Proceedings of the 4th National Conference; INDIACom-2010
Computing For Nation Development, February 25 – 26, 2010
Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi

Data Stream Mining Using Landmark Stream Model for Offline Data Streams: A Case Study of Health Care Unit.

P. K. Srimani[1] and **Malini M. Patil**[2]
[1]F.N.A.Sc.M.E. (CSE), Director, R&D Division, B.U., DSI, Bangalore
[2]Assistant Professor and HOD, Department of Information Science and Engineering , JSS Academy of Tech. Education, Bangalore-60, Karnataka, Research scholar, Bhartiyaar University, Coimbatore, Tamilnadu.
E-[1]mailprofsrimanipk@gmail.com, [2]patilmalini31@yahoo.com.

ABSTRACT
There exists much different kind of applications in data stream mining. Few main areas of applications include sensor networks, internet packet streams, web logs, medical data and many more. Data streams typically arrive in high speeds and change in data distributions. A different approach is used to mine them. The paper proposes a model for mining such data streams using an application of 24 hours multi-speciality health care unit. The main aim of the paper is to propose a model for identifying the symptoms of patient registrations and proposing an immediate first hand information for the patients using landmark window data stream model and K-means clustering algorithm.

KEYWORDS
Data mining, data streams, offline data streams, online data streams, landmark window, clustering.

1. INTRODUCTION

In recent years many patient monitoring systems are available. The diagnosis done under such systems are used to monitor the patients. Such systems generate different types of data continuously. The data generated can be in the form of a signal, a digital reading or a pathological laboratory result, which can be used for diagnosis. But the literature review suggests different methods, models and approaches only to the monitoring of patients who are in ICU. The analysis of data collected continuously for series of patient registrations is addressed in this paper. The data so stored in traditional static databases, is analyzed using query processing. But when the query processing technique is applied to stream data a different approach is followed. This paper proposes a new model for specific areas of medical diagnosis of patients especially for a super-specialty health care unit (SSHCU) using clustering technique.

The rest of the paper is organized as follows. In section 2, we first give the preliminaries related to basic definitions, types and other features of data streams and also about three different types of window models for data streams. In section 3, specifically more detailed discussion on landmark model and its application to real time example of SSHCU. Section 4, discuses about basics of clustering technique and K–means clustering algorithm. In section 5, we discuss more about the application of the technique of clustering to data streams by taking an example of a SSHCU and explain about the model which can be developed to handle data streams. Section 6, discusses about the conclusion and Section 7 is about future scope of the work and. Finally, the relevant references are appended.

2. PRELIMINARIES

A data stream is an ordered sequence of items that arrives in timely order. Data streams are different from data in traditional databases. They are continuous, unbounded usually come in high speed and have a data distribution which often changes with time [Guha 2001].

Data streams can be further classified into offline streams and online streams [1] . Offline streams are characterized by regular bulk arrivals. For example, web logs. Web logs are considered as offline data streams because most of the reports are generated in a certain period of time. Other examples of offline data streams include queries on data warehouses, medical data in which patient's case history is collected day-wise which we are discussing in this paper as a case study for mining offline data streams.

Online data streams [1] are characterized by real time updated data that come one by one in time. Examples for online data streams are, frequency estimation of internet packet streams, stock market data, and sensor data. Such data should be processed online. Another very important feature of online data streams is they should be processed online with the rapid speed with which they arrive and should be discarded immediately after being processed. Yet another important feature is bulk data processing is not possible in online data streams where as it is possible in offline data streams.

Few important challenges of data streams are summarized as follows: **Firstly**, since the data collected is huge, multiple scans are not possible in data streams mining as compared with traditional data mining algorithms. **Secondly,** the mining method of data streams should handle the change in data distribution. **Thirdly,** in case of online data streams mining methods should be more faster than the speed of incoming data. **Fourthly,** memory management issues related to data storage and CPU speed also matter more in data stream mining.

Data stream processing models are also key features in data stream mining. They are application dependent. The three basic

data processing models are **landmark model, damped model, sliding windows model**. The **landmark model** mines all the frequent itemsets over the entire history of stream data from a specific time point called landmark to present. This model is suitable for applications such as stock monitoring systems where people are interested only in the most recent information of data streams. **Damped model** mines frequent itemsets in stream data in which each transaction has a weight and this weight decreases with age. Older transactions contribute less weight towards itemset frequencies. This is suitable for applications in which old data has an effect on the mining results, but the same effect decreases as time goes on. The **sliding windows** model finds and maintains frequent itemsets in sliding windows. Only part of the data streams within the sliding window are stored and processed at the time when the data flows in. The size of the sliding window is application and machine dependent.

