

APPLICATION OF DATA MINING TOOLS FOR FRAUD DETECTION IN HEALTH INSURANCE SCHEMES

Fashoto Stephen Gbenga,Okeyinka E.A and Jayeoba Ayobami

Redeemer's University Redemption camp, Mowe Ogun state Nigeria

Department of Mathematical Sciences

gbengafash@yahoo.com

ABSTRACT

Getting good health care is an expensive but necessary service, so people rely on government-managed health insurance systems, private health insurance systems, or both to share the expensive healthcare costs. With such an intensive need for health insurances, health care service providers have found means of exploiting health insurance schemes by their fraudulent and abusive behavior.

Nigeria was recently placed in the top three position of scam prevalent countries in the world. With this statistic it is impossible to be absolutely certain about the legitimacy of and intention behind an application or claim made to Health Insurance Schemes. The best practice therefore is to tease out possible evidences of fraud from the available datasets using Data Mining techniques like the clustering algorithm. This paper has attempted to create outliers by analyzing health insurance claims using the clustering algorithms. Clusters were generated using the K-MEANS algorithm; the clusters were further analyzed resulting in some selected numbers of fraudulent patterns. Our analyses results in a good division of the sample dataset into regular cases and cases interesting to investigate further. It is hoped that this study will serve as a formal documentation for system analysts, system designers, programmers, researchers, Data Miners etc for detection of insurance fraud using data mining techniques.

Keywords : Health insurance, Data mining, Health insurance scheme and Clustering

1. INTRODUCTION

1.1 Overview

Financial fraud has become a common phenomenon in this day and age; this monster has

smeared the great name of Nigeria in the international community, collapsed many companies, and has weakened numerous great economies. Fraud includes an array of irregularities and illegal acts characterized by

intentional deception.

Fraud is a multi-million Naira business and it is increasing every year, this trend is rearing its ugly head in the insurance sector of the economy, especially the health insurance. The healthcare environment is generally perceived as being 'information rich' yet

'Knowledge poor'. There is a wealth of data available within the healthcare systems however these vast resources are not utilized efficiently especially for fraud detection in the health care sector, there have been recorded cases of Doctors claiming a larger historical base of data and thereby submitting false bills to their insurance company.

Fraud is an elusive enemy in that it is a human nature issue. Therefore, as quickly we can build a person-trap to catch a particular fraudster, just as quickly will the person find a new means to extract funds from a company. While this may be a pessimistic attitude, I consider it more of a realistic situation assessment. Also, by admitting that technology alone will not stop fraud, it brings us down the road to better detecting it. Technology needs to be paired with the human mind to make a new form, a bionic one, that stands the best chance to detecting fraud in organizations.

It is well known that in an Information Technology (IT) driven society, knowledge is one of the most significant assets of any organization, despite the pool of data available in the health care sector,

there is a lack of effective analysis tools to discover hidden relationships and trends in data. The use of systems which are able to build user profiles and then use this knowledge to extract patterns of fraud is better for adapting patterns of fraudulent activities, this system are generally built using Data Mining techniques.

1.2 Background of study

Health insurance in a narrow sense would be 'an individual or group purchasing health care coverage in advance by paying a fee called *premium*'. In its broader sense, it would be any arrangement that helps to defer, delay, reduce or altogether avoid payment for health care incurred by individuals and households. Given the appropriateness of this definition in the Nigerian context, this is the definition, we would adopt. The health insurance market in Nigeria is very limited covering less than 20% of the total population. The existing schemes can be categorized as:

- 1) **Voluntary health insurance schemes or private-for-profit schemes:** In private insurance, buyers are willing to pay premium to an insurance company that pools people with similar risks and insures them for health expenses. The key distinction is that the premiums are set at a level, which provides a profit to third party and provider institutions.
- 2) **Employer-based schemes:** Employers in both the public and private sector offers

employer-based insurance schemes through their own employer-managed facilities by way of lump sum payments, reimbursement of employee's health expenditure for outpatient care and hospitalization, fixed medical allowance, monthly or annual irrespective of actual expenses, or covering them under the group health insurance policy.

- 3) **Insurance offered by NGOs / community based health insurance:** Community-based funds refer to schemes where members prepay a set amount each year for specified services. The premia are usually flat rate (not income-related) and therefore not progressive. Making profit is not the purpose of these funds, but rather improving access to services.
- 4) **Mandatory health insurance schemes or government run schemes (e.g. NHIS):** Since 2005, all employees of the Federal Government of Nigeria and their dependencies; some autonomous and semi-government organizations and some individuals that voluntarily subscribed to the scheme are covered under National Health Insurance Scheme (NHIS).

The term 'fraud' here refers to the abuse of a profit organization's system without necessarily leading to direct legal consequences.

In the health insurance sector, fraud has become a

business critical problem because of its prevalence and the fail-safe nature of the fraud detection and prevention procedures available. Typically, auditors use past experience and intuition to create profiles that describe a fraudulently filed claim or bill. But these unscientific methods can miss opportunities to catch fraudulent claims while auditors unknowingly review legitimate ones.

However, data mining techniques have been proven to improve detection of fraud and abuse. Data mining combines powerful analytical techniques with knowledge to turn the data already acquired into the information and insight needed to identify probable instances of fraud.

