AN IMPROVED MODEL FOR MEDICAL RECORDS CLASSIFICATION

C.G ILOFULUNWA [(1)] S. OOKIDE [(2)] AND I. JMGBEAFULIKE [(3)]
[1, 3] Dept. of Computer Science, Chukwuemeka Odumegwu Ojukwu, University, Anambra State
[2]Dept. of Computer Science, Nandi Azikiwe University, Awka, Anambra State

ABSTRACT
This research deals with an improve model for medical records classification, Shanahan hospital is used as a case study. In this context, many classification models and relevant systems are still being developed in order to assist the strategic management mechanisms. Data mining techniques have been used to uncover hidden patterns and relations, to summarize the data in a novel ways that both understandable and useful to predict future trends and behaviour in business. The major aim of this project is to address the inadequacies of the records keeping using data mining techniques. The hospital services are studied and relevant information on patient's records is collected and categorized.This project proposed a data mining model that can classify patients medical records based on the similarity of previous illness and prescription using clustering technique. This system manages metadata that are extracted from the medical data files automatically or are created by researchers. The results obtained from the web-based system ascertain that a great improvement has been made on the way patient's records are managed.

Keywords: Medical records, Data mining, Classification, Clustering, Hybrid, Clustering Algorithmsand Predictive

1. Introduction

Data Mining (DM) is a discipline that seeks to discover patterns in large datasets for consultation and analysis. Data mining is an essential part of knowledge management (KM). Efficient use of big data is definitely improving health care in the broad sense, including the way doctors and hospital work or treat patients, as well as boasting research, saving costs at the enterprise level, and improving public or private health police management It has been proven that data mining can enhance the KM process with better knowledge. Wang, and Wang, (2008) pointed that data mining can be useful for KM to share common knowledge of business intelligence context among data miners and to use data mining as a tool to extend human knowledge. It is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. (Kovalerchuk, Vityaev, and Ruiz, 2001).

Data mining aims to analyze a set of given data or information in order to identify novel and potentially useful patterns (Fayyad, Piatetsky-Shapiro and Smyth, 1996). Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge driven decisions. Its' tools can answer business questions that traditionally were time consuming to resolve. If data mining will be followed by the clinicians instead of their own decision-making, then better and cheaper care would ensure. Data mining techniques have been used to discover various biological, drug discovery, and patient care knowledge and patterns using selected statistical analyses, machine learning, and neural networks methods.

Data mining systems can be categorized according to various criteria, the classification is as follows: Classification of data mining systems according to mining techniques used. This classification is according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database oriented or data warehouse-oriented, etc. It is a technique used to extract valid novel, potential patterns and useful information from complex and huge amount of data set. The two fundamental goals of Data Mining are prediction and description otherwise known as verification model and discovery model. Some

important data mining tasks are association rules, classification rules, outlier analysis and clustering. A cluster is a collection of data objects that are more similar to one another within the same cluster and are dissimilar to the objects in other clusters.

The goal of clustering and cluster analysis is to group and distinguish comparable units and to separate them from differing units. Towards this end, cluster analysis encompasses a wide range of statistical techniques. In cluster analysis, one attempts "to group large numbers of persons, jobs, or objects into smaller numbers of mutually exclusive classes in which the members have similar characteristics. The ultimate dispassionate is to develop clusters whose configurations would be such that each entity in the analysis would be classified into only a single, unique cluster. We proposed a hybrid clustering algorithm for decision support system for records management system in health care records.

Clustering is the most frequently used technique to implement the segmentation operation. Though a lot of data sets dealt in data mining have categorical attributes, most existing clustering algorithms are limited to numeric attributes. The traditional approach is to convert category attributes into binary attributes and to treat the binary attributes as numeric in the clustering algorithms developed for numeric attributes.

