Guideline of Data Mining Technique in Healthcare Application.

Mr.Swapnil D. Raut, Prof. Avinash Wadhe.

Abstract— **In today's world, healthcare is the most important factor affecting human life. The management of health care database is the most challenging subject of this era. For this the data mining has been used intensively and extensively by many organizations which are related to healthcare. In healthcare, the need of data mining is increasing rapidly. There are various algorithms of data mining used on healthcare databases. This paper presents an overview on different types of data mining algorithms like K-means and D-stream algorithm, also a comparative study on these data mining algorithms. From that study we found the effectiveness and limitations of these data mining algorithms.**

Index Terms— k-means, d-stream, cluster, feature Extraction.

I. INTRODUCTION

The successful application of data mining in highly visible fields like e-business, marketing and retail has led to its application in other industries and sectors. This seminar intends to provide a survey of current techniques of knowledge discovery in databases using data mining techniques that are in use today in medical research and public health. We also discuss some critical issues and challenges associated with the application of data mining in the profession of health and the medical practice in general. An overview of the current research being carried out using the data mining techniques for the diagnosis and prognosis of various diseases. Data mining is defined as "a process of nontrivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database" by Fayyad [1]. The data mining is the process of extracting or mining the knowledge from the large amounts of data, database or any other data base repositories. Healthcare databases have a huge amount of data but however, there is a lack of effective analysis tools to discover the hidden knowledge. Appropriate computer- based information and/or decision support systems can help physicians in their work.

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. Data stream can be conceived as a continuous and changing sequence of data that continuously arrive at a system to store or process. Whereby data streams can be produced in many fields, it is crucial to modify mining techniques to fit data streams. Data stream mining has many applications and is a hot research area [2]. Data stream mining is the extraction of structures of knowledge that are represented in the case of models and patterns of infinite streams of information. These data stream mining can be used to form the clusters of medical health data. Various data mining techniques are available for predicting diseases namely Classification, Clustering, Association rules and Regressions. From previous research it is seen that K-means algorithm and density based clustering algorithm are two most important algorithms in data mining.

The K-means clustering algorithm is incompetent to find clusters of arbitrary shapes and cannot handle outliers. Further, they require the knowledge of k and user-specified time window. 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 algorithm adopts a density decaying technique to capture the dynamic changes of a data stream. Exploiting the intricate relationships between the decay factor, data density and cluster structure, our algorithm can efficiently and effectively generate and adjust the clusters in real time. Further, a theoretically sound technique is developed to detect and remove sporadic grids mapped to by outliers in order to dramatically improve the space and time efficiency of the system. The technique makes high-speed data stream clustering feasible without degrading the clustering quality. The experimental results show that our algorithm has superior quality and efficiency, can find clusters of arbitrary shapes, and can accurately recognize the evolving behaviors of real-time data streams [3].

II. LITERATURE REVIEW

Fast retrieval of relevant information from the databases has always been a significant issue. The application of data clustering technique for similarity searching in medical databases lends itself into many different perspectives. In Vikram Singh, Sapna Nagpal *et.al* [5] presents an experiment based on clustering data mining technique to discover hidden patterns in the dataset of liver disorder patients. The system uses the SOM network's internal parameters and k-means algorithm for finding out patterns in the dataset. The research has shown that meaningful results can be discovered from clustering techniques by letting a domain expert specify the input constraints to the algorithm.

P.Santhi *et al.* [8], proposed a framework where they used the heart attack prediction data for finding the performance of clustering algorithm. In 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.

RifatShahriyar *et al.* [6] 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.

Daniele Apiletti *et al.* [7], 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. Models are exploited to perform real time classification of physiological signals and continuously assess a person’s health conditions. The proposed framework allows both instantaneous evaluation and stream analysis over a sliding time window for physiological data. But dynamic behavior of the physiological signals is not analyzed also the framework is not suitable for ECG type of signals.

In Shing-Hong Liu, *et.al* [4] has built an automatic disease classification system using pulse waveform analysis, based on a Fuzzy c-means (FCM) clustering algorithm. A self designed three-axis mechanism was used to detect the optimal site to accurately measure the pressure pulse waveform (PPW). A fuzzy c-means algorithm was used to identify myocardial ischemia symptoms in 35 elderly subjects with the PPW of the radial artery.