1.3 Statement of Problem

Fraud is a multi-million Naira business and the rate at which it's increasing is startling. With most companies especially health insurance companies facing liquidation threats due to the present global economic recession, these companies can no longer put up with wastage, fraud and abuse. Unfortunately over one third of all frauds are detected by 'chance' means with the second best detection method as internal control, both detection methods have been proven to be insufficient in detecting health insurance fraud.

Deceptive health care providers, doctors and even recipients/enrollees use intricate, diverse and elaborate fraud schemes to milk the health care

sector of its derisory funds, various health insurance schemes armed with limited resources to process millions of claims, find it very difficult to wage war against this fraud and abuse. Fraud is detected in many ways, or at least one tries to detect it in many ways, traditionally, a company relies most on its internal control activities and the internal auditor to prevent and detect fraud. If this isn't sufficient, external audit, risk management systems, new technologies or other measures are installed, corresponding to the need the company experiences, but it is rather unfortunate that most of these solutions are not efficient. Therefore, this project would attempt to provide a relatively efficient solution/system for the national health insurance scheme (NHIS) to identify fraudulent claims using clustering algorithm.

Grossman *et al.* (1999) have indicated that fraud detection is one of the areas where data mining is considered a successful tool.

1.4 Aim and Objectives

Aim: The aim of this project is to develop a system/solution that would attempt to maximally detect health insurance fraud by generating clusters and consequently analyzing the clusters generated from the dataset for anomaly.

Objectives: The solution is developed in order to:

- Detect fraudulent activities before it's perpetrated

- Discover patterns and information relatively unknown.
- Make the information and pattern discovered available to expert systems(if any) for further analysis
- Uncover abnormal relationships between claims, care, providers and enrollee data.
- Compare data with known fraud incidents to uncover potential risk indicators for future claims and enrollments.

1.5 Research Methodology

The following steps would be followed for the research

- 1) Data collection and gathering from authentic sources.
- 2) Data analysis.
- 3) Data cleaning
- 4) Data transforming
- 5) Construction/Extraction
- 6) Data structuring and organization
- 7) Data Mining
- 8) Analysis of Result

1.6 Limitations

While data mining products can be very powerful tools, they are not self sufficient applications. To be successful, data mining requires skilled technical and analytical specialists who can structure the analysis and interpret the output

that is created. Consequently, the limitations of data mining are primarily data or personnel related, rather than technology-related.

Data mining techniques requires huge amount of data and selection of appropriate variables. The input data has to be cleaned and transferred and hence limiting the use of online data mining. Also patterns derived from data mining in the past may not continued to be the same in future. Certain types of frauds are difficult to be detected by data mining tools.

A greater limitation is the availability of data in a usable format. Data keeping in most Nigerian health insurance companies is done using files in cabinet and for the few using the electronic data; they are largely not Relational Database Management System (RDBMS) friendly.

2 LITERATURE REVIEW

2.1 Health Insurance

Insurance is a veritable tool for healthcare financing that comes in different models. It has been used by most advanced countries in its various forms to fund healthcare. Insurance in one form or the other remains a genuine and sustainable tool for financing the hard ware, delivery structure and systems of healthcare in Nigeria. It is only recently being applied by poorer developing nations to address the glaring problem of inadequate healthcare provision, which was hitherto financed exclusively from public taxation.

Insurance is a risk transfer mechanism wherein the proposer (the insured), agrees to make small periodic payments called premium to another person (the insurer) in return for the payment of a larger sum (benefit) on the occurrence of a specified event.

The basis for insurance is to protect the individual from the financial consequences of events with a low probability of happening but with the potential to cause substantial loss.

Health Insurance is a social device for pooling the health risks and costs of an exposure unit with view towards predictability. However in Health Insurance, the probability of illness is not low hence the actuarial determinations allow for more variance.

In the context of Health Insurance, the Premium is the amount charged by the insurance company with the promise to pay for any eventual “covered medical treatment” for the designated coverage.

Consequently, health insurance makes it possible to substitute a small but certain cost (premium) for a large but uncertain loss (claim), a claim is a bill for health care services that your health care provider turns in to the insurance company for payment, under an arrangement in which the healthy majority compensate for the risks and costs of the unfortunate ill minority. Pooling of health risk is a fixture of every society and takes many forms. It was even practiced in our traditional societies. The overall contributions are placed into a pool of

funds from which payment is made. The prominent health insurance in Nigeria is the ones provided by National Health Insurance Scheme, which would be the case study for fraud detection.

The parties involved in a Health Insurance Scheme include:

The Regulator: They provide oversight functions to the organs that will be involved in direct delivery of services to members i.e. HMOs and Providers.

The Health Maintenance Organization (HMO): This is an organization that offers pre-paid, comprehensive health care coverage for doctors' and hospital services. The financial burden of risks of over-using health services are borne by the HMO, its service providers or a combination.

The Providers: They are the health care providers. They are the mix of public and private facilities in the spirit of Public-Private partnership.

The Payers: They pay for the services rendered by the providers to the HMOs

The Users/Enrollees: They are the ones who receive the healthcare

2.1.1 National Health Insurance Scheme

A national health insurance scheme was first proposed in Nigeria in 1962 under a bill that was introduced to Parliament in the same year by then Federal Minister of Health, Dr. M. A. Majekodunmi. The scheme was to commence in Lagos area and provide health services through salaried doctors. The bill was opposed at the time by the Nigeria

Medical Association whose membership was influenced by Private medical practitioners in Lagos.