With these few important features of data streams, we are discussing the mining of offline data streams, using an application of SSHCU in this paper.

Mathematical model : Let I be a set of items. An itemset (or a pattern), $I = \{x_1, x_2, \ldots, x_k\}$, is a subset of I. An itemset consisting of k items is called a k-itemset and is written as $x_1, x_2, \ldots, x_k$. We assume that the items in an itemset are lexicographically ordered. A transaction is a tuple, (tid, Y), where tid is the ID of the transaction and Y is an itemset.

A transaction data stream is a sequence of incoming transactions and an **extracted part of the stream** is called a **window**. A window, W, can be either time-based or count-based according to the number of transactions that are updated each time and either a landmark window or a sliding window. W is time-based if W consists of a sequence of fixed-length time units, where a variable number of transactions may arrive within each time unit. W is count-based if W is composed of a sequence of batches, where each batch consists of an equal number of transactions.

W is a landmark window if $W = (T_1, T_2, \ldots, T_n)$. W is a sliding window if $W = (T_{n-w+1}, \ldots .T_n)$ where each T_i is a time unit or a batch, T_1 and T_n are the oldest and the current time units or batches, and w is the number of time units or batches in the sliding window, depending on whether W is time-based or count-based.

3. CLUSTERING TECHNIQUE: K–MEANS ALGORITHM

K-means and K-medoid are two important prototype-based clustering techniques. Such techniques create a one-level partitioning of the data objects. **K-means** defines a prototype in terms of a centroid, which is usually the mean of a group of points, and is typically applied to objects in a **continuous n-dimensional space.**

K-medoid defines a prototype in terms of a **medoid**, which is the most representative point for a group of points, and can be applied to a wide range of data since it requires only a proximity measure for a pair of objects. While a centroid almost never corresponds to an actual data point, a medoid, by its definition, must be an actual data point. In this section we are discussing only K-means clustering algorithm in detail.

The Basic K-means Algorithm

The K-means clustering technique is simple and also one of the oldest and most widely used algorithms. The description of the algorithm is given below.

- We first choose K initial centroids, where K is a user specified parameter, namely, the number of clusters desired.
- Each point is then assigned to the closest centroid.
- Each collection of points assigned to a centroid is a cluster.
- The centroid of each cluster is then updated based on the points assigned to the cluster.
- We repeat the assignment and update steps until no point changes clusters, or equivalently, until the centroids remain the same.

K-means is formally described by Algorithm 1. The operation of K-mean is illustrated in Figure 3.1, which shows how, starting from three centroids, the final clusters are found in four assignment-update steps. In these and other figures displaying K-means clustering, each subfigure shows (1) the centroids at the start of the iteration and(2) the assignment of the points to those centroids. The centroids are indicated by the "+" symbol; all points belonging to the same cluster have the same marker shape.

Algorithm 1: Basic K-means algorithm.

1: Select K points as initial centroids
2: Repeat
3: Form K clusters by assigning all points to the closest Centroid.
4: Re-compute the centroid of each cluster
5: Until the centroids don't change

Demonstration of the Basic K-means algorithm is shown in the following figures:

Figure 1 (a)

1) *k* initial "means" (in this case *k*=3) points are randomly selected from the data set (shown in colors).

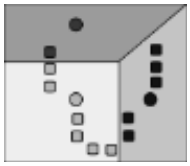


Figure 1 (b)

2) ***k* clusters** are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

Figure 1 (c)

3) The **centroid** of each of the *k* clusters becomes the new **means**.

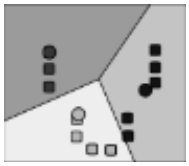


Figure 1 (d)

4) Steps 2 and 3 are repeated until the convergence is achieved.

In the first step, shown in Figure 1(a), iteration 1, points are assigned to the initial centroids, which are all in the larger group of points. For this example, we use the **mean** as the centroid. After points are assigned to a centroid, the centroid is then updated. Again, the figure for each step shows the centroid at the beginning of the step and the assignment of points to those centroids. In the second step, points are assigned to the updated centroids, and the centroids are updated again. In steps which are shown in iterations in the figures respectively, two of the centroids move to the two small groups of points at the bottom of the figures. K-means algorithm terminates since no more changes occur, the centroids have identified the natural groupings of points as shown in figure 1(d).

