Volume 3, Issue 12, December 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Scrutiny Prediction of Quad-Clustering Algorithms of Healthcare Application

D.Nithya
Research Scholor,
Psg College of Arts and Science, India

Dr.R.Shanmugavadivu
Assistant Professor,
Psg College of Arts and Science, India

Abstract: ***Health care industry is considered one of the largest industries in the world. The health care industry is same as the medical industries having the large amounts of health related and medical related data. It includes thousands of hospitals, clinics and other types of facilities which provide primary, secondary & tertiary levels of care. A health care provider is an institution or person that provides preventive, curative, promotional or rehabilitative health care services in a systematic way to individuals, families or community. This paper deals with a preventive measure for determining whether a person is fit or unfit based on his/her historical and real time data by applying clustering algorithms viz. Simple K-means, D-stream, K-means++ and global K-means. These clustering algorithms are applied on patient's biomedical historical database. Here, Simple K-means, D-stream, Global K-means and K-means++ algorithms are compared and evaluated successfully. The Prediction of health status is very sensitive job. In that, K-Means++ algorithm's prediction accuracy is better when comparing other clustering algorithms.***

Keywords: Healthcare –Prediction - Clustering – Data Modeling – Data Preprocessing.

I. INTRODUCTION

Data mining has been used intensively and extensively by many organizations. In healthcare, data mining is becoming increasingly popular, if not increasingly essential. Data mining applications can greatly benefit all parties involved in the healthcare industry. Data mining can help healthcare insurers detect fraud and abuse, healthcare organizations make customer relationship management decisions, physicians identify effective treatments and best practices, and patients receive better and more affordable healthcare services. The huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. Data mining provides the methodology and technology to transform these mounds of data into useful information for decision making. The purpose of this paper is to cluster the healthcare dataset efficiently using clustering algorithms. Here, Simple K-means, D-stream, Global K-means and K-means++ algorithms are compared and evaluated successfully. In that, K-Means++ algorithm's prediction accuracy is better when comparing other clustering algorithms. The proposed system uses historical biomedical data which is very useful for prediction of current health status of a patient by using clustering algorithms like K-means, D-stream, and Global K-Means. The major goal is to cluster the patient's records into different groups with respect to the test report attributes which may help to diagnose the patient's disease in proficient way.

II. RELATED WORKS

Health Informatics is a rapidly growing field that is concerned with applying Computer Science and Information Technology to medical and health data. With the aging population on the rise in developed countries and the increasing cost of healthcare, governments and large health organizations are becoming very interested in the potential of Health Informatics to save time, money, and human lives. The application of data clustering technique for fast retrieval of relevant information from the medical databases lends itself into many different perspectives. Intelligent Mobile Health Monitoring System (IMHMS), [6] Author proposed the system which can provide medical feedback to the patients through mobile devices based on the biomedical and environmental data collected by deployed sensors. The system uses the Wearable Wireless Body/Personal Area Network for collecting data from patients, mining the data, intelligently predicts patient's health status and provides feedback to patients through their mobile devices. Real-Time analysis of physiological data to support medical applications [8], proposed a flexible framework to perform real-time analysis of physiological data and to evaluate people's health conditions. Patient or disease-specific models are built by means of data mining techniques. Performance of Clustering Algorithms in Healthcare Database [9], proposed a framework where they used the heart attack prediction data for finding the performance of clustering algorithm. Final result shows the performance of classifier algorithm using prediction accuracy and the visualization of cluster assignments shows the

relation between the error and the attributes. The comparison result shows that, the make density based clusters having the highest prediction Accuracy.

III. PROBLEM DEFINITION

The main purpose of the data mining is to find out the hidden knowledge from the data base. In health care industry, the data may have some unwanted data, missing values and noisy data. Those unwanted data will be removed by using preprocessing techniques in data mining. Preprocessing is the process of removing noise, redundant data and irrelevant data. After the preprocessing the data will be used for some useful purpose. In recent years diverse approaches are proposed to overcome the challenges of storing and processing of fast and continuous streams of data.

