International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 3, Issue 7, July 2014 **ISSN 2319 - 4847**

Application of The K-Means Clustering Algorithm In Medical Claims Fraud / Abuse Detection

[1]Leonard Wafula Wakoli, Abkul Orto[2] and Stephen Mageto[3]

[1]Department of Computer Science, Kenya Methodist University - Kenya

[2]Department of Information Technology, Meru University of Science and Technology - Kenya

[3]Department of Information Technology, Meru University of Science and Technology - Kenya

ABSTRACT

This paper is about a system which applies a modified K-Means algorithm[12] to flag out suspicious claims for further scrutiny has been developed. The Java programming Language and mySQL database tools were used. The K-Means algorithm is well known for its efficiency in clustering large data sets. However, a major limitation of this algorithm is that it works only with numeric values, thus the method cannot be used to cluster real-world data containing categorical values. To counter this, data sets were converted to numeric data whereby ailments were listed and matched with patients. The presence of the ailment was represented by a one (1) and the absence was represented by a zero (0). To get the data, a total of 15 insurance companies in Kenya out of 31 were randomly selected and a pre-tested questionnaire was used to collect data. 15 insurance companies out of 31 is close to 50%, which is a very good representative of the entire population. 67 % of the respondents indicated that the people involved in the processing of claims were billing for services that were not rendered. The results also showed that all the companies had internal control mechanisms to address the problem and 47% of the respondents said the internal controls were not efficient. 87% of the respondents indicated that the common member fraud cases involved mmembership substitution including card abuse.

Keywords: billing , K- Means, 0s and 1s, clustering, Euclidean distance

1. INTRODUCTION

To develop a medical fraud detection system applying the K-Means [12], the data collected were converted to 0s and 1s and the Euclidean distances [14],[16] were computed and these distances were used to cluster given data sets. The average claim amount for a given cluster was computed and claims that very high figures far away from the computed average claim within that cluster were flagged for further scrutiny or rejected altogether. Clustering [1],[10] is a fundamental operation in data mining[8]. It is useful in a number of tasks, such as classification, aggregation and segmentation or dissection. For example, by partitioning objects into clusters, interesting object groups may be discovered, such as the groups of motor insurance policy holders with a high average claim cost [3], or the groups of clients in a banking database having a heavy investment in real estate. Clustering [1],[10],[12]. is a popular approach to implementing the partitioning operation. Clustering methods partition a set of objects into clusters such that objects in the same cluster are more similar to each other than objects in different clusters according to some defined criteria. Statistical clustering methods [11] partition objects according to some (dis)similarity measures, whereas conceptual clustering methods cluster objects according to the concepts objects carry [9] data sets. The data sets to be mined often contain millions of objects described by tens, hundreds or even thousands of various types of attributes or variables (interval, ratio, binary, ordinal, nominal, etc.). This requires the data mining operations and algorithms to be scalable and capable of dealing with different types of attributes. This paper focuses on the K-Means [12] clustering method with a minor modification to achieve more efficiency. This technique is efficient for processing large data sets. Therefore, it is best suited for data mining. However, the K-Means algorithm[12] only works on numeric data, i.e., the variables are measured on a ratio scale , because it minimises a cost function by changing the means of clusters. This prohibits it from being used in applications where categorical data are involved. Unlike statistical clustering methods, the K-Means algorithm [12] is based on a search for objects which carry the same or similar concepts. Therefore, its efficiency relies on good search strategies. Setting the number of clusters before hand is another major drawback when using the K-Means algorithm since it involves guess work as there is no formula that can be used to calculate the exact number of clusters required for a given data. The K-Means algorithm is well known for its efficiency in clustering large data sets. The algorithm's limitation of working on numeric data only was addressed by converting the data sets to numeric data whereby ailments were listed and matched with patients. The presence of the ailment was represented by a one (1) and the absence was represented by a zero (0) and then these numeric data are fitted in the Euclidean formula for distance measure which is used in clustering the records.