The k-means algorithm is well known for its efficiency in clustering large data sets. However, working only on numeric values prohibits it from being used to cluster real world data (business processes dataset) containing categorical values. In this work we present two algorithms which extend the k-means algorithm to categorical domains and domains with mixed numeric and categorical values. The k-modes algorithm uses a simple matching dissimilarity measure to deal with categorical objects, replaces the means of clusters with modes, and uses a frequency-based method to update modes in the clustering process to minimize the clustering cost function. With these extensions the k-modes algorithm enables the clustering of categorical data in a fashion similar to k-means. This measure is often referred to as simple matching.

In this research, we use a modified version of value difference metric, k-representatives algorithm, first introduced by (Stanfill, 1986) and k-means algorithm to proposed hybrid clustering algorithm to cluster large volume of mixed data of patient information. Thus, reducing time wastage and improve service delivery and discover hidden pattern in large volume of medical records, which will help in good decision making in the hospital management System.

2. REVIEW OF RELATED WORKS FOR CLUSTERING ALGORITHM

Modified k-means algorithm for clustering mixed data sets (Ahmad et al. 2007). They also proposed a modified representation for the cluster center. Though similar representations have been used for fuzzy clustering, they used it in a novel manner to compute the distance of an object from a cluster center. The significance of this representation is that it captured cluster characteristics very effectively, since it contained the distribution of all categorical values in a cluster. The results obtained with their algorithm over a number of real-world datasets are highly encouraging. Although their results were encouraging, their methods could not address the discretized numeric valued attributes, which seldom leads to information loss.

A new distance measure based on the weight age, which is automatically generated and applied to incremental clustering algorithm. Their results show that the combination of the two algorithms proved to be more effective in handling mixed datasets consisting of numerical and categorical attributes only. That notwithstanding, the issue of defining the right distance measure and choosing different K-values and threshold are still open problems (Sowjanya et al. 2011)
Unified metric for categorical and numerical attributes in data clustering, in which the attributes are in either one of the three: numerical, categorical and mixed. Their experimental results show

the efficacy of the proposed approach, there is no user-assigned parameter in the proposed algorithm (Yiu-ming et al. 2011).

Presentation of a Fuzzy C-means (FCM) clustering algorithm, which allows one piece of data to belong to two or more clusters. Their new fuzzy k-modes algorithm is effective and better than the other existing k-modes algorithms, their algorithm could not combine the two implementation of fuzzy c-means and fuzzy k-modes for mixture of data items set. Moreover, they could not find out the cluster area and center of cluster. They suggested that the Fuzzy C-mean (FCM) algorithm for numeric data should be implemented and the Fuzzy K-modes algorithm for categorical data be also implemented separately. Then combine the both implementation for mixture of data items that is numeric as well as categorical data items to produce the final results. (Dewangan et al., 2010)

Proposed cluster of mixed numeric and categorical datasets in an efficient manner. They used a clustering algorithm based on similarity weight and filter method paradigm, that works well for data with mixed numeric and categorical features. Although their approach is very efficient in solving any number of dimensions, reduce time complexity and the irregular boundaries of the filter algorithms, the problem is that they could not mix the different clustering datasets with different algorithms (Asadi et al., 2012).

The proposed efficient algorithm that uses the techniques of Divide and Conquer to cluster large datasets. He applied the Squared Euclidean Distance in the measuring the similarity between data points. Their approach is very efficient in identifying the data points and assigning the data points to the best clusters, he could not read the entire data files at once (Ahirwar, 2014).

Implementation of a parallel k-means algorithm to cluster large datasets using agricultural databases. Their improved the time complexity inherent in k-means algorithm when applied on large datasets since it used to be computationally expensive. The performance of the parallel k-means algorithms proved to be better in terms of efficiency and time complexity compared with the normal sequential k-means, the problem is that their processing speed was disturbed with the bandwidth of the network available (Ramesh et al, 2013).