The NHIS idea was resurrected again in 1988 by another Minister, Prof. Ransome-Kuti. This effort resulted in the Eronini Report (1989) on the NHIS which formed the template of the present day scheme. The scheme had been bedeviled with lack of political will by the successive governments and inter-professional rivalry within the ranks of stakeholders.

However concrete steps were taken with the passage of Decree 35 (National Health Insurance Scheme) of 1999 by the government of then General Sanni Abacha and the first launch of the scheme took place. This was followed by a period characterized by administrative fumbling and pilot schemes that were ill advised and not backed by legislation.

However, the private sector wherein most of the activity in health takes place took the bull by the horn and launched private health insurance schemes in 1998. There are now over 30 HMOs providing private health insurance schemes in Nigeria. The pressure from the private sector and other stakeholders along with the enthusiasm of the then Federal Minister of Health, Prof. Eytayo Lambo led to the present new-launch of the formal public sector program of the NHIS in June 2005.

The major roles of the NHIS include:

1. Registration of HMOs and Providers

2. Setting standards for HMOs and Health care providers.
3. Ensure compliance with standards
4. Mobilization and sensitization of all stakeholders
5. Defining the minimum benefit packages.
6. Actuarial determination of premium.
7. Drawing up contracts between stakeholders
8. Training
9. Monitoring and evaluation

2.2 Health Insurance Fraud and its Detection.

The Concise Oxford Dictionary defines fraud as 'criminal deception; the use of false representations to gain an unjust advantage'. Fraud is as old as humanity itself, and can take an unlimited variety of different forms. However, in recent years, the development of new technologies (which have made it easier for us to communicate and helped increase our spending power) has also provided yet further ways in which criminals may commit fraud. Traditional forms of fraudulent behavior such as money laundering have become easier to perpetrate and have been

joined by new kinds of fraud such as mobile telecommunications fraud, computer intrusion and health insurance fraud.

Fraud detection is a continuously evolving discipline. Whenever it becomes known that one detection method is in place, criminals will adapt their strategies and try others. Of course, new criminals are also constantly entering the field. Many of these will not be aware of the fraud detection methods which have been successful in the past, and will adopt strategies which lead to identifiable frauds. This means that the earlier detection tools need to be applied as well as the latest developments. The development of new fraud detection methods is made more difficult by the fact that the exchange of ideas in fraud detection is severely limited. It does not make sense to describe fraud detection techniques in great detail in the public domain, as this gives criminals the information that they require in order to evade detection. Data sets are not made available and results are often censored, making them difficult to assess. Overall, data mining techniques has stood the test of time in the detection of fraud.

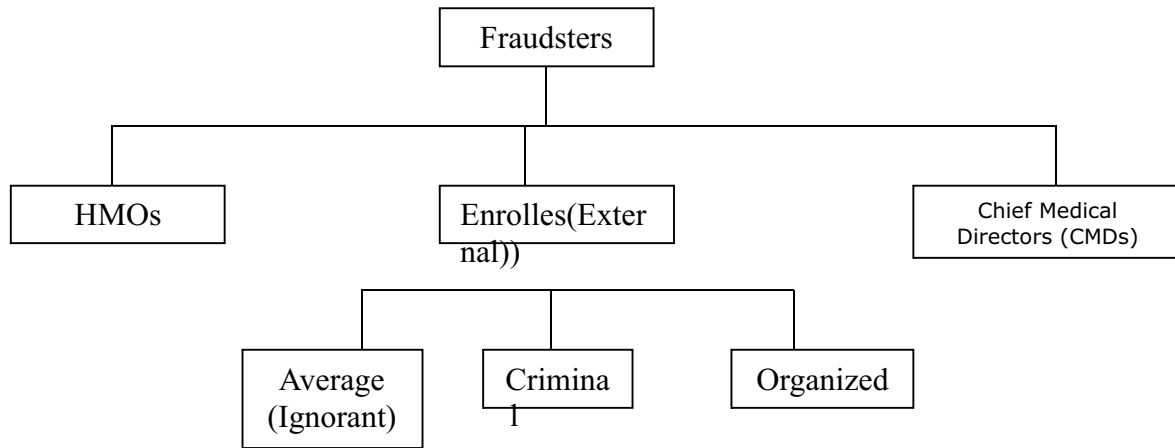


Figure 2.1: Hierarchy chart of perpetrators of fraud from the provider, Management and Community level perspectives.

1.1 Data Mining

Data mining involves the use of data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. Data mining is an essential process where intelligent methods are applied in order to extract data patterns with an objective to mine large amount of data to discover previously unknown, action oriented, hidden trends, patterns and complex relationships.

Clustering: This is an unsupervised classification technique that seeks to describe dataset in terms of natural clusters of cases.

Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

Genetic algorithms: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

Decision trees: Tree-shaped structures that

represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) . CART and CHAID are decision tree techniques used for classification of a dataset.

Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where $k \geq 1$). Sometimes called the k -nearest neighbor technique.

Rule induction: The extraction of useful if-then rules from data based on statistical significance.

Data visualization: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

The diagram below shows the architecture of a typical data mining system.