2. LITERATURE REVIEW

The K-Means Clustering Technique

The K-Means clustering technique is a non-hierarchical approach used to form good clusters by specifying a desired number of clusters, say, k, then assigning each case (object) to one of k clusters so as to minimize a measure of dispersion within the clusters [2] A very common measure is the sum of distances or sum of squared Euclidean distances from the mean of each cluster [5]. The problem can be set up as an integer programming problem but because solving integer programs with a large number of variables is time consuming, clusters are often computed using a fast, heuristic method that generally produces good (but not necessarily optimal) solutions. The K-Means algorithm is one such method. The K-Means method is a very popular algorithm for clustering high-dimensional data. Initiated with k arbitrary cluster centres, it assigns every data point to its nearest center, and then readjusts the centers, reassigns the data points, until it stabilizes. The K-Means method terminates in a local optimum, which might be far worse than the global optimum. However, in practice it works very well. It is particularly popular because of its simplicity and its speed: "In practice, the number of iterations is much less than the number of samples"[6].The K-Means clustering algorithm divides the data set into a predetermined number, k, of clusters. These clusters are centred at random points in the record space[4][7]. Records are assigned to the clusters through an iterative process that moves the cluster means (also called cluster centroids) around until each one is actually at the centre of some cluster of records. (See Figures 2.1, 2.2 and 2.3

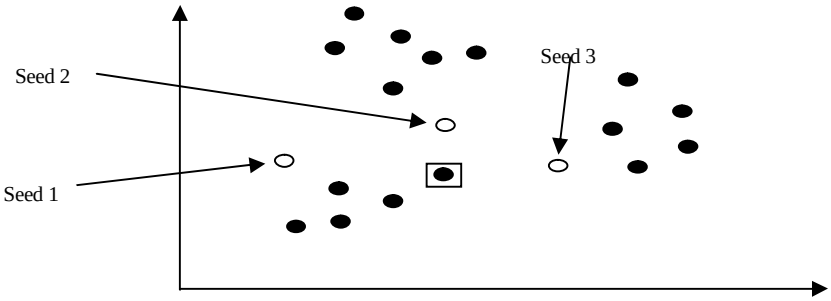

Figure 2.1 Initial cluster seeds (from Berry & Linoff 2000).

In the first step, k data points are selected to be the seeds more or less arbitrarily. Each of these seeds is an embryonic cluster with only one element. In the example shown in figure 1, k is 3.

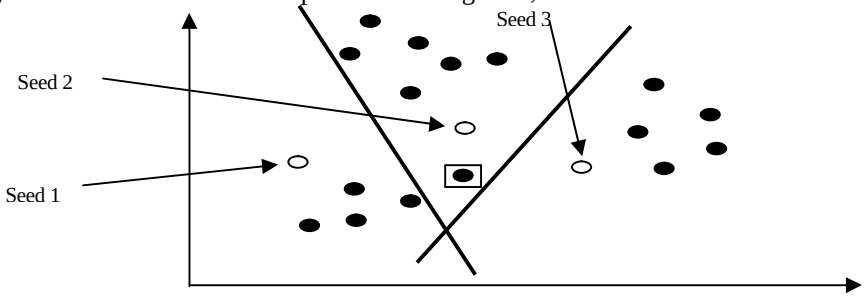

Figure 2.2: Initial clusters and inter-cluster boundaries (from Berry & Linoff 2000).

In the second step, each record is assigned to the cluster whose centroid is nearest to that record. This forms the three clusters shown in Figure 2.2 with the new inter-cluster boundaries. Note the boxed record which was assigned to cluster 2 (Seed 2) initially now becomes part of cluster 1

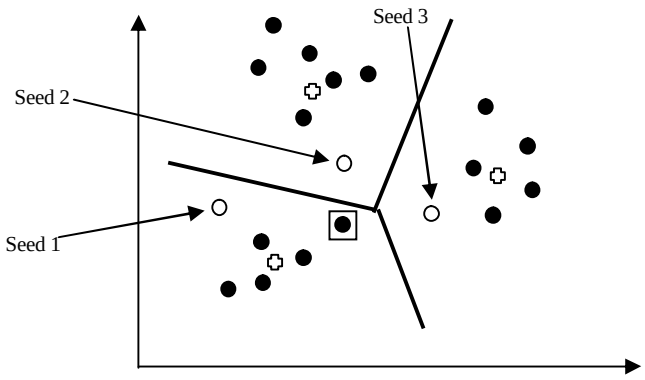

Figure 2.3: New clusters, their centroids marked by crosses and inter-cluster

2.3 Related Efforts

In [17], an on-line discounting learning algorithm to indicate whether a case has a high possibility of being a statistical outlier in data mining applications such as fraud detection is used for identifying meaningful rare cases in health insurance pathology data from Australia's Health Insurance Commission (HIC). The performance of a k -Nearest Neighbor (kNN) algorithm with the distance metric being optimized using a genetic algorithm was applied in a real world fraud detection problems faced by the HIC. In Chile, a single neural network to detect fraudulent medical claims was implemented in another healthcare insurance company. This scheme utilizes all the data available in arriving medical claim for constructing a unique vector which is evaluated by the single neural network.