Data stream can be conceived as a continuous and changing sequence of data that continuously arrive at a system to store or process. Data streams can be produced in many fields, it is crucial to modify mining techniques to fit data streams. These data stream mining can be used to form the clusters of medical health data. The K-means clustering algorithm is incompetent to find clusters of arbitrary shapes and cannot handle outliers. To address these issues, D Stream, a framework for clustering stream data using a density-based approach. The algorithm uses an online component which maps each input data record into a grid and an offline component which computes the grid density and clusters the grids based on the density.

The k-means algorithm is a local search procedure and it is well known that it suffers from the serious drawback that its performance heavily depends on the initial starting conditions. To treat this problem several other techniques have been developed that are based on stochastic global optimization methods. However, it must be noted that these techniques have not gained wide acceptance and in many practical applications the clustering method that is used is the k-means algorithm with multiple restarts.

IV. MAJOR CONTRIBUTION

The following evaluation techniques are the main contributions of the work, which has been screened out in this paper:

A. *Data Set Collection* - The data must be collected in a resourceful manner. The data set contains 7 attributes, SpO2, ABPsys, ABPdias, HR, heredity, obesity, cigarette smoking. These attributes are the risk factors that can help in predicting the patient's health status. Attributes such as SpO2, ABPsys, ABPdias, HR can be collected form MIMIC database and the other attributes are influenced by the person's behavior.

B. *Model Building* - In model building phase features of the available data will be extracted and then clustering algorithm will be applied on extracted features.

- Feature Extraction

  In Feature Extraction, for each physiological signal x among the X monitored vital signs, the following features are extracted.

  1. *Offset* - The offset feature measures the difference between the current value x(t) and the moving average (i.e., mean value over the time window). It aims at evaluating the difference between the current value and the average conditions in the recent past.
  2. *Slope* - The slope function evaluates the rate of the signal change. Hence, it assesses short-term trends, where abrupt variations may affect the patient's health.
  3. *Dist* - The dist feature measures the drift of the current signal measurement from a given normality range. It is zero when the measurement is inside the normality range.

- Risk Components

  In Risk Components, the signal features contribute to the computation of the following risk components

  1. *Sharp Changes* - The z1 component aims at measuring the health risk deriving from sharp changes in the signal (e.g., quick changes in the blood pressure may cause fainting)
  2. *Long-Term Trends* - The z2 component measures the risk deriving from the h weighted offset over the time window. While z1 focuses on quick changes, z2 evaluates long-term trends, as it is offset-based.
  3. *Distance From Normal Behavior* - The z3 component assesses the risk level given by the distance of the signal from the normality range. A patient with an instantaneous measurement outside the range may not be critical, but her/his persistence in such conditions contributes to the risk level

C. *Data Preprocessing*- If there is much irrelevant and redundant information present or noisy and unreliable data, and then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. Data pre-processing includes cleaning, normalization, transformation, feature extraction and selection, etc.

D. *Cluster Formation* - The proposed flow of the system uses four algorithms K-means and D-stream, Global k-means and K-means++. The comparison between these clustering algorithms will be performed using the above described attributes.

V. EXPERIMENTAL RESULTS

The objective is to cluster the patient's records into different groups with respect to the test report attributes which may help the clinicians to diagnose the patient's disease in efficient manner. The flow of this research work is depicted in Fig 1. The specified framework will perform clustering of dataset available from medical database in effective manner.