III. METHOD

In this framework that will perform clustering of dataset available from medical database effective manner. The flow of the system is depicted in Figure 3.1.

The target 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 and The evaluation steps are the following-

3.1 Data set collection:

In the data set contains some attribute like 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 [8] and the other attributes are influenced by the person’s behavior. These all attributes values are discrete in nature .The dataset will be in preprocessed format.

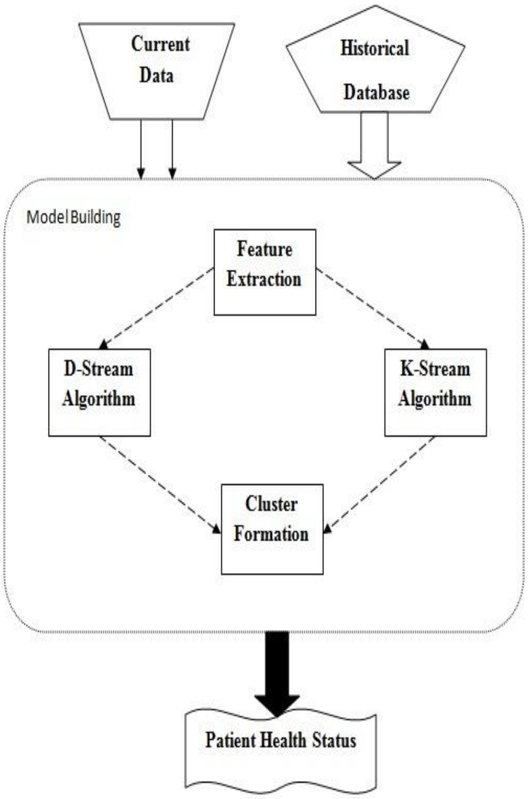

Figure 3.1: Flow of the system

3.2 Model Building

In model building phase features of the available data will be extracted and then clustering algorithm will be applied on extracted features.

3.2.1. Feature Extraction

For each physiological signal x among the X monitored vital signs, we extract the following features [9].

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.

3.2.2. 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 From above risk components, risk functions and global risk components will be calculated. These values will be further used in clustering algorithms as an input for cluster formation.

3.2.3 Cluster formation

The proposed flow of the system uses two algorithms K-means and D-stream. The comparison between two clustering algorithms will be performed using the above described attributes.

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 but the inter cluster similarity is low. Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the cluster's centroid or center of gravity [7].

The *k*-means algorithm proceeds as follows,

- First, it randomly selects *k* of the objects, each of which initially represents a cluster mean or center.
- Then, 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 square-error criterion is used, defined as

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2,$$

where *E* is the sum of the square error for all objects in the data set; ***p*** is the point in space representing a given object; and ***mi*** is the mean of cluster *Ci* (both ***p*** and ***mi*** are multidimensional). In other words, for each object in each cluster, the distance from the object to its cluster center is squared, and the distances are summed. This criterion tries to make the resulting *k* clusters as compact and as separate as possible. *k*-means procedure is summarized in above.

The **D-stream** algorithm is explained as follows [11]

```
1. procedure D-Stream
2. Tc = 0;
3. Initialize an empty hash table grid list;
4. while data stream is active do
5.     read record x = (x1, x2, · · · , xd);
6.     determine the density grid g that contains x;
7.     if(g not in grid list) insert g to grid list;
8.     update the characteristic vector of g;
9.     if tc == gap then
10.        call initial clustering(grid list);
11.    end if
12.    if tc mod gap == 0 then
13.        detect and remove sporadic grids from grid list;
14.        call adjust clustering(grid list);
15.    end if
16.    tc = tc + 1;
17. end while
18. end procedure
```

Figure 3.2: The overall process of D-Stream.

The overall architecture of D-Stream, which assumes a discrete time step model, where the time stamp is labeled by integers 0, 1, 2, · · · , n, D-Stream has an online component and an offline component. The overall algorithm is outlined in Figure 3.2.

For a data stream, at each time step, the online component of D-Stream continuously reads a new data record, place the multi-dimensional data into a corresponding discretized density grid in the multi-dimensional space, and update the characteristic vector of the density grid (Lines 5-8 of Figure 3.2). The density grid and characteristic vector are to be described in detail later. The offline component dynamically adjusts the clusters every gap time steps, where gap is an integer parameter. After the first gap, the algorithm generates the initial cluster (Lines 9-11). Then, the algorithm periodically removes sporadic grids and regulates the clusters(Lines12-15).

