International Journal of Technical Research and Applications e-ISSN: 2320-8163,
www.ijtra.com Volume 2, Issue 2 (March-April 2014), PP. 07-10

K-MEANS AND D-STREAM ALGORITHM IN HEALTHCARE

Ms Samidha N. Kalwaghe
M.E I year
Shram Sadhana Bombay Trust's College of Engineering and Technology
Bambhori, Jalgaon,
Maharastra, INDIA

***Abstract*— The healthcare industry is considered one of the largest industry in the world. The healthcare industry is same as the medical industries having the largest amount of health related and medical related data. This data helps to discover useful trends and patters that can be used in diagnosis and decision making. Clustering techniques like K-means, D-streams, COBWEB, EM have been used for healthcare purposes like heart disease diagnosis, cancer detection etc. This paper focuses on the use of K-means and D-stream algorithm in healthcare. This algorithms were used in healthcare to determine whether a person is fit or unfit and this fitness decision was taken based on his/her historical and current data. Both the clustering algorithms were analyzed by applying them on patients current biomedical historical databases, this analysis depends on the attributes like peripheral blood oxygenation, diastolic arterial blood pressure, systolic arterial blood pressure, heart rate, heredity, obesity, and this fitness decision was taken based on his/her historical and current data. Both the clustering algorithms were analyzed by applying them on patients current biomedical historical databases, this analysis depends on the attributes like peripheral blood oxygenation, diastolic arterial blood pressure, systolic arterial blood pressure, heart rate, heredity, obesity, cigarette smoking. By analyzing both the algorithm it was found that the Density-based clustering algorithm i.e. the D-stream algorithm proves to give more accurate results than K-means when used for cluster formation of historical biomedical data. D-stream algorithm overcomes drawbacks of K-means algorithm**

***Key words*— K-means, D-stream, healthcare, biomedical data, fitness etc.**

I. INTRODUCTION

Along with the application in various fields like e-business, marketing and retail data mining has proved its importance in the field of healthcare also. This paper intends to provide a new look to the field of healthcare The Traditional datamining and OLAP techniques are not useful for continuous data i.e. stream data clustering as it requires multiple scans of data and it becomes tedious to scan the continuous data. We will use K-means and D-stream algorithm for the cluster formation the input will be various signals and general health habits of patient. Signals include blood pressure such as systolic arterial blood pressure and diastolic arterial blood pressure, peripheral blood oxygenation, blood sugar level, heart rate, heredity, obesity, cigarette smoking in this.

II. BACKGROUND

DM came into prominence in mid 90s because computers made possible the fast construction of huge data warehouses, containing potentially large amounts of information. The modern day statistical techniques and the advances in probability theory offered the necessary analytical tools.

John Snow detected the source of cholera counted the number of deaths and plotted the victim's addresses on the map as dots. He found that most of the deaths were clustered towards a specific water pump in London, this water pump was the main source of disease spread

Related Work

Patient Clustering in Healthcare Monitors: A system was develop using several clustering mechanisms experiments were conducted on the data collected from the online questionnaire monitors which validated the feasibility of patient clustering, and at the same time, indicate directions of further work.[3]
Learning and Sequential Decision Making For Medical Data Streams Using Rl Algorithm: Here using reinforcement learning algorithm on diabetes data helped in making effective decision about giving specific quantity of insulin dose that should be given to the patient at particular time.[4]

Osama Abu Abbas in his article "comparison between data clustering algorithms" compared four different clustering algorithms (K-means, hierarchical, SOM, EM) according to the size of the dataset, number of the clusters and type of S/W. Osama tested all the algorithms in LNKnet S/W[5]

T. velmurgun in his research paper "performance evaluation of K-means & Fuzzy C-means clustering algorithm for statistical distribution of input data points" studied the performance of K-means & Fuzzy Cmeans algorithms. The cluster centers were calculated for each clusters by its mean value and clusters were formed depending upon the distance between data points [6]

III. METHODOLOGY

Cluster analysis or clustering is the process of partitioning a set of data objects into subsets. Each subset is a cluster, such that objects in a cluster resemble one another, yet dissimilar to objects in other clusters.In this context, different clustering methods may generate different clusterings on the same data set. The partitioning is not performed by humans, but by the clustering algorithm

Both K-mean and D-stream clustering algorithms were used for formation of clusters on medical database. The data was collected from the mimic data set. The data set contains the 110 instances and the 12 attributes. The attributes are age, sex, Blood Pressure, Cholesterol, Chest Pain and etc. The

performance of these algorithms will be computed by using correctly predicted instance. [1].

The flow of the system is depicted in Fig.1

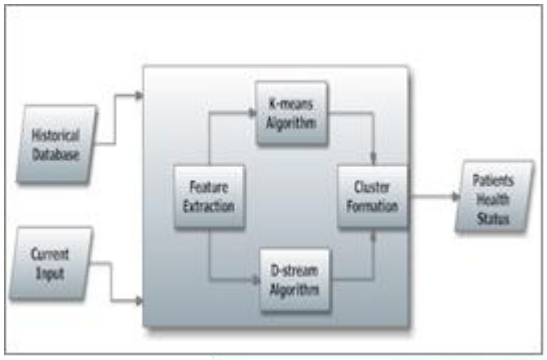

Fig 1.Flow of the system

We will cluster patient's record this clusters will help doctors to diagnose the disease of the patient the steps are as follows

1) Historical database: The MIMIC database [7] will seem to be our historical database from which Attributes such as SpO2, ABPsys, ABPdias, HR are collected

2) Current input: Other attributes such as heredity, obesity, cigarette smoking etc are computed by the person's behavior and this will be our current input

3) Model Building

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

4) Feature Extraction

For each physiological signal x among the X monitored vital signs, we extract the following features [8].