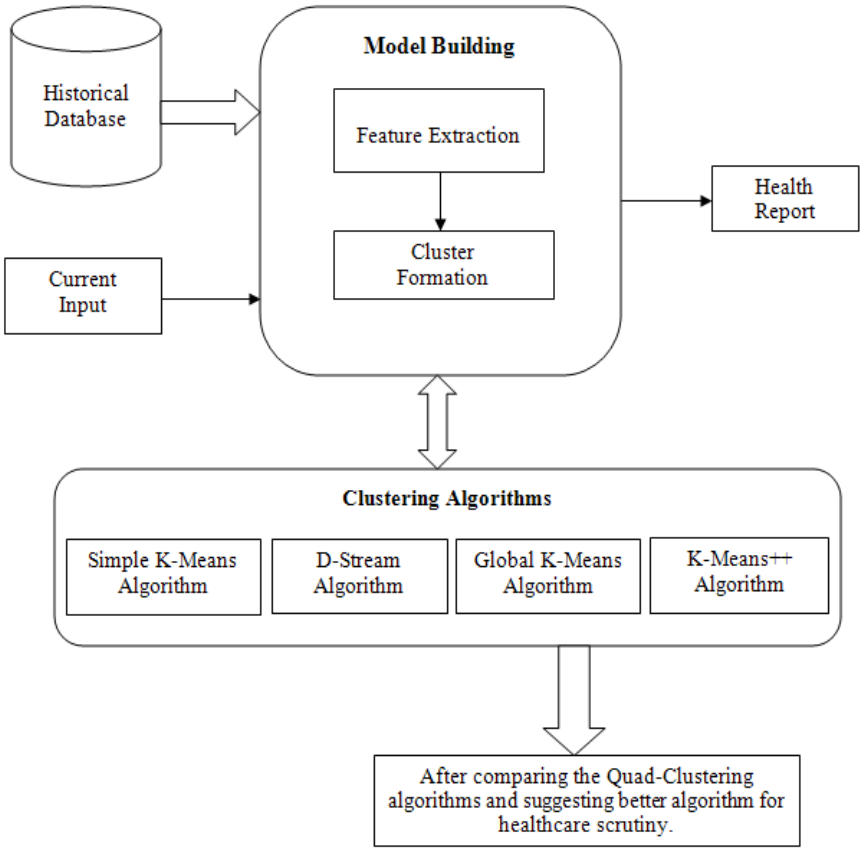

Fig.1 Flow of the System

The global k-means algorithm is an incremental approach to clustering that dynamically adds one cluster center at a time through a deterministic global search procedure consisting of N executions of the k-means algorithm from suitable initial positions. The basic idea underlying this algorithm is that an optimal solution for a clustering problem with M clusters can be obtained using a series of local searches. The cluster formation of global k-means algorithm is screened out in Fig.2.

The cluster formation using global k-means algorithm is shown in Fig 6.4. The red color represents one individual cluster and black and green color represents other individual clusters that were predicted. The cluster formation of k-means++ algorithm is demonstrated in the Fig.3. The classification of clusters are efficiently predicted using k-means++ clustering algorithm. In this cluster, each color represents individual clusters.

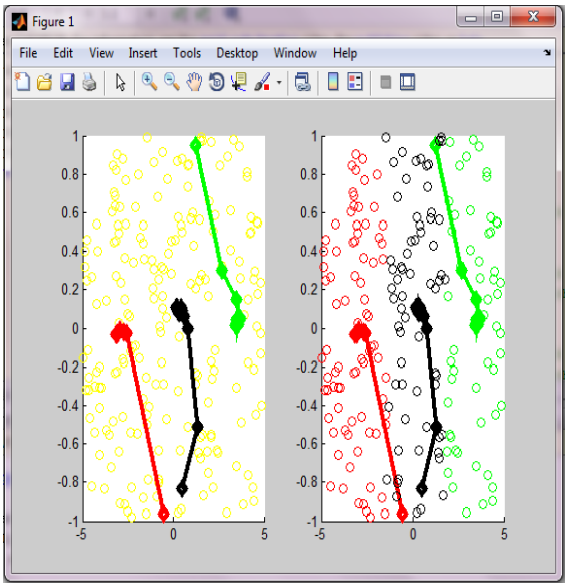

Fig.2 Cluster Formation of Global K-Means Algorithm

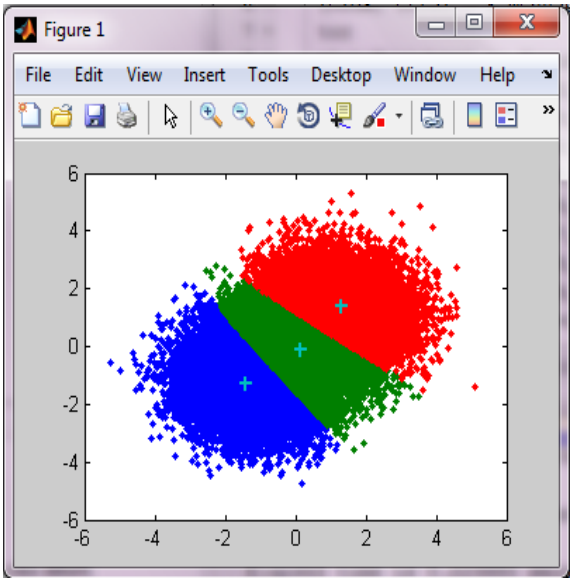

Fig.3 Cluster Formation of K-Means++ Algorithm

The cluster formation of each color represents each cluster which is implemented using heart dataset in Matlab. It works better while comparing with other clustering algorithms. The performance of k-means++ is comparably high.