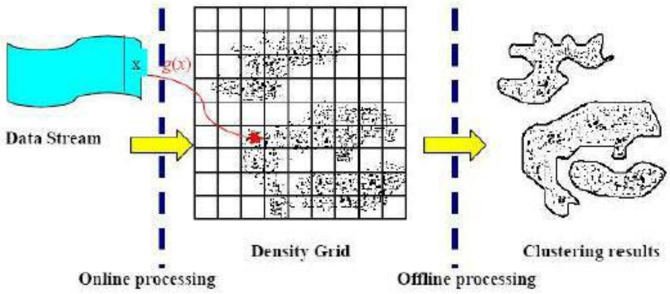

Figure 3.3: Illustration of the use of density grid.

For a data stream, at each time step, the online component of D-Stream continuously reads a new data record, place the multi-dimensional data into a corresponding discretized density grid in the multi-dimensional space, and update the characteristic vector of the density grid (Lines 5-8 of Figure 3.2). The density grid and characteristic vector are to be described in detail later. The offline component dynamically adjusts the clusters every gap time steps, where gap is an integer parameter. After the first gap, the algorithm generates the initial cluster (Lines 9-11). Then, the algorithm periodically removes sporadic grids and regulates the clusters (Lines 12-15).

D-Stream partitions the multi-dimensional data space into many density grids and forms clusters of these grids. This concept is schematically illustrated in Figure 3.3. The input data has d dimensions, and each input data record is defined within the space

$$S = S_1 \times S_2 \times \cdots \times S_d, \quad \text{…... (1)}$$

where Si is the definition space for the ith dimension. In D-Stream, we partition the d−dimensional space S into density grids. Suppose for each dimension, its space Si, I = 1,.. d is divided into pi partitions as

$$S_i = S_{i,1} \bigcup S_{i,2} \bigcup \cdots \bigcup S_{i,p_i},$$

-----(2)

then the data space S is partitioned into N density grids. For a density grid g that is composed of

$$S_{1,j_1} \times S_{2,j_2} \cdots \times S_{d,j_d},\ j_i = 1, \ldots, p_i,$$

we denote it as

$$|g = (j_1, j_2, \cdots, j_d). \qquad \text{--------------(3)}$$

A data record x = (x1, x2, · · · , xd) can be mapped to a density grid g(x) as follows:

$$g(x) = (j_1, j_2, \cdots, j_d)$$ where xi 2 Si,ji .

$$D(x,t) = \lambda^{t-T(x)} = \lambda^{t-t_c}, \qquad \text{----------(4)}$$

where λϵ (0, 1) is a constant called the decay factor. Definition (Grid Density) For a grid g, at a given time t, let E(g, t) be the set of data records that are map to g at or before time t, its density D(g, t) is defined as the sum of the density coefficients of all data records that mapped to g. Namely, the density of g at t is:

$$D(g,t) = \sum_{x \in E(g,t)} D(x,t).$$

```
procedure initial clustering (grid list)
  update the density of all grids in grid list;
  assign each dense grid to a distinct cluster;
  label all other grids as NO CLASS;
  repeat
    for each cluster c
      for each outside grid g of c
        for each neighboring grid h of g
          if (h belongs to cluster c′)
            if (|c| > |c′|) label all grids in c′ as in c;
            else label all grids in c as in c′;
          else if (h is transitional) label h as in c;
  until no change in the cluster labels can be made
end procedure
```

Figure 3.4: The procedure for initial clustering.

```
procedure adjust clustering (grid list)
  update the density of all grids in grid list;
  for each grid g whose attribute (dense/sparse/transitional) is changed since last call to adjust clustering()
    if (g is a sparse grid)
      delete g from its cluster c, label g as NO CLASS;
      if (c becomes unconnected) split c into two clusters;
    else if (g is a dense grid)
      among all neighboring grids of g, find out the grid h whose cluster ch has the largest size;
      if (h is a dense grid)
        if (g is labeled as NO CLASS) label g as in ch;
        else if (g is in cluster c and |c| > |ch|)
          label all grids in ch as in c;
        else if (g is in cluster c and |c| <= |ch|)
          label all grids in c as in ch;
      else if (h is a transitional grid)
        if ((g is NO CLASS) and (h is an outside grid if g is added to ch)) label g as in ch;
        else if (g is in cluster c and |c| >=|ch|)
          move h from cluster ch to c;
    else if (g is a transitional grid)
      among neighboring clusters of g, find the largest one c′ satisfying that g is an outside grid if added to it;
      label g as in c′;
  end for
end procedure
```

Figure 3.5: The procedure for dynamically adjusting clusters.