3. METHODOLOGY

The methodology describes the research design, subjects, target population, sampling procedures and sample size informed consent, Research instruments, validity of the Research instrument, reliability of the instrument, Data collection, Survey results and data analysis the statistical techniques that were used to analyse the data collected as well as software development tools. Potential problems, limitations and ethical considerations are presented.

Research Design

Research design can be thought of as the structure of research -- it is the "glue" that holds all of the elements in a research project together. It is the investigator's action plan for answering the research questions and realizing the objectives [15]. Survey designs are often known as correlational designs to denote the tendency to reveal relationships between variables and to draw attention to their limited capacity in connection with elucidation of casual processes. The design was very appropriate because it was able to elicit a diverse range of information about the area of study.

Subjects

The unit of analysis is the study fraud cases in the health care sector in Kenya. Focus is based on the assumption that medical claims can be clustered using the K-Means technique and detecting fraud using the Euclidean distance measure.

Target population

The thirty six (36) active insurance companies in Kenya

Sampling procedures and Sample size

Subjecting the whole target population to investigation is usually an impossible taskas a result of prohibitive costs and the time involved. This then calls for a sample, which is a subset of the target population through which the requisite information can be obtained at reasonable costs [15]. Samples should be as representative as possible, because too-small-a-sample is likely to yield under-estimated information that may not reflect the actual population characteristics or perceptions . In situations where a population is too small to be sampled, it is logical to sample all the elements [15].

This study applied random sampling procedure to obtain samples of the insurance companies as the ultimate units of study. Out of the target population of 36 insurance companies, 15 were randomly selected for study, being close to half the total population of Insurance companies in Kenya As proposed in Chapter one, this study was designed to examine medical insurance claim fraud/abuse cases at various service providers in Kenya and the subsequent designing and development of a system to address the problem.

Research Instrument

The study had only one set of the research instrument which focused the establishment of the extent of medical fraud in Kenya, current mechanisms of detecting the medical claim fraud and the root causes of this type of fraud. The instrument had 15 questions

Validity of the Research instrument

Validity of research instruments is a key element of an accomplished study. It denotes the extent to which an instrument captures it purports to measure[15]. The acceptable level of validity largely depends on logical reasoning, experience and professionalism of the investigator [16]. Pilot testing is a crucial step in the research process because it helps in refining instruments so that they capture the intended data. Pilot testing reveals what works and what do not, for example, vague questions and unclear instructions[13].

Reliability of Instruments/tools of the survey

The only instrument/tool that was used to collect data was the Questionnaire which was field-tested after it was designed. The questionnaire included a mix of check -the-box items, fill-in blanks and closed ended questions.

Data Collection

This started with the designing of the research instrument and recruitment of one research assistant to assist with data collection activities. This followed consensus building involving the investigator and the research assistant. The purpose was to discuss the items contained in the instrument for familiarity, logical requirements and acquisition of the research permit. The data from each of the questions were summarized and presented in summary tables then subjected to statistical analysis techniques.

4. FINDINGS AND DISCUSSION

4.1 Findings

Common types of fraud that make claims to be rejected:

A. Billing for services that were not rendered.
B. Billing for more expensive services than were actually provided
C. Performing medically unnecessary services so as to generating insurance payments
D. Falsifying a patient's diagnosis to justify tests, surgeries or other procedures that are not medically necessary
E. Double billing
F. Other (specify)

Table 4.1: Common reasons for claim rejection

Reason for Claim Rejection	No. of respondents	% age of respondents
A	10	67
B	8	53
C	3	20
D	9	60
E	2	13
F	1	06