The proposed idea of the K-means clustering algorithm analysis. They offered two methods of improving the K-means clustering algorithm based on improving the initial focal point and secondly, determining the value of K (where K is the number of clusters). Their simulation experiments proved that the improved clustering algorithm was not only stable in the clustering process but reduced or even avoided the impact of the noise data in the data set object. Their experimental results shows an improvement of K-means clustering algorithms are still not solved completely, rather, it requires further attempt and exploration.(Zhang et al., 2013),

New algorithm and applied same to mixed types of dataset, which consisted of numeric and categorical types. He opined that most data in the actual world is always mixed of two types, the numeric and categorical one. The experimental results show work better than K-means even though their input is a mixed dataset. And their experimental results are better in terms of speed and performance than K-means, (Kanjanawattana, 2012).

The proposed combination of Supervised and Unsupervised learning of clustering data without any preliminary assumption on the cluster shape. The experimental results show that the clusters are created correctly when the attributes are petal length verses petal width good accuracy. Thus, their results were encouraging and trustable. (Kolbe et al., 2010)

3. DATA MINING MODEL

Data Mining is the process of extracting knowledge hidden from large volumes of raw data. The knowledge must be new, not obvious, and one must be able to use it. Data mining has been

defined as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. It is "the science of extracting useful information from large databases".
Data mining is one of the tasks in the process of knowledge discovery from the database. Shows the process of knowledge discovery. The steps involved in Knowledge discovery are:

i. **Data Selection**: The data relevant to the analysis is decided and retrieved from the various data locations.
ii. **Data Preprocessing**: In this stage the process of data cleaning and data integration is done.
    a) Data Cleaning: It is also known as data cleansing; in this phase noise data and irrelevant data are removed from the collected data.
    b) Data Integration: In this stage, multiple data sources, often heterogeneous, are combined in a common source.
iii. **Data Transformation**: In this phase the selected data is transformed into forms appropriate for the mining procedure.
iv. **Data Mining**: It is the crucial step in which clever techniques are applied to extract potentially useful patterns. The decision is made about the data mining technique to be used.
v. **Interpretation and Evaluation**: In this step, interesting patterns representing knowledge are identified based on given measures.

The discovered knowledge is visually presented to the user.This essential step uses visualization techniques to help users understand.