The calculated values of z1, z2, z3 components will be applied as an input for both the clustering algorithms to form the clusters based on their risk level.

3.3 Patient's Health status

Using clustering algorithm we form the clusters for attributes stated above. And then for patient's current input we predict patient's health status i.e. patient is fit or unfit.

IV. THEORETICAL ANALYSIS

The K-means & D-means algorithms used for formation of clusters on medical database. The data will be collected from the data set. The data set contains the number of instances and the number of attributes. The attributes can be like age, sex, Blood Pressure, Cholesterol, Chest Pain and etc. The performance of these algorithms will be computed by using correctly predicted instance. [3]

Performance Accuracy= correctly predicted Instance / Total Number of Instance

Cluster Category	Cluster Algorithm	Measures		
		Correctly Classified Instance	In correctly Classified Instance	Prediction Accuracy
Clusters	Simple K-means	89	18	83.18
	D-stream	94	13	87.85

Table 4.1: Performance of clustering algorithm

From above formula we can observed that the performance between density based algorithm or simple K-means algorithm and also can be found Accuracy of D-stream and K-means algorithm.

V. FUTURE SCOPE

K-means is unable to handle arbitrary cluster formation because prediction of the number of classes to be formed is not fixed. The D-stream algorithm has superior quality and efficiency, can find clusters of arbitrary shapes, and can accurately recognize the evolving behaviors of real-time data streams. Therefore, D-stream will perform better in biomedical applications. This system can be further developed for real time analysis of biomedical data to predict patient's current health status.

This concept can be used for monitoring elderly people, Intensive Care Unit (ICU) Patient. Also the system gives the health status of patient, it can be used be used by clinicians to keep the records of patients.

V. Conclusion

This concept is adaptive, since it can handles more than one physiological signal. This concept 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, etc. Prediction of health status is very sensitive job; D-stream will perform better here, as it supports arbitrary cluster formation which is not supported by K-means. Also D-stream is particularly suitable for users with little domain knowledge on the application data that means it won't require the K-values. Hence D-stream is parameter free and proves to give more accurate results than K-means when used for cluster formation of historical biomedical data.

REFERENCES

[1] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy , R. G. R.: Advances in Knowledge Discovery and Data Mining. AAAI Press / the MIT Press, Menlo Park, CA. (1996).

[2] Mahnoosh Kholghi, Mohammadreza Keyvanpour, "An Analytical Framework For Data Stream Mining Techniques Based On Challenges And Requirements", Mahnoosh Kholghi et al. / International Journal of Engineering Science and Technology (IJEST), Vol. 3 No. 3 Mar 2011.

[3] Yixin Chen, Li Tu, "Density-Based Clustering for Real-Time Stream Data"in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 2007.

[4] Liu S.H., Chang K.M. and Tyan C.C. (2008). Fuzzy c-means Clustering for Myocardial Ischemia Identification with Pulse Waveform Analysis. Proceedings of the 13th International Conference on Biomedical Engineering, Singapore, Vol. 23, 485-489.

[5] Vikram Singh, Sapna Nagpal "A Guided clustering Technique for Knowledge Discovery – A Case Study of Liver Disorder Dataset", International Journal of Computing and Business Research, Vol.1, no. 1, Dec 2010.

[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] 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.

[8] 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

[9] The MIMIC database on PhysioBank (2007, Oct.) [Online]. Available: http://www.physionet.org/physiobank/database/mimicd

[10] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", Second Edition

First Author Mr. Swapnil D. Raut, M.E. Scholar, Department of Computer Science and Engineering,,G. .H. Raisoni College of Engineering and Management, Amravai.

Second Author Prof. Avinash Wadhe , Lecturer, Department of Computer Science and Engineering,,G. .H. Raisoni College of Engineering and Management, Amravai.