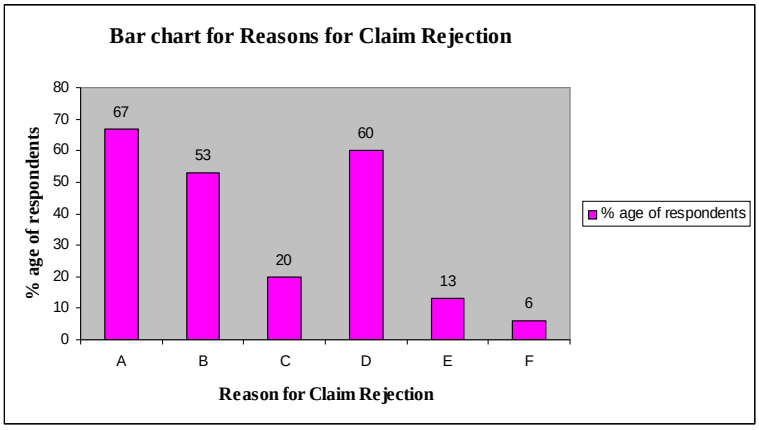

Figure 4.1: Common Reasons for Claims Rejection

Figure 4.1 shows that billing for services that were not rendered and falsifying a patient's diagnosis to justify tests, surgeries or other procedures that are not medically necessary are the most common fraudulent activities that cause the rejection of medical claims.

How companies determine fraudulent claims:

Control	No. of Respondents	% age of Respondents
Internal RiskControl	15	100
By chance	13	87
Other	5	33

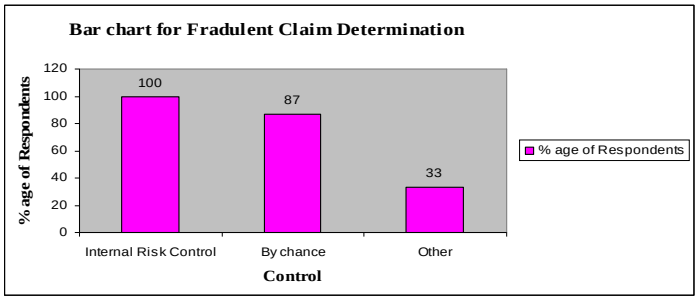

Figure 4.2: Fraudulent Claim determination

The results show that most of the insurance companies have put in place internal controls to try to address the issue of medical claim fraud

4: Efficiency of internal controls in fraud detection:

Table 4.3: Efficiency of Internal control mechanisms

Type of Response	No. of Respondents	% age of Respondents
Efficient	7	47
Not efficient	8	53
Total	15	100

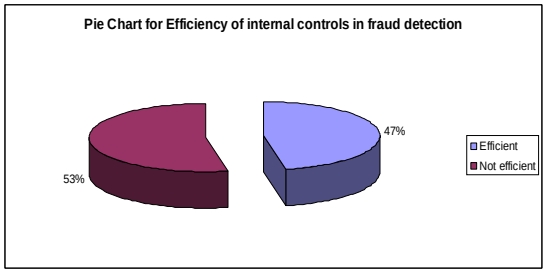

Figure 4.3: Efficiency of Internal control mechanisms

Method(s) used for uncovering fraud in organisation

A. Internal controls
B. Internal Audit
C. Notification of employee
D. Accident
E. Anonymous tip
F. Notification by customer
G. Notification by a Regulatory or Law enforcement agency
H. Notification by Vendor
I. External Audit

Table 4.4: Methods for uncovering fraud

Method for uncovering fraud	No. of Respondents	% age of Total No of respondents
A	15	100
B	15	100
C	9	60
D	7	47
E	8	53
F	8	53
G	4	27
H	3	20
I	5	33

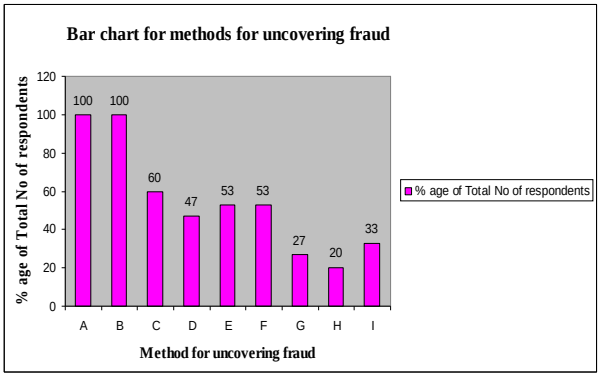

Figure 4.4: Methods for uncovering fraud

Internal controls and internal audits were found to be the most commonly used methods for uncovering fraud.

4.2 Summary of the Results Analysis

The results clearly show that insurance companies in Kenya are finding it difficult to address the problem of Medical Claim Fraud. However, most of the companies are trying to put up a spirited fight to combat the said fraud using various approaches to combat the Insurance Claims Fraud.