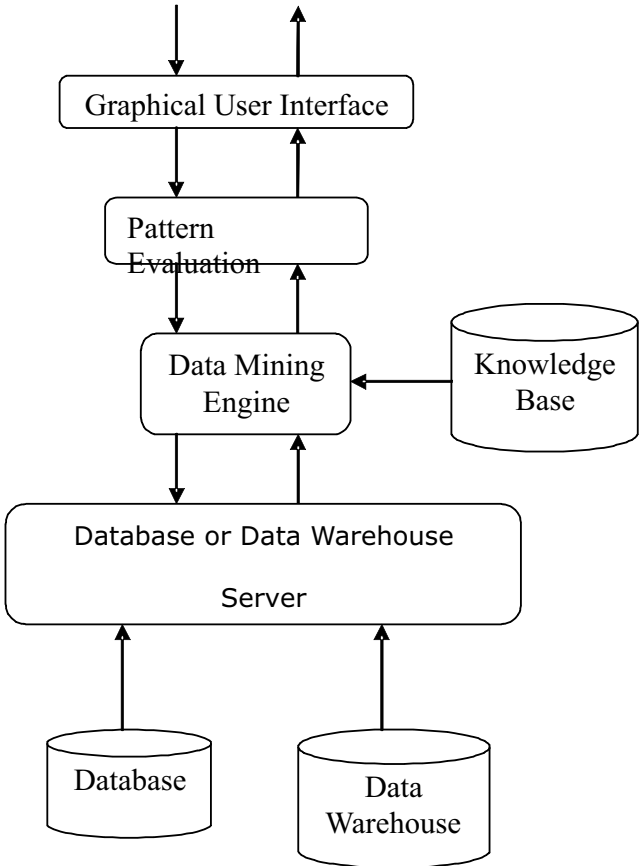


Fig 2.2: Architecture of a typical data mining system

1.1 Existing Data Mining Solutions for Detecting Fraud in Health Insurance

Data mining tools for fraud detection have crowded the health insurance market but due to changes in user specifications, these solutions become obsolete and there is a need for another optimal solution to cater for the new specification. It is estimated that Data mining algorithms change almost every ten years, better ways of solving the same problem evolve every decade.

Broadly Data Mining has two main techniques for health insurance fraud detection;

The first been straight forward statistical

techniques: Weisberg and Derrig (1993) have used linear regression models to analyze the statistical significance of fraud indicators in injury sustained on the body in Massachusetts, some other investigators developed a model and an expert system that can detect potentially fraudulent claims based on a probit model using a number of fraud indicators.

The second main data mining technique used in insurance fraud detection is clustering algorithms. Derrig and Oztaszewski (1995) used fuzzy set theories to classify and identify characteristics of fraudulent claims. They found that fuzzy techniques could efficiently classify claims based on the suspected fraudulent content. Brockett, Xia, and Derrig (1998), on the other hand, apply neural networks and back-propagation algorithms to cluster claims based on the degree of fraud suspicion. They demonstrated that this data mining technique performs better than the combination of an insurance adjuster's fraud assessment and an insurance investigator's fraud assessment with respect to consistency and reliability.

The use of data mining for fraud detection is investigated in several studies. An intensively explored method is neural networks. The studies of Barson et al. (1996), Fanning and Cogger (1998) and Green and Choi (1997) all use neural network technology for detecting respectively fraud in mobile phone networks (Barson et al.) and

financial statement fraud. Lin et al. (2003) apply a fuzzy neural net, also in the domain of fraudulent financial reporting. Both Brause et al. (1999) and Estevez et al. (2006) use a combination of neural nets and rules. The latter used fuzzy rules, where the former use traditional association rules. Also He et al. (1997) apply neural networks in the supervised component of their study. A Bayesian learning neural network is implemented for credit card fraud detection by Maes et al. (2002) (aside to an artificial neural network), for telecommunications fraud by Ezawa and Norton (1996) and for auto claim fraud detection by Viaene et al. (2005).

In the same field as Viaene et al. (2005), insurance fraud, Major and Riedinger (2002) presented a tool for the detection of medical insurance fraud. They proposed a hybrid knowledge/statistical-based system, where expert knowledge is integrated with statistical power. Another example of combining different techniques can be found in Fawcett and Provost (1997). A series of data mining techniques for the purpose of detecting cellular clone fraud is used. Specifically, a rule-learning program to uncover indicators of fraudulent behavior from a large database of customer transactions is implemented. From the generated fraud rules, a selection has been made to apply in the form of monitors. This set of monitors profiles legitimate customer behavior and indicate anomalies. The outputs of the

monitors, together with labels on an account's previous daily behavior, are used as training data for a simple Linear Threshold Unit (LTU). The LTU learns to combine evidence to generate high confidence alarms. The method described above is an example of a supervised hybrid as supervised learning techniques are combined to improve results. In another work of Fawcett and Provost (1999), Activity Monitoring is introduced as a separate problem class within data mining with a unique framework.