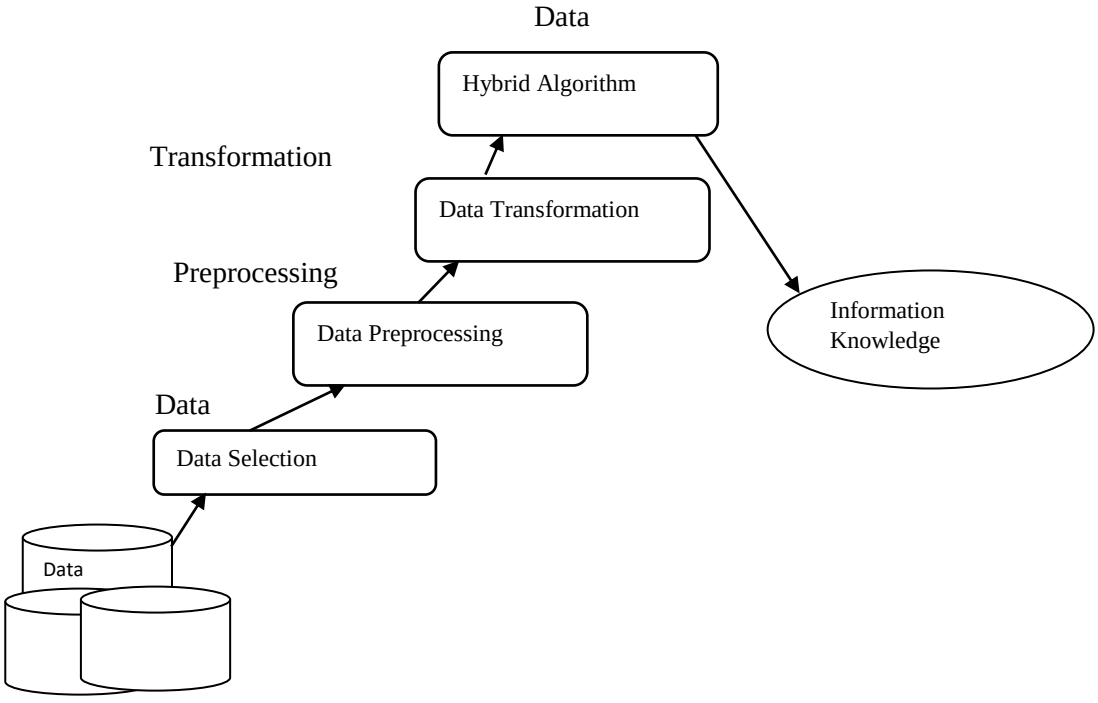

Figure 1: Data Mining Model

3.1 Predictive Modeling

Predictive modeling uses statistics to predict outcomes. Most often the event one wants to predict is in the future, but predictive modelling can be applied to any type of unknown event, regardless

of when it occurred. For example, predictive models are often used to detect crimes and identify suspects, after the crime has taken place. Raw health data can be in various formats, including narrative/textual data (e.g., history of a present illness), numerical measurements (e.g., laboratory results, vital signs, and measurements), recorded signals (e.g., electrocardiograms), and pictures (e.g., radiologic images). Numerical measurements and recorded signals were the format used most in the reviewed articles. Three main approaches were used for extracting predictors from raw health data. First, for some variables, including age and gender, the exact/coded value was used as a predictor. Second, in articles employing recorded signals and/or longitudinal numerical measurements, numeric variables, including wavelet coefficients, minimums, maximums, means, and variances, were extracted from within particular time windows. Third, three studies employed the term frequency-inverse document frequency (TF-IDF) technique from text mining to produce predictors for their model.

Although predictor extraction affects the performance of the model, one of the challenging tasks in patient similarity-based predictive modeling is identifying the most relevant and important patient characteristics for patient similarity assessment. Patient similarity assessment is generally defined as investigating the similarity of patients' data in terms of their symptoms, comorbidities, demographics, and treatments, but there is no predefined list of predictors to be considered. Most of the studies proposed an arbitrary list of predictors or limited their work to the available predictors, but selected predictors must be representative of patient's condition in each particular application.

3.2 Concept of Classification

One of the important task of data mining is data classification which is the process of finding a valuable set of models that are self-descriptive and distinguishable data classes or concept to predict the set classes with an unknown class label.

For example in the transportation network all highway with the same structural and behavioral properties can be classified as class highway. From the application point of view, classification helps in credit approval, product marketing and medical diagnosis. So many techniques such as decision tree, neural network, nearest neighbor method and rough set-base methods enable the creation of classification model. Classification techniques are also applied to analyze various signals and their relationships with particular diseases or symptoms (Demighab 2015).

Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects. For Example, after starting a credit policy, the video store managers could analyze the customers? Behaviors' vis-à-vis their credit, and label accordingly the customers who received credits with three possible label safe, risky and very risky. The classification analysis would generate a model that could be used to either accept or reject credit requests in the future (Yanthy W. etal., 2009). In case of classification following terms are important:

- Accuracy: It gives a measure for the overall accuracy of the classifier:
- Accuracy (%) = number of correctly classified instances * 100 / Number of instances.
- Precision and recall: With respect to classifier:
- Precision(X) = Number of correctly classified instances of class X / Number of instances in class X.
- Recall(X) = Number of correctly classified instances of class X / Number of instances in class X.

"Classification "is the most frequently used data mining function with a predominance of the implementation of Bayesian classifiers, networks, and SVMS (Support Vector Machines).

