auditors normally use past experiences with intuition to create profiles that describe fraudulently filled claim or bill. Unfortunately, these unscientific methods can miss opportunities to detect fraudulent claim while the auditors merely review the legitimate ones.

Healthcare Fraud:

As a precautionary measure, the plan of action employed by an audit system should look further for all possible means of fraud within its specified scope of audit. A comprehensive understanding of means by which the audit systems have been circumvented in the past will provide adequate skill to deter future occurrence. Some of the means (fraudulent schemes) employed to defraud health One common example of fraudulent schemes in health insurance is Phantom Billing (Liberman and Rolle 1998). This is when an unscrupulous healthcare provider charges for services that are not provided to patients. Phantom billing could take any of these two forms. Firstly, healthcare practitioners sometimes exploit the ignorance of patients due to their lack of awareness of the services they are billed for. The healthcare practitioners usually accomplish this with the use of codes. This type of fraud usually goes unnoticed because there is no communication link between the patients and the health Insurance provider that does the payment for the bills. The second way is when healthcare practitioners place code for ghost patients (non-existing patients). The report from [Business and Health, 2000] shows that, Phantom billing accounts for approximately 34 percent of healthcare fraud.

Bogus billing is another form of fraudulent scheme in health insurance. This is falsification or alteration of billing code to cover services which were not covered for in the insurance policy (Liberman and Rolle 1998 ). This type of fraud is perpetrated whenever new drugs or procedures are yet to be covered for under the implemented insurance policy adopted in a healthcare system. The common practice by some practitioners is that these new drugs or procedures will be billed just like the similar drugs or procedures that are covered for in the currently implemented health insurance policy.

According to (Liberman and Rolle, 1998), unnecessary billing or unneeded service billing is the common fraud that happens in ambulatory service facilities. Some physicians recommend services to perfectly healthy patients in order to generate bills for insurance claim. This type of fraud is hard to point out because physicians
usually claim that they were just observing safety measure. Unnecessary services billing accounts for 18 percent of healthcare fraud (Datawatch, 2000).

Other forms of healthcare fraudulent scheme are double billing and pharmacy fraud (Liberman and Rolle, 1998). Double billing is when multiple bills are sent to a health insurance policy provider for claim or when the same bills are sent to more than one health insurance policy providers for claim. An example of pharmacy fraud is a case whereby a generic drug is dispensed and insurance provider is being charged for expensive branded product.

**Data Mining:**

According to Kantardzic (2011) Data mining is an iterative process within which progress is defined by discovery, through either automatic or manual methods. Data mining is most useful in an exploratory analysis scenario in which there are no predetermined notions about what will constitute an "interesting" outcome. Data mining is the search for new, valuable, and nontrivial information in large volumes of data. It is a cooperative effort of humans and computers. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers.

In practice, the two primary goals of data mining tend to be *prediction* and *description*. *Prediction* involves using some variables or fields in the data set to predict unknown or future values of other variables of interest. *Description*, on the other hand, focuses on finding patterns describing the data that can be interpreted by humans. Therefore, it is possible to put data-mining activities into one of two categories:

1) Predictive data mining, which *produces the model* of the system described by the given data set, or
2) Descriptive data mining, which *produces new, nontrivial information* based on the available data set.

On the predictive end of the spectrum, the goal of data mining is to produce a model, expressed as an executable code, which can be used to perform classification, prediction, estimation, or other similar tasks. On the other, descriptive, end of the spectrum, the goal is to gain an understanding of the analyzed system by uncovering patterns and relationships in large data sets. The relative importance of prediction and description for particular data-mining applications can vary considerably.

The goals of prediction and description are achieved by using data-mining techniques, and the following are the primary data-mining tasks:

1. **Classification** - discovery of a predictive learning function that classifies a data item into one of several predefined classes.
2. **Regression** - discovery of a predictive learning function, which maps a data item to a real-value prediction variable.
3. **Clustering** - a common descriptive task in which one seeks to identify a finite set of categories or clusters to describe the data.
4. **Summarization** - an additional descriptive task that involves methods for finding a compact description for a set (or subset) of data.
5. **Dependency Modeling** - finding a local model that describes significant dependencies between variables or between the values of a feature in a data set or in a part of a data set.
6. **Change and Deviation Detection** - discovering the most significant changes in the data set.

**K-Means Algorithm:**

The k-means algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intra-cluster similarity is high whereas the inter-cluster similarity is low. Cluster similarity is measured in regard to the mean value of the objects in d cluster, which can be viewed as the cluster’s "centre of gravity".

The algorithm proceeds as follows: First it randomly selects k of the objects which each represents a cluster mean or center. For each of the remaining objects, an object ia assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean. It then computes the new mean for each cluster. This process iterates until the criterion function converges. Typically, the squared-error criterion is used, defined as

\[ E = \sum_{i=1}^{K} \sum_{x \in C_i} \| x - m_i \|^2 \]

Where \( x \) is the point in space representing the given object, and \( m_i \) is the mean of cluster \( C_i \) (both \( x \) and \( m_i \) are multidimensional). This criterion tries to make the resulting \( k \) as compact and as separate as possible.