Another framework presented, for the detection of healthcare fraud, is a process mining framework by Yang and Hwang (2006). The framework is based on the concept of *clinical pathways* where structure patterns are discovered and further analyzed. The fuzzy expert systems are also experienced within a couple of studies. So are there Pathak et al. (2003), Bordoni et al. (2001) and Deshmukh and Talluru (1998). Stolfo et al. and Lee et al. delivered some interesting work on intrusion detection. They provided a framework, MADAM ID, for Mining Audit Data for Automated Models for Intrusion Detection. Next to this, the results of the JAM project are discussed. JAM stands for Java Agents for Meta-Learning. JAM provides an integrated meta-learning system for fraud detection that combines the collective knowledge acquired by individual local agents.

There is a special focus on fraud detection models because white collar fraudster daily research on

ways to cheat the available optimal systems that detect frauds so algorithms are rewritten almost on a daily basis to detect fraud.

The systems that would be considered are:

Clementine, Developed by Integrated Solutions Limited now SPSS

Darwin, Developed by Thinking Machines now ORACLE

1.1.1 Clementine

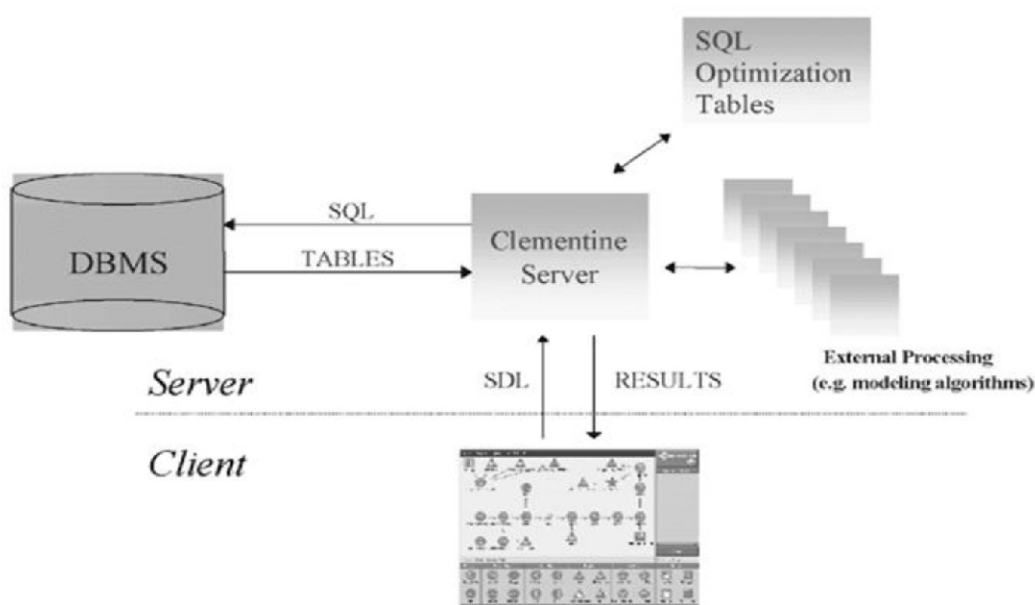


Fig 2.3: Clementine Server Architecture

Clementine also comes with the ability to use more than one building model. Some of the common models include Rule Induction, Graph, Clustering, Association Rules, Linear regression, Neural Networks etc

Limitations of the Clementine Package

The Clementine Server is a robust system but has the following limitations

Clementine does not have effective incremental methods to update mining

Clementine Server is a large scale, distributed Data mining software package with fast and scalable performance, a user-friendly visual workflow interface, and powerful analytical techniques.

Clementine provides Interactive and efficient mining process with Visual Workflow Interface using Visual Programming (that enables non-technical users to solve fraud problems) and also Easy and intuitive desktop.

results, this means if there is a better way of getting the same result a whole new dataset would be mined and a new set of result.

Clementine does not support certain data sources i.e. Dbase, Foxpro.

1.1.1 Darwin

Darwin is a parallel, scalable, data mining architecture with a Graphical User Interface “client” that provides access to the data mining functions and structured templates called Mining

Activities that automatically prescribe the order of operations, perform required data transformations, and set model parameters. The user interface also allows the automated generation of Java and/or SQL code associated with the data mining activities.

Darwin makes provide tools to transform data, including row operations (randomizing, splitting into training and testing sets, sub-sampling) and column operations (merging fields, creating new fields). The models used by Darwin for data Mining includes Decision Trees, Neural Networks, Regression, Kohonen Self-Organizing Maps, Clustering, Association Rules.

Limitations of the Darwin

1. There are a bit of hiccups with Contradiction and redundancy elimination.
2. Darwin is best when network bandwidth is at a premium (say, on very large databases). So it doesn't work optimally for small databases.

1.1.1 Internal Control

A system of well designed processes and procedures used for the purpose of fraud prevention and detection is generally referred to as Internal Control.

Internal controls can be split into two groups: active and passive internal control systems.

Active internal controls are signatures, passwords, segregation of duties etc. As Davia et al. (2000) put it; these can be compared with fences. They may

appear insurmountable at first sight, but like all fences, they have their weakness to be defeated by clever fraud perpetrators. And like a fence, once evaded, there is little or no continuing value in preventing or detecting fraud. (Davia et al., 2000)

Passive internal controls operate at a different level. Instead of preventing fraud, like active controls attempt to, the emphasis here is on deterring. Passive internal control systems induce a state of mind in the would-be perpetrator that strongly motivates him "not to go there". Examples of passive control systems are surprise audits, customized controls and audit trails. Passive control systems, when turned active if a company feels the need to do so (they suspect fraud), mainly make use of reporting tools, like providing different numbers and statistics for manual analysis.