4.3 Clustering using the Euclidean distances concept

a) N-dimensional distance

The Euclidean distance between points P = (p_1, p_2, …,p_n) and Q = (q_1,q_2, …q_n), in Euclidean *n*-space is defined as:

$$\sqrt{((p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2)} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \quad (1)$$

Using this Euclidean algorithm, medical claim records are clustered using the distance measure technique.

Consider two records: Record 1 and Record 2 as shown in (b) below

b) Euclidean and Manhattan Distance between two records

Table 4.5 : Euclidean and Manhattan Distance between two records

Variable	Record 1 sore throat swollen neck, and fractured shoulders	Record 2 Absolute Difference Square ruptured toe and broken elbow	Absolute Difference	Square
Shoulder	1	0	1	1
Knee	0	0	0	0
Ankle	0	0	0	0
Fracture	1	0	1	1
Back	0	0	0	0
Right	0	0	0	0
Left	0	0	0	0
Cut	0	0	0	0
Elbow	0	1	1	1
Toe	0	1	1	1
Finger	0	0	0	0
Thumb	0	0	0	0
Eye	0	0	0	0
Ear	0	0	0	0
Throat	1	0	1	1
Leg	0	0	0	0
Hand	0	0	0	0
Lower	0	0	0	0
Trauma	0	0	0	0
Wrist	0	0	0	0
Hip	0	0	0	0
Broken	0	1	1	1
Strain	0	0	0	0
Stress	0	0	0	0
Palm	0	0	0	0
Sore	1	0	1	1
Swollen	1	0	1	1
Ruptured	0	1	1	1
Head	0	0	0	0
Scratch	0	0	0	0
Foot	0	0	0	0
Neck	1	0	…..1	1
Total	**6**	**4**	**10**	**10**
Distance Measure	3.16	10		

The Euclidean distances are calculated using the following formula:

$$d_{i,j} = \left(\sum_{k=1}^{n} (x_{i,k} - x_{j,k})^2 \right)^{½} \quad (2)$$

where i,j = records and n = number of variables

c) Comparing two medical claims

Record 1 shows sore throat, swollen neck and fractured shoulders whereas Record 2 shows ruptured toe and broken elbow. The system checks these ailments against the column for variables and the existence of the same is marked with a one (1) and the absence is marked with a zero (0) as shown in Table 4.5. The absolute difference and corresponding square are computed for each row. The totals and corresponding Euclidean distances are then calculated. These distances are the ones that will then be used for clustering.

d) The clustering process and claim analysis process

Medical claim details are entered one by one. When all the working day's medical claim data are entered into the system, the Euclidean distance between each pair of the claim forms is calculated as shown in Figure 4.5. The user of the system sets the desired maximum distance for the records to belong to the same cluster. The sum claimed for all the claim forms is computed and the average amount determined. If the amount for a given claim exceeds a set amount for that cluster, then that particular claim form is rejected. The same process is repeated for different set distances and the rejected claims is listed.

4.4 Entry of Claim details

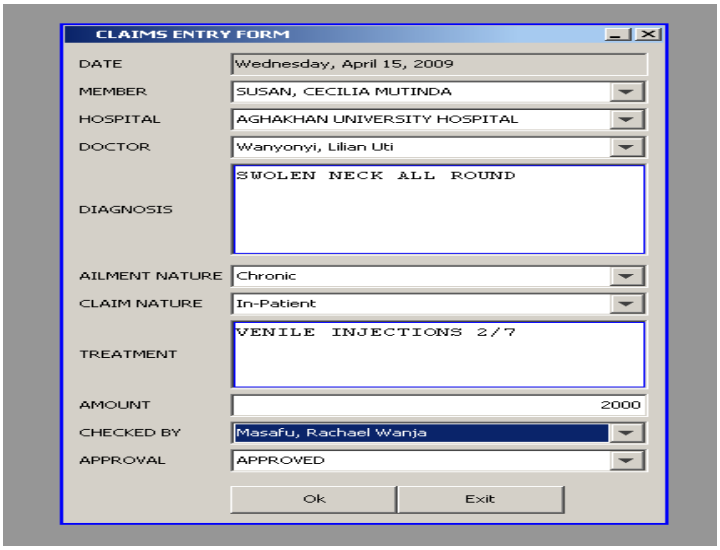

Figure 4.5: Entry of Claim details

4.5 Claims Analysis

The Analysis process:

The Analysis process starts when the user clicks the "Analyse" button shown in Figure 6.3 The program then computes the Euclidean distance between each pair of the claim forms using the Euclidean distance formula:

$$d_{i,j} = \sum_{k=1}^{n} |x_{i,k} - x_{j,k}| \quad (3)$$

where i,j = records, n = number of variables

The user of the system sets the desired maximum distance for the records to belong to the same cluster, Varying the distance also changes the number of clusters that are formed. The assumption here is that the greater the distance, the lower the number of clusters for the given data set and vice versa. The sum claimed for all the claim forms is computed and the average amount determined. If the amount for a given claim exceeds the set amount for that cluster, then that particular claim form is rejected.

CLAIMS ANALYZER

```
package com.wakoli.claimsanalyser;
import javax.swing.*;
import javax.swing.event.*;
import java.awt.*;
import java.awt.event.*;
import java.sql.SQLException;
import com.wakoli.claimsanalyser.methods.*;
```

```
public class ClaimsAnalyzer extends JFrame{
JPanel
flags = new JPanel();
JLabel
flag[ ]= {
                    new JLabel(new ImageIcon("com/wakoli/claimsanalyser/support/images/flag.gif")),
                    new JLabel(new ImageIcon("com/wakoli/claimsanalyser/support/images/flag.gif")),
                    new JLabel(new ImageIcon("com/wakoli/claimsanalyser/support/images/flag.gif")),
                    new JLabel(new ImageIcon("com/wakoli/claimsanalyser/support/images/flag.gif")),
                    new JLabel(new ImageIcon("com/wakoli/claimsanalyser/support/images/flag.gif")),
                    new JLabel(new ImageIcon("com/wakoli/claimsanalyser/support/images/flag.gif")),
                    new JLabel(new ImageIcon("com/wakoli/claimsanalyser/support/images/flag.gif")),
                    new JLabel(new ImageIcon("com/wakoli/claimsanalyser/support/images/flag.gif")),
                    new JLabel(new ImageIcon("com/wakoli/claimsanalyser/support/images/flag.gif"))},

logo = new JLabel(new ImageIcon("com/wakoli/claimsanalyser/support/images/eck.gif"));
static JDBCAdapter dataModel;
JDesktopPane desktop = new JDesktopPane();Login pass = new Login();
ChangePassword change = new ChangePassword();
boolean changed=false;
boolean loaded=false;
public ClaimsAnalyzer(){
                  Color background=new Color(150,160,130);
                  flags.setBackground(background);
                  flags.setLayout(new GridLayout(8,1));
                  int i;
                  for(i=0;i<=7;i++){
                  //flags.add(flag[i]);
     }
   ContentPanel contentPane = new ContentPanel("com/wakoli/claimsanalyser/support/images/eck.gif");
                                  desktop.setBackground(background);//(150,160,130));
                                   desktop.add(pass,JLayeredPane.MODAL_LAYER);
getContentPane().add(flags,BorderLayout.WEST);
getContentPane().add(desktop);//Can either add JDesktopPane or set it to be ContentPane
//getContentPane().add(contentPane); //this sets a background image
contentPane.setOpaque(false);
setDefaultCloseOperation(JFrame.EXIT_ON_CLOSE);
setTitle("THE CLAIMS ANALYZER");
setSize(1024,768);//800,500);
Label c=new Label();setLocationRelativeTo(c);
setUndecorated(true);
setVisible(true);
ChildFrame.select(pass);
//AudioPlayer.play("com/wakoli/claimsanalyser/support/audio/wape vidonge.wav");
```

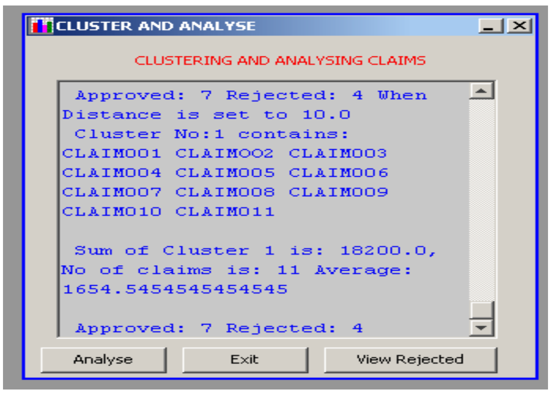

Figure 4.6: Claims Analysis

Figure 4.6 shows the dialogue box for clustering and analyzing the claims in which case the distance was set to 10 units whereby the cluster had 11 records out of which 7 were approved and 4 rejected. Clicking on the "View Rejected" button displays a screen as the one shown in Figure 4.7