4. APPLICATION OF LANDMARK WINDOW MODEL ON SSHCU USING CLUSTERING TECHNIQUE

The definition of landmark model given in section 1, is with respect to association rule mining of data streams. An attempt is made to apply the clustering technique in a **continuous 2-dimensional space** to the same window model which is **time based** in the current scenario for the data collected.

The real time activities of a SSHCU are mapped according to the definition of landmark window. As it is quite commonly noted, the 24 hours service of the unit is divided in to three shifts of 8 hrs each. Let's assume that the time slots are fixed as first shift: 12:00 PM to 8:00 A.M. Second shift: 8 A.M. to 4 P.M. third Shift: 4:00 P.M. to 12:00 P.M. These three shifts are treated as three different windows or batches.

On daily basis the patient registrations are observed and recorded at the registration desk with respect to time and date of registration. The shifts are assumed as three different slots, and in future three different windows as per landmark window. Such patient registrations here are considered as stream data as they arrive continuously. This approach helps in analyzing data easily.

Now first we will consider one batch of data say B2 that is time slot of Second shift: 8 A.M. to 4 P.M. Let the set of patients be represented as P= $\{a_1,\ldots,a_n , b_1,\ldots,b_n,\ldots z_1,..,z_n\}$ $W_1=\{a_1, b_1, c_1, d_1, e_1\ldots.z_1\}$, $W_2= \{a_2, b_2, c_2, d_2, e_2\ldots.z_2\}$, $W_3= \{a_3, b_3, c_3, d_3, e_3\ldots.z_3\}$ where W_1, W_2, W_3 are the three consecutive windows. Here, every patient's registration is treated as one transaction.

In this part of the paper we will try to map the real time data of patient registration with the theoretical definition of data streams and also try to apply clustering technique on the same. If patient 'a_1' admits with a complaint of severe stomach pain, he is diagnosed by a physician and advised immediate scanning which results in a minor operation of appendicitis. If patient 'a_2' registers with a complaint of severe thirst and frequent urination, after pathological tests he is diagnosed with diabetes mellitus. If patient 'a_3' complaints with sever chest pain and no signs of consciousness, he will be referred to the emergency unit, and will be sent to ICU for further treatment.

The above-referred three cases are few cases in health care unit. Each patient's registration is considered as single transaction. The list will be endless if we list manually. The details of data collected and their mapping with windows is shown in table 1. Each row in the table represents data collected in three consecutive batches per day. Since the diagnosis is time dependent and past history of patients is needed and preserved for future analysis the data stream is considered as offline data stream. These typical tables are best suited for mining data streams using association rule mining. As the aim of the paper is in applying the clustering technique, a still more micro level mapping of data is necessary which is represented in table 2 (by adding one more column **date**). This table will be used for the present analysis using the clustering technique.

Shift times →	**12:00P.M. to 8:00A.M.**	**8:00A.M. to 4:00 P.M**	**4:00P.M. to12:00P.M**
Date /batch	**W_1**	**W_2**	**W_3**
8/11/09	a_1 b_1	a_2 c_2	a_3 b_3 d_3
9/11/09	d_1 f_1	d_2 b_2	a_3 b_3 d_3 c_3
10/11/09	a_1 s_1	a_2 c_2	d_3 b_3 a_3
11/11/09	b_1 c_1 z_1	a_2 b_2 y_2	a_3

Table 1: Representation of data streams using Landmark window.

5. IMPLEMENTATION OF OUR METHOD

Our method of using the landmark window model for the patient diagnosis using clustering algorithm is shown in figure 2. Implementation of K-means clustering algorithm needs one more level of data mapping as discussed earlier.

For this lets take the data from first row from all the windows. Say $W_1 = \{a_1,b_1\}$,$W_2=\{a_2 , c_2\}$, $W_3 = \{a_3 , b_3 ,d_3\}$. The following table is created with respect to the symptoms of all the patients in all three windows. After applying K-means

algorithm the relative patients are grouped into three different clusters as **cluster 1, cluster 2, cluster 3.**

Window	Patient	Symptoms	Clusters
W_1	a1	Frequent urination	Cluster 1
	b1	Chest pain	Cluster2
W_2	a_2	Thirst, sweating	Cluster1
	c_2	Scanning	Cluster2
W_3	a_3	High Blood sugar	Cluster1
	b_3	Blood pressure	Cluster1
	d_3	Stomach pain	Cluster3

Table 2 : Data mapping for applying K-means Algorithm and clusters.