2 METHODOLOGY

Davia et al. (2000) compared the art of fraud detection, with the art of fishing. In their work, they cited that an expert fishermen never simply go fishing for fish. Rather, they first decide what type of fish they have a taste for. Next, they decide the how, with what equipment, and where they will expertly search for that type of fish and that type alone.

Following this advice of first deciding what sort of fraud you are looking for, the health insurance fraud has to be narrowed down. In this project, we

will narrow it down to fraudulent claims.

The most common method of fraud detection using data mining generally adopted is the CRISP-DM (CRoss Industry Standard Process for Data

Mining). The diagram below gives a graphical overview of the processes involved in using the CRISP-DM.

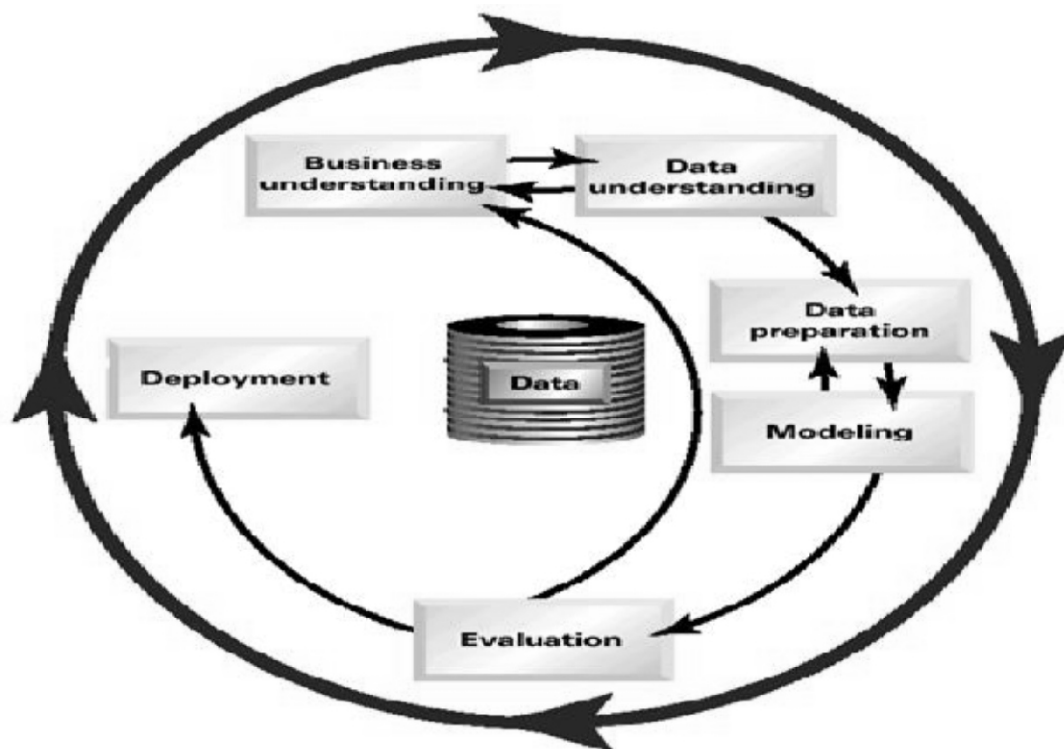


Fig 3.1: Data Mining using the CRISP-DM methodology

By utilizing the CRISP-DM process, the data mining process can implement more data mining goals faster, it would also be easier to introduce the workflow to a new person entering the project, and there is provision for the update and change of the system when the actions of the fraudsters change.

A data mining methodology used for fraud detection is a structured step-by-step approach to identifying transactions consistent with a fraud scenario, as described through the fraud data profile.

1.1 Identify the inherent fraud scheme:

The first step is to establish the scope of the fraud audit. Each business system has five to seven inherent fraud schemes. The audit plan should identify which inherent fraud schemes are within the fraud audit scope. Basically the Nigerian Health Insurance Scheme (NHIS) implements the managed care system which is a general classification used for a type of health insurance that provides comprehensive medical care based on a prepaid contract as opposed to traditional fee-for-service health insurance. Below are the identified

fraudulent schemes.

1.1.1 Billing Fraud

Some medical providers knowingly submits false medical bills by billing for services not rendered, billing for wrong procedure codes or billing for procedures of a medical necessity when procedures may have been elective or cosmetic in nature and not covered by health insurance.

1.1.2 Inflated Billing

This includes overbilling for services provided by any medical facility, doctor, chiropractor, laboratory, pharmacist etc. Generally there is an agreed amount earmarked by National Health Insurance Scheme for most services rendered by the health providers.

1.2 Building the Model

Building a data-mining model typically starts with identifying recurring patterns in the data, and then distilling those patterns in a way that helps communicate them to humans or other machines. Models can take the form of a graphical representation, a set of equations, a neural network, or even a collection of rules. Models can be applied to new data, or evaluated and refined in the presence of ever larger data sets. For this project the model would be based on clustering algorithms.

A Cluster is a collection of data objects that are

similar to one another within the same group and dissimilar to the objects in other groups. Clustering analysis is a form of learning by observation rather than by example hence the reason for its classification as an unsupervised learning.