3.3 Concept of Clustering

Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects. A cluster is a collection of data objects that are similar to one another with the same cluster and are dissimilar to the objects in other clusters. A cluster of data objects can be treated collectively as one group in many applications. Clustering is a form of learning by observation rather than learning by examples. Cluster analysis is an important human activity in which we indulge since childhood when we learn to distinguish between animals and plants etc. by continuously improving subconscious clustering schemes. It is widely used in numerous applications, including pattern recognition, data analysis, image processing, and market research (Han et al., 2006), Clustering is a very important application area but widely interdisciplinary in nature, that makes it very difficult to define its scope. It is used in several research communities to describe methods for grouping of unlabeled data. Now, these communities have different terminologies and assumptions for the components of the clustering process and the contexts in which clustering are used. Cluster analysis has been studied extensively for years, focusing mainly on distance-based cluster analysis. Many clustering tools were made based on k-means, k-medoids, and some of the methods were incorporated in many statistical analysis software packages (Perumal, 2011). The major clustering steps are preprocessing and feature selection, similarity measure, clustering algorithms, result validation, and result interpretation.

Clustering can be used in designing a triage system. Triage helps to classify patients at emergency departments to make the most effective use of resources distributed. What is more important is that accuracy in carrying out triage matters greatly in terms of medical quality, patient satisfaction and life security. The study is made on medical management and nursing, with the knowledge of the administrative head at the Emergency Department, in the hope to effectively improve consistency of triage with the combination of data mining theories and practice. The purposes are as follows:

i. Based on information management, the information system is applied in triage of the Emergency Department to generate patients' data.
ii. Exploration of correlation between triage and abnormal diagnosis; cluster analysis conducted on variables with clinical meanings.
iii. Establishing triage abnormal diagnosis clusters with hierarchical clustering (Ward's method) and partitioning clustering (K-means algorithm); obtaining correlation law of abnormal diagnosis with decision trees.
iv. Improving consistency of triage with data mining; offering quantified and scientific rules for triage decision-making in the hope of serving as a foundation for future researchers and clinical examination.
v.

3.1 Hybrid Clustering Algorithms

It is straightforward to integrate the k-means and K-Representatives algorithms into the Hybrid clustering algorithm that is used to cluster the mixed-type objects. The Hybrid clustering algorithms is practically more useful because frequently encountered objects in real world databases are mixed-type objects.The dissimilarity between two mixed-type objects X and Y, which are described by attributes A^r_1, A^r_2,... ;A^r_p,… A^c_{p+1},A^c_m, can be measured by

$$d_2(X,Y) = \sum_{j=1}^{p}(x_j - y_j)^2 + y\sum_{j=p+1}^{p}\partial(x_j,y_j) \qquad 1$$

Where the first term is the squared Euclidean distance measure on the numeric attributes and the second term is the simple matching dissimilarity measure on the categorical attributes. The weight

α is used to avoid favoring either type of attribute. The influence of α in the clustering process is discussed in (Huang, 1997a) Using (9) for mixed-type objects.

The k-means algorithm (Macqueen, 1967) was introduced to cluster numerical datasets. It minimizes the objective function J for hard (non-fuzzy) k-partition of a dataset into k clusters (Bezdek, 1980): Formula2.1Here uim is an element of the partition matrix. The condition uim=1 means that the record Xi is assigned to cluster m with prototype (center) Qm.

4. DISCUSSION AND RESULT

This improved model for medical records classification system help medical doctors organize previous medical prescriptions on a particular ill health into similar drugs components to reduce information overload and improve access to the most effective drug in treating illness. The application of clustering as a data mining technique in solving this problem, help to classify patient records and separate them for easy extraction.

5. CONCLUSION

In this research we have made modifications in the k-means method while apply the method to the problem of clustering patient dataset for decision making in medical records. The clustering performance of the proposed algorithm is demonstrated with real world dataset of inpatient and outpatient treatment medical records in the hospital. The experimental results have shown that the proposed clustering algorithm gives pure clustering results, and is accurate for clustering mixed datasets.