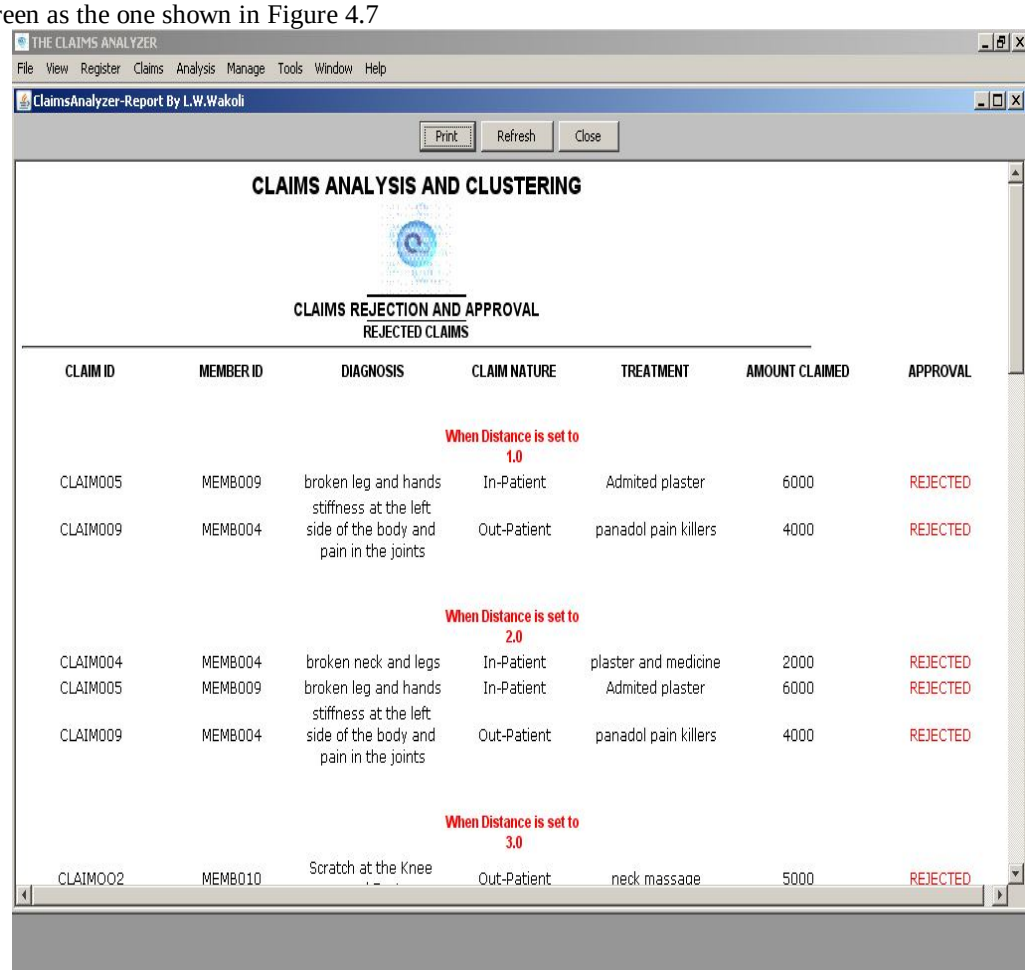

Figure 4.7: Rejected Claims

Figure 4.7 also shows that the greater the distance, the higher the rejection rate. This is because increasing the Euclidean distance means having more records within a cluster; hence the chances of netting more fraudulent ones are higher.

5. CONCLUSIONS

This paper shows the successful application of the K-Means clustering algorithm to medical claims records. The medical claims were successfully clustered and the average amount claimed per cluster was computed. Claims that were far away from the average were flagged for further scrutiny. Hence the prototype can be used isolate flag suspicious claims that can be subsequently rechecked. This prototype can immensely increase the medical claim fraud detection rate which in turn will yield savings that cover operational costs and allowed to increase the quality of the health care coverage, fully justifying the investment.