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.

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.

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.

5) K-means and D-stream algorithm: This two algorithms are used for Data Set formation

6) 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.

K-Means: A Centroid-Based Technique

The k-means algorithm defines the centroid of a cluster as the mean value of the points within the cluster. First, it randomly selects k of the objects in D, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the Euclidean distance between the object and the cluster mean. The k-means algorithm then iteratively improves the within-cluster variation. For each cluster, it computes the new mean using the objects assigned to the cluster in the previous iteration clusters formed in the current round are the same as those formed in the previous round. The k-means procedure along with algorithm is given below

Algorithm K-means:

Input = K: The number of clusters
= D: A dataset containing n objects
Output = A set of K clusters
Method:

(1) Arbitrarily choose K-objects from D as the initial cluster centers

(2) Repeat

(3) Re-assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster

(4) Update the cluster means, i.e. calculate the mean value of the objects for each clusters

(5) Until no changer

The time complexity of the k-means algorithm is O(nkt),where n is the total number of objects, k is the number of clusters, and t is the number of iterations. Normally, $k \ll n$ and $t \ll n$. Therefore, the method is relatively scalable and efficient in processing large data sets

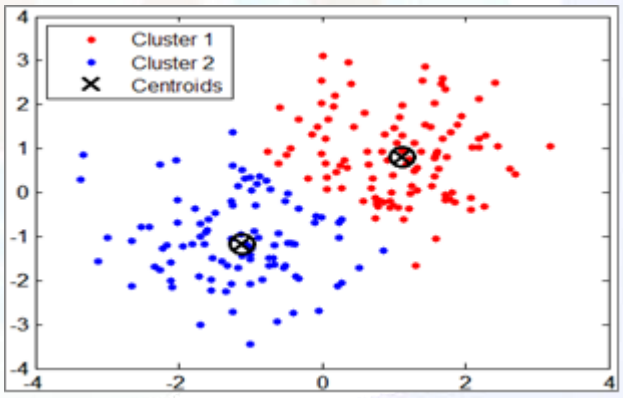

Fig 2: K-means clustering view

D-stream Clustering Algorithm

D-Stream has an online component and an offline component. 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. 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. Then, the algorithm periodically removes sporadic grids and regulates the clusters.

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.

The D-stream algorithm is explained as follows [2]

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

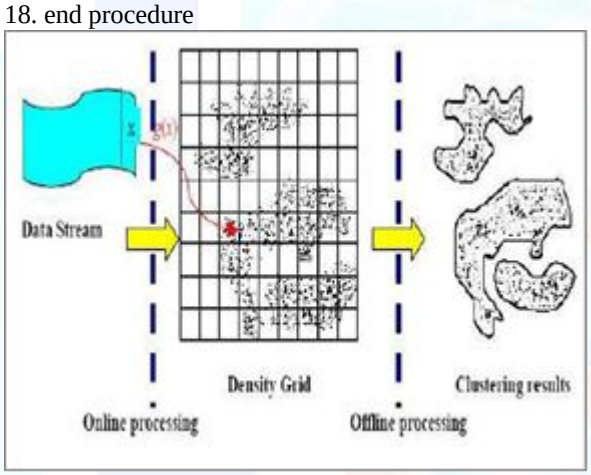

Fig 3: Illustration of the use of density grid.

The input data has d dimensions, and each input data record is defined within the space

$S = S_1 U S_2 * \cdots * S_d, \ldots\ldots$ (1)

here 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

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

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

$$S_{i,j1} * S_{2,j2} \cdots S_{d,jd}, \quad j_i = 1, \ldots, p_i,$$

We denote it as

$$g = (j_1, j_2, \cdots j_d). \quad \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

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 .

For each data record x, we assign it a density coefficient which decreases with as x ages. In fact, if x arrives at time tc, we define its time stamp T(x) = tc, and its density coefficient D(x, t) at time t is

$$D(x,t) = \lambda^{t-T(x)} = \lambda^{t-tc}, \ldots\ldots\ldots (4)$$

Where $\lambda \epsilon (0, 1)$ 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:

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

IV. DISCUSSION

Both the clustering algorithms were analyzed by applying them on patients current biomedical historical databases, this analysis depends on the attributes described previously. From this the performance accuracy of both the algorithm was calculated. It was found that performance of D-stream algorithm is greater than K-means algorithm

Performance Accuracy = correctly predicted Instance/ Total Number of Instance

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

Table 1: Performance of k-mean and D-stream algorithm

V. CONCLUSION

After Studying and analyzing both the algorithm it was found that K-means clustering algorithm faces difficulty in comparing quality of the clusters produced, does not work well with non-globular clusters and is unable to select variables automatically. All this disadvantages of K-means algorithm are removed by the D-stream algorithm as it superior quality and efficiency ,can find clusters of arbitrary shape and can accurately recognize the evolving behavior of real-time data streams and it is parameter free

REFERENCES

[1] 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

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

[3] Xiao Hu Patient Clustering in Healthcare Monitors:LIS429 Implementation of Information Storage and Retrieval Spring 2004

[4] Prof Pramod P, Dr. Parag K, Ms. Rachana S,"Learning and Sequential Decision Making For Medical Data Streams Using Rl Algorithm", International Journal of Research in Computer andCommunication Technology, Vol 2, Issue 7, July-2013

[5] Osama A. Abbas(2008),Comparison between data clustering algorithm, The International Arab journal of Information Technology, vol 5, N0. 3

[6] Velmurugan T., T. Santhanam(2010), performance evaluation of k-means & fuzzy c-means clustering algorithm for statistical distribution of input data points., European Journal of Scientific Research, vol 46

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

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