REFERENCE

Ahmad A. and L. Dey, (2007), A k-mean clustering algorithm for mixed numeric and categorical data', Data and Knowledge Engineering Elsevier Publication, vol. 63, pp 503-527.

Ahirwar, D. K., Saxena, S. K., and Sisodia, M. S. (2012), Anomaly detection by naive Bayes & RBF network. International Journal of Advanced Research in Computer Science and Electronics Engineering 1, 1, 14{18. Academic Press,}

Demigha[b], S. (2015) "Data Mining for Breast Cancer Screening," *In the 10th IEEE International Conference on Computer Science & Education, IEEE ICCSE*, 59–63, Fitzwilliam College, Cambridge University, Cambridge, UK.

Dewangan, D., Kumar, M., and Qureshi, M. (2010) "Power system transient stability analysis based on interval type-2 fuzzy logic controller and genetic algorithms", *International Journal of Innovative Science Engineering and Technology*, Vol. 1, No. 4, pp. 103-120

Fayyad, U., Shapiro, G.P. and Smyth, P. (1996) "Knowledge Discovery and Data Mining: Towards a Unifying Framework," *In Proc. 2nd International Conference on Knowledge Discovery and Data Mining. AAAI Press*, 82–8.

Han, R. Ng (2006), Efficient and effective clustering method for spatial data mining, in: Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile, pp. 144–155.

Huang, Z. (1997) 'Extensions to the k-means algorithm for clustering large data sets with categorical values', *Data Mining and Knowledge Discovery*, Vol. 2, No. 3, pp.283–304.

Kanjanawattana, M. R. Anderberg (2012). *Cluster Analysis for Application*. Academic Press,

Kolhe, Dr. Ch. G.V.N. Prasad, HanumanthaRao, DepaPratima and B.N. Alekhya, (2010), Unsupervised Learning Algorithms to Identify the Dense Cluster in Large Datasets", International Journal of Computer Science and Telecommunications [Volume 2, Issue 4,]

Kovalerchuk, B., Vityaev, E., and Ruiz, J. F. (2001) "Consistent and Complete Data and 'Expert' Mining in Medicine," In Cios, K. J. (Ed.), *Medical Data Mining and Knowledge Discovery*, New York, USA: Physica- Verlag

MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281–297.* MR 0214227. Zbl 0214.46201. Retrieved 2009-04-07.

Perumal (2011) *A Clustering Technique for Summarizing Multivariate Data*, Behavioural Science, **12**, pp. 153-155.

Ramesh D, B Vishnu Vardhan, (2013). Data Mining Techniques and Applications to Agricultural Yield IJCATM :www.ijcaonline.org

Saad ,F. H., B. de la Iglesia, and G. D. Bell,(2006) ― A Comparison of Two Document Clustering Approaches for Clustering Medical Documents, Proceedings of the 2006 International Conference on Data Mining (DMIN-06).

Sowjanya, A.M. and M. Shah, (2011). Cluster feature based incremental clustering approach (CFICA) for numerical data. IJCSNS Int. J. Comp. Sci. Network Security, 10.

Yanthy W., Skiya T., Yamaguchi K., (2009), Mining Interesting Rules by Association and Classification Algorithms in the proceeding of International Conference on frontier of Computer Science and Technology, pp.177182.

Yiu Ming Cheung and Hong Jia in "Categorical and numerical data clustering based on a unified similarity metric without knowing cluster number" in Pattern Recognition 46 (2013) 2228–2238, Elsevier (2011).

Zhang, Chen; Xia Shixiong (2001). "K-means Clustering Algorithm with improved categorical values. Data Mining and Knowledge Discovery, 2(3),

Wang, H. & Wang, S. (2008) "A knowledge management approach to data mining Processor business intelligence. Industrial Management &Data Systems, 108(5), 622-634.https://en.m.wikipedia.org/wiki/Predictive modelling