VI. PERFORMANCE EVALUATION

K-Means++ is parameter free and proves to give more accurate results than K-means and D-Stream and Global k-means algorithm, when used for cluster formation of historical biomedical data. Prediction of health status is very sensitive job, K-Means++ will perform better here, as it supports arbitrary cluster formation which is not supported by other algorithms.

TABLE I
PERFORMANCE EVALUATION

Cluster Category	Cluster Algorithm		Measures
Clusters	**Correctly Classified Instance**	**Incorrectly Classified Instance**	**Prediction Accuracy**
Simple K-Means	89	18	83
D-Stream	94	13	87
Global K-Means	96	10	90
K-Means++	99	7	93

While comparing the results of quad clustering algorithm, the prediction of accuracy becomes high. The accuracy of correctly classified instance increases and incorrectly classified instance decreases while cluster formation. By analyzing the cluster formation, k-means++ algorithm perform better than other three algorithms. The performance of cluster formation is efficiently implemented and screened out in the Table I.

VII. CONCLUSIONS

Healthcare is the most important factor affecting human life. Due to heavy work load, personal healthcare is not a possible thing. All Several health care projects are in full swing in different universities and institutions, with the objective of providing more and more assistance to the elderly. The application of data clustering technique for fast retrieval of relevant information from the medical databases lends itself into many different perspectives. The basic idea deals with the dissertation, finding the accuracy of the clustering algorithms in healthcare application. By calculating the performance of simple K-Means Algorithm, D-Stream Algorithm, Global K-Means Algorithm and K-Means++ Algorithm and uncovering the accuracy of the clustering algorithm are discussed in previous chapters. The developed applications were tested using small number of users. While testing the performance of the work, it found to be good. In future, the application can be implemented in large healthcare databases.

REFERENCES

[1] American Medical Informatics Association, http://www.amia.org/informatics/.
[2] Canada's Health Informatics Association, http://www.coachorg.com/.
[3] National Library of Medicine,http://www.nlm.nih.gov/tsd/acquisitions/cdm/subjects58.html.
[4] Nuria Oliver , Fernando Flores-Mangas, "HealthGear: A Real-time Wearable System for Monitoring and Analyzing Physiological signals" in proceeding BSN'06 Proceedings of the International Workshop on Wearable and Implantable Body Sensor Networks, 2006
[5] Vesanto J., Himberg J., Alhoniemi E., Parhankangas J. (1999). Self-organising map in Matlab: the SOM Toolbox. Proceedings of the Matlab DSP Conference, Finland, 35–40.
[6] RifatShahriyar, Md. Faizul Bari, GourabKundu, Sheikh IqbalAhamed and Md. Mustofa Akbar 5,"Intelligent Mobile Health Monitoring System(IMHMS)", International Journal of Control and Automation, vol 2,no.3, Sept 2009, pp 13-27.
[7] Illhoi Yoo, Patricia Alafaireet, Miroslav Marinov, Keila Pena-Hernandez, Rajitha Gopidi, Jia-Fu Chang, Lei Hua, "Data Mining in Healthcare and Biomedicine: A Survey of the Literature", Journal of Medical Systems August 2012, Volume 36, Issue 4, pp 2431-2448
[8] Daniele Apiletti, Elena Baralis, Member, IEEE, Giulia Bruno, and Tania Cerquitelli, "Real-Time Analysis of Physiological Data to Support Medical Applications", IEEE Transactions On Information Technology In Biomedicine, Vol. 13, No. 3, May 2009.
[9] P.Santhi, V.Murali Bhaskaran Computer Science & Engineering Department Paavai Engineering College, "Performance of Clustering Algorithms in Healthcare Database", International Journal for Advances in Computer Science, Volume 2, Issue 1 March 2010
[10] The MIMIC database on PhysioBank (2007, Oct.) [Online]. Available: http://www.physionet.org/physiobank/database/mimicdb