Table 2 shows the possible clusters using K-means algorithm which is used on the sample data collected. The model is shown in figure 2. The designed model is capable of handling offline data streams for a SSHCU. The model comprises of seven main modules mainly, data preprocessing module where the preprocessing activity takes place; visualization tools module provides user interfaces; medical rule manager module is the decision making module; Stream data manager module is the data storage module; the last important module is the stream data query manager module which mainly performs the data analysis of patients.

6. CONCLUSION

The results of present investigation high light the following important features.

Firstly, since the data collected is huge, multiple scans are not possible in data streams mining as compared with traditional data mining algorithms.

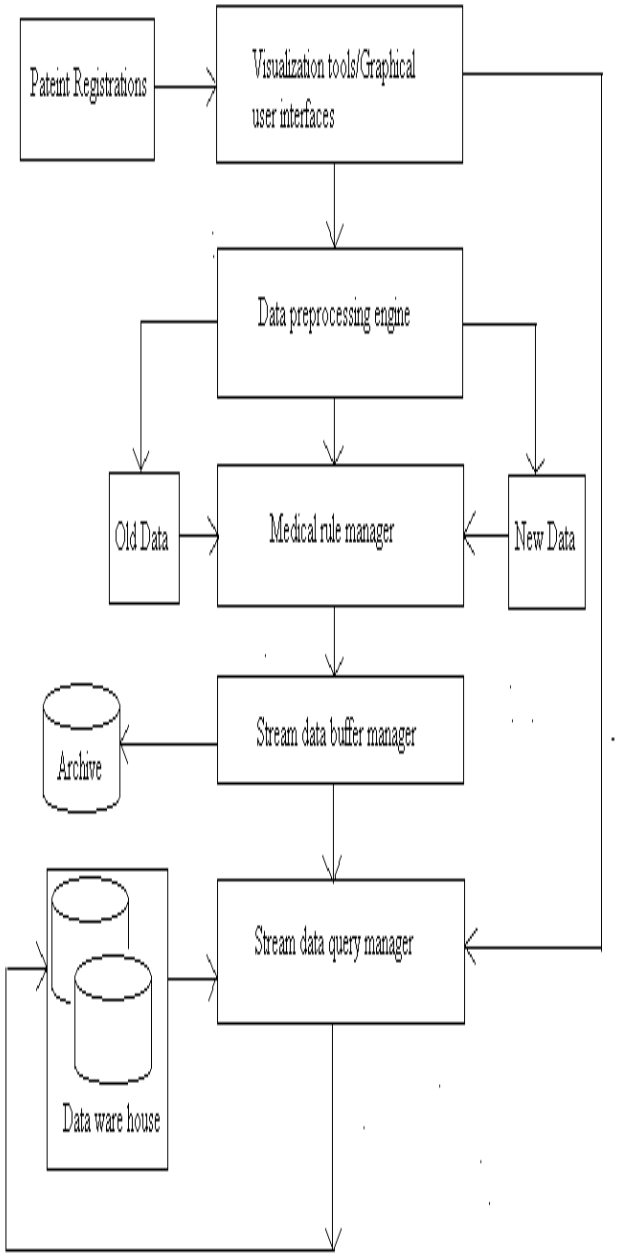
Secondly, the traditional algorithms require lot of memory management issues and results in CPU overhead.

Thirdly, for the landmark model proposed association rule mining technique is best suited as per its definition stated in section 2. but in the present situation it results in the above mentioned disadvantages.

Finally , the proposed technique discussed in this paper can overcome the above said drawbacks no doubt and also is capable of generating possible, valid clusters

7. FUTURE SCOPE

Currently extensive research work is going on with regard to the optimization of the health care predictions, and patient monitoring systems. But the results are not optimal. Hence, the results of the present investigation will be of great practical interest.



8. REFERENCES

[1]. Nan Jiang and Le Gruenwald, "research issues in data stream association rule mining " Procedings of SIGMOD record, vol.35 No. 1, Mar. 2006.

[2]. R. Agrawal, T. Imielinski, and A. N. Swami. Mining Association Rules between Sets of Items in Large Databases. In Proc. of SIGMOD, 1993.

[3]. R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules. In Proc. ofVLDB, 1994.

[4]. Mohamed Medhat Gaber, Arkady Zaslavsky and Shonali Krishnaswamy; Resource-Aware Knowledge Discovery in Data Streams; Int'l Workshop on Knowledge Discovery in Data Streams; September 2004.