There exists a large number of clustering algorithms in the literature. The choice of clustering algorithm depends both on the type of data available and on the particular purpose and application. Broadly, there are two types of data that can be used for clustering analysis and they use the following main memory based data structures.

1. **Data Matrix:** also called object-by-object variable structure, it represents object, such as persons, with p variables (attributes), such as age, height, weight, gender, race etc. This structure is in the form of relational table, or by n -by- p matrix (n objects $\times$ variables), as shown below.

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

1. **Dissimilarity Matrix:** also called object-by-object structure, it stores a collection of proximities that are available for all pairs of n objects. It is often represented by an n -by- n table as shown below.

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

Clusters of objects are computed based on their similarities or dissimilarities. Measures of dissimilarity or dissimilarity coefficients can be used to measure the quality of clustering. In general, dissimilarity $d(i, j)$ is a non-negative number that is close to 0 when i and j are near each other, and becomes larger when they are more different. Dissimilarities can be computed from correlation coefficients. Given n objects for clustering, the parametric **Pearson product-moment correlation** between two variables f and g is defined in the equation below, where f and g are variables describing the objects, m_f and m_g are the mean values of f and g , respectively, and x_{if} is the value of f for the i th objects and x_{ig} is the value of g for the i th objects.

$$R(f, g) = \frac{\sum_{i=1}^n (x_{if} - m_f)(x_{ig} - m_g)}{\sqrt{\sum_{i=1}^n (x_{if} - m_f)^2} \sqrt{\sum_{i=1}^n (x_{ig} - m_g)^2}}$$

The equation above would be used in this paper to verify the similarities or dissimilarities of the clusters generated from the claims, separating suspicious health insurance claims and the genuine ones into groups using the *k-means*

Algorithm.

1.1.1 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 is 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.
- 1) Arbitrary choose k objects as the initial cluster centers;
 - 2) Repeat
 - 3) (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.
 - 4) Update the cluster means, i.e., calculate the mean value of the objects for each cluster.
 - 5) Until no change
 - 6)

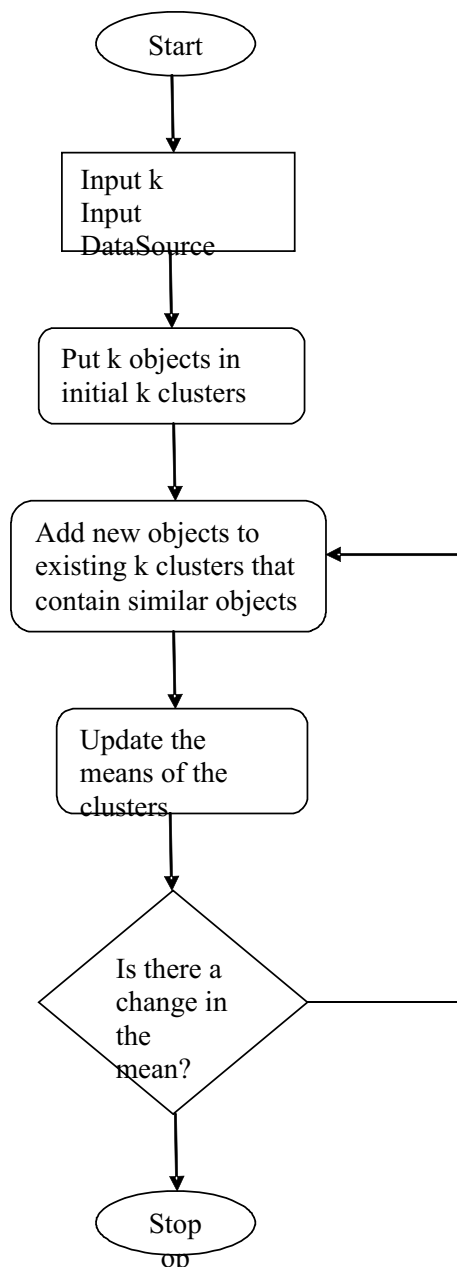


Fig 3.2: Flowchart for the K-means algorithm

1.1.1 Data Feed

The data source would be introduced to the system via this interface. The supported file types

would be loaded to the system via a file uploader.

The figure below shows the proposed interface.

The interface consists of a rectangular box containing the following elements:

- Data Source**: A label followed by a text input field with a dropdown arrow 'V' and a **Browse** button.
- Number of Clusters**: A label followed by the text *(optional)* and a small text input field.
- Generate Clusters**: A button located at the bottom left of the interface.

Fig 3.3 : Interface for the data feed

1.1 Appropriate Technologies and Tools

The tools and technologies available that would be appropriate to implement the desired solution for health insurance fraud detection are discussed below according to their classes of need.

1.1.1 Database Development and Model

- **ORACLE RDBMS 10g**

- The Oracle Database (commonly referred to as Oracle RDBMS or simply Oracle) consists of a relational database management system (RDBMS). The Oracle RDBMS stores data logically in the form of tablespaces and physically in the form of data files. Tablespaces can contain various types of memory segments, such as Data Segments, Index Segments, etc. Segments in turn comprise one or more extents.

Extents comprise groups of contiguous data blocks. Data blocks form the basic units of data storage. At the physical level, datafiles comprise one or more data blocks, where the block size can vary between data-files.

