Application of The k-means Clustering Algorithm In Medical Claims Fraud / Abuse Detection

Leonard Wafula Wakoli

A thesis submitted in partial fulfillment for the degree of Masters of Science in Software Engineering in the Jomo Kenyatta University of Agriculture and Technology

2012

ABSTRACT

There is a serious threat to the survival of the insurance industry in Kenya due to fraudulent medical insurance claims. Several studies have shown that unsuspecting insurance companies lose huge sums of money annually settling fraudulent claims. It is feared that if this issue is not stemmed, many insurance companies offering the medical covers may collapse. The aim of this thesis was to determine the extent of fraudulent medical claims within insurance companies in Kenya, establishing the measures in use to counter the problem and thereby develop a system to detect these fraudulent activities.

The systems was developed applying the k-means algorithm using mySQL and Java software as development tools. The *k*-means algorithm is well known for its efficiency in clustering large data sets. 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. However, this limitation 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).

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. All the data were analyzed using Microsoft Excel. The results showed that all the respondents who filled the questionnaire confirmed that there were cases where claims could be rejected. 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. To develop a medical fraud detection system applying the kmeans, the data collected were converted to 0s and 1s and the Euclidean distances are 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 were flagged for further scrutiny or rejected altogether.