The k-means algorithm for partitioning based on the mean value of the objects in the cluster is given below.

**Function k-means**

**Input:** The number of clusters k, and the database containing the n objects (claims)

**Output:** A set of k clusters which minimizes the squared-error detection

**Method:** The k-means algorithm is implemented as follows.

Arbitrary choose k objects as the initial cluster centers;
Repeat
(re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster.
Update the cluster means, i.e., calculate the mean value of the objects for each cluster.
Until no change

The new approach is knee-point k-means clustering Algorithm:
Input : S(instance set)
Output: clusters
1: initialize k,p,x,sse
2: while k=2 to 15
3: \(X_k = \text{kmeans}(s,k)\)
4: End while
5: \(P = \text{kneepoint}(k_{2:15},X_k)\)
6: \(\text{sse} = \text{kmeans}(s,p)\)

Results:
Below is the result obtained from WEKA 3.6.0:

==== Run information ====
Scheme: weka.clusterers.SimpleKMeans -N 6 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation: Test2-weka.filters.unsupervised.attribute.Remove-R5-weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0-weka.filters.unsupervised.attribute.Remove-R3-4-weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0
Instances: 2477
Attributes: 2
Total_Amount_Billed
Total_Amount_Approved
Test mode: evaluate on training data
==== Clustering model (full training set) ====
kMeans

Number of iterations: 49
Within cluster sum of squared errors: 2.474946333312165
Missing values globally replaced with mean/mode
Cluster centroids:
Cluster#
Attribute                Full Data          0          1               2             3           4            5
(2477)        (7)      (404)     (1838)       (45)       (46)      (137)
============================================================================

Total_Amount_Billed 0.0338 0.8361 0.0414 0.009 0.2356 0.427 0.104
Total_Amount_Approved 0.0275 0.5462 0.0308 0.0073 0.1949 0.3763 0.0894
Time taken to build model (full training data) : 0.62 seconds

Discussion

Jain (2010) specified that the partitional clustering technique, k-means, is the most computationally simple and efficient clustering method. Wu et al., (2008) have shown that the k-means was one of the top ten algorithms in data mining. Although the k-means method has a number of advantages over other data clustering methods, the specification of number of clusters is a priori, which is usually unknown.

The sample data contained 2,474 enrollee claims with 8 attributes from an HMO in Lagos, Nigeria. The eight attributes used were enrollee ID, Hospital Name, Diagnosis, Class of treatment, Total Amount Billed, Total Amount Approved, Month Group and Year Group. The data was used to classify the output into fraudulent and Non-fraudulent case based on two attributes: the Total Amount billed and Total Amount approved. These attributes were normalized for comparison in order to reduce the impact of scale differences or biases. Using these two attributes, six clusters were formed. The claims were grouped into the following clusters: cluster 0 - 0%, cluster 1 - 16%, cluster 2 - 74%, cluster 3 – 2%, cluster 4 - 2%, and cluster 5 - 6%. The number of iterations was 49. The number of cluster selected was 6. Within clusters sum of squared errors was 2.474946333312165. The total number of claims identified as possible anomalies from cluster-based outliers was 7.
This scheme is, however, known to be beset by fraudulent claims from health practitioners within the system. To reduce or possibly eliminate this fraud, we applied Knee-point K-means Clustering method, which is capable of detecting fraudulent claims from Health providers.

Cluster-based outliers were examined. Health providers claims submitted to Health Maintenance Organization (HMO) were grouped into clusters. Claims with similar characteristics were grouped together. Clusters with small populations were flagged for further investigations.

**Conclusion:**

Since it is not easy to identify suspicious claims, clustering technique as an unsupervised learning algorithm is a good method for fraud detection in Health insurance scheme.

Cluster-based outliers were examined. Health providers claims submitted to Health Maintenance Organization (HMO) were grouped into clusters. Claims with similar characteristics were grouped together. Clusters with small populations were flagged for further investigations.

Clustering differs from classification and regression by not producing a single output variable, which leads to easy conclusions, but instead requires that you observe the output and attempt to draw your own conclusions. From the result above, we conclude that the model produced six clusters, but it was up to us to interpret the data within the clusters and draw conclusions from this information. In this respect, it can be difficult to get your clustering model correctly without determining the value of k clusters first; we were able to carve out some interesting information from the results on health insurance claims.

The results from the data collected from an HMO in Lagos Nigeria show that the total number of claims identified as possible anomalies from cluster-based outliers is 7 in Nigeria health insurance using probability of 0.6 as the cutoff point.

**REFERENCES**

Dogo Muhammad, 2012. Expanding health insurance coverage in Nigeria,