- **Microsoft SQL Server 2008**

SQL Server 2008 is an RDBMS that stores data from structured, semi-structured, and unstructured documents, such as images and rich media, directly within the database. SQL Server 2008 delivers a rich set of integrated services that enables users do more with your data such as query, search, synchronize, report, and analyze.

- **MYSQLRDBMS 5.0**

The MySQL database has become the world's most popular open source database because of its consistent fast performance, high reliability and ease of use. Not only is MySQL the world's most popular open source database, it's also become the database of choice for a new generation of applications built on the LAMP stack (Linux, Apache, MySQL, PHP / Perl / Python.) MySQL runs on more than 20 platforms including Linux, Windows, OS/X, HP-UX, AIX, Netware, giving you the kind of flexibility that puts you in control.

1.1.2 Application Development Suites

- **Microsoft Visual Studio**

Visual Studio (VS) is an integrated development tool that allows the development of console and GUI applications on platforms that support the Microsoft .NET framework. VS contains the Microsoft programming languages of Visual Basic, Visual C#, Visual C++ and Visual J# and the functionality to create web applications. The advantages of VS are that GUI creation is made very simple with a drag and drop process. The parsing and processing of XML is made easy with powerful built in libraries. Dot Net's most obvious advantage is that it offers a fully integrated range of technologies all designed by the same vendor. This means that from the database through to the development environment there is a lot of

integration between the different parts of the framework. Disadvantages of VS include the .NET framework is limited to mostly Windows based operating systems. The enterprise suite of VS is a licensed application of around 120,000 Naira.

- **Netbeans & Eclipse**

Netbeans and Eclipse are integrated development environments for the development of Java applications (although C/C++ development plugins are available). Netbeans has better support for J2EE, J2ME, XML and XSLT processing and debugging within the web applications server, Tomcat. Eclipse is probably more widely used and has better support. Both are open-source meaning no license costs. In addition Java runs on a virtual machine that is supported by the majority of platforms.

4.0 Conclusion

Having implemented and tested this data mining solution, it can be concluded that using data mining techniques to detect health insurance is a viable tool to combat fraudulent activities in the Health care sector.

Data Mining is the extraction of hidden predictive information from large databases. This is a new powerful technology with great potential to help health insurance schemes detect health insurance fraud, fish out shady claims and make excellent decisions in providing health care services. "The

automated, prospective analyses offered by data mining move beyond the analyzes of past events provided by retrospective tools typical of decision support systems" (Pilot G)

4.1 Recommendations

Although the Statement of Problem is about fraud detection, presented results are only the first part of fraud detection, namely outlier detection (anomaly detection). After examining manually the selected observations, we can then clarify if they were fraudulent, errors, extreme values or procedural errors.

Human intervention would be needed to verify that a suspected claim is fraudulent. For easier predictive analysis, the next version should implement more graphical user interfaces in the form of pie charts, scatter diagrams. As a way of recommendation, subsequent systems could merge the capabilities in WEKA (an open source data mining tool) and ORACLE data Miner (ODM).

Data Mining as a fraud detection tool is novel to the Nigerian health care sector, it is therefore recommended that data mining be integrated into the individual Information Management Systems (IMS) implemented by different Health Management Organizations (HMO) as opposed to the systems being run presently where data mining is an alienated process from the management of Information in the Health care sector. This singular step can transform the rich

information in the Managed care to useful knowledge.

REFERENCES

- 1) EA Lisbon et al (2007) *Data Mining for Fraud Detection: Toward an Improvement on Internal Control Systems?*
- 2) Bologna, G. and R. Lindquist (1995). *Fraud Auditing and Forensic Accounting*. John Wiley & Sons.
- 3) Davia, H. R., P. Coggins, J. Wideman, and J. Kastantin (2000). *Accountant's Guide to Fraud Detection and Control (2 ed.)*. John Wiley & Sons.
- 4) Leonard W. Vona (2008) *Fighting Fraud with Data Mining & Analysis*
- 5) Awosika L (2005) *Health Insurance and Managed Care in Nigeria*
- 6) Frieden, J. (1992). *Health care fraud detection enters the information age*. *Business and Health*
- 7) <http://www.dinkla.net/fraud/> *Artificial Intelligence and Fraud detection management*
- 8) Maes, S., et al (2002). *Credit card fraud detection using bayesian and neural networks*.
- 9) Major, J. and D. Riedinger (2002). *EFD: a hybrid knowledge/statistical-based system for the detection of fraud*. *The Journal of Risk and Insurance* 69 (3), 309{324.
- 10) Michalski, R. S., I. Bratko, and M. Kubat (1998). *Machine Learning and Data Mining - Methods and Applications*. John Wiley & Sons Ltd.
- 11) Murad, U. and G. Pinkas (1999). *Unsupervised profiling for identifying superimposed*

fraud. Lecture Notes in Computer Science 1704, 251{262.

12) Shao, H., H. Zhao, and G. Chang (2002). *Applying data mining to detect fraud behaviour in customs declaration*.

13) Viaene, S., G. Dedene, and R. Derrig (2005). *Auto claim fraud detection using*

bayesian learning neural networks. Expert Systems with Applications 29, 653{666.

14) Wasserman, S. and K. Faust (1998). *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

15) Wells, J. (2005). *Principles of Fraud Examination*. John Wiley & Sons.