REFERENCES

[1]. Anderberg, 1973: Anderberg, M. R. (1973) Cluster Analysis for Applications.
[2]. Andrew Moore: "K-Means and Hierarchical Clustering - Tutorial Slides" http://www-2.cs.cmu.edu/~awm/tutorials/kmeans.html (Accessed on 17 December 2008).
[3]. Berry, Michael J. A., and Gordon Linoff (2000), Mastering Data. Mining)
[4]. Brian Everitt, Sabine Landau, Morven Leese, Cluster analysis Edition: 4, illustrated Published by Arnold, 2001 ISBN 0340761199, 9780340761199
[5]. Brian T. Luke: "K-Means Clustering http://fconyx.ncifcrf.gov/~lukeb/kmeans.html (Accessed on 23 January 2009).
[6]. DUDA, R. and HART, P. 1973. Pattern Classification and Scene Analysis. John Wiley & Sons, New York, NY.
[7]. Dunn J.C. (1973): "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", Journal of Cybernetics 3: 32-57

[8]. Fayyad, Usama; Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996). "From Data Mining to Knowledge Discovery in Databases". http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf. Retrieved on 2008-12-17. (Accessed on 23 January 2009).
[9]. Hand D.J., Adams N.M., and Bolton R.J. (eds.) (2002) Pattern detection and discovery. Springer .
[10]. Hans-Joachim Mucha and Hizir Sofyan: "Nonhierarchical Clustering" http://www.quantlet.com/mdstat/scripts/xag/html/xaghtmlframe149.html (Accessed on 23 January 2009).
[11]. Hartigan, J. A. (1975). Clustering Algorithms. Wiley. MR0405726. ISBN 0-471-35645-X
[12]. MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observations." In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, edited by L. M. Le Cam and J.
[13]. Mugenda, O. M. and Mugenda, A. G. (1999) Research Methods: Quantitative and. Qualitative Approaches.
[14]. Paul E. Black, "Euclidean distance", in Dictionary of Algorithms and Data Structures [online], Paul E. Black, ed., U.S. National Institute of Standards and Technology. Available from: http://www.itl.nist.gov/div897/sqg/dads/HTML/euclidndstnc.html (Accessed on 23 January 2009).
[15]. UNESCO, 2004 (Best and Khan, 2004; Mugenda and Mugenda, 1999; Nachmias and Nachmias, 1996).
[16]. Wikipedia, http://en.wikipedia.org/wiki/Euclidean_distance (Accessed on 23 January 2009).
[17]. Yamanishi K., Takeuchi J., G. Williams, and P. Milne, "On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms," In Data Mining and Knowledge Discovery vol. 8, pp. 275–300, 2004.

Mr. Leonard .W. Wakoli is a Lecturer, Kenya Methodist University (Department of Computer Science and Business Information Systems) – Nakuru Campus With well over 10 years of experience. He is also a Cyber Security Consultant, PhD Candidate (Business Information Systems – Security), Jaramogi Oginga Odinga University – Bondo – Kenya. He holds a MSc. Software Engineering – JKUAT- Kenya, a Post Graduate Dip. in Management of Information Systems (MIS) – Greenwich University – UK, a BSc. in Mathematics and Computer Science – JKUAT- Kenya, a Dip. in Science Education – KSTC-Kenya. He is also an Environmental Impact Assessment Auditor , a Motivational Speaker and a Certified Public Accountant K – Finalist. His area of interest in research is Cyber Security, now working on "Harnessing the Power of Intrusion Detection Systems".

Mr. Abkul Orto is a Lecturer in the Department of Information Technology in the School of Information Technology and Engineering and Director of Open, Distance and eLearning , Meru University of Science and Technology, Kenya. He holds a Master of Science (Computer Based Information Systems) from University of Sunderland, United Kingdom. He previously taught at KCA University for seven years as a full time faculty and also at Institute of Computer Science and Information Technology and IT Centre of the Jomo Kenyatta University of Agriculture and Technology, Kenya. His research interests include Human Centred Computing, Spoken Language processing focusing on African Languages, Information retrieval and security. He has a teaching experience of 12 years research interests include Human Centred Computing, Spoken Language processing focusing on African Languages, Information retrieval and security. He has a teaching experience of 12 years.

Mr. Stephen Mageto is a Lecturer in information Technology and Management and Chairman of Department at Meru University of Science and Technology. He holds a Master of Studies (Information Technology and Management) degree from Madurai Kamaraj University, 2004. He has over 8 years of teaching experience. He has a strong research interest in the impact and integration of ICT applications on community development and strategic information systems.