ISSN 2348 – 9928
IJAICT Volume 6, Issue 3, March 2019 Doi:01.0401/ijaict.2015.05.08 Published on 05 (3) 2019

A HYBRID CLUSTERING ALGORITHM FOR DATASTREAMS

Ms. Anagani. Bhanu Sree
Student of M.Tech., Dept. of CS & SE,
Andhra University College of Engineering (A),
Visakhapatnam, Andhra Pradesh, India

Mrs. A. Mary Sowjanya
Assistant Professor, Dept. of CS & SE,
Andhra University College of Engineering (A),
Visakhapatnam, Andhra Pradesh, India

Abstract*— *Data mining is an extensively studied field of research area, where most of the work involves knowledge discovery. Data stream is a dynamic research area of data mining. A data stream is an enormous sequence of data elements continuously generated at a fast rate. In data streams, huge amount of data continuously inserted and queried, such data has very large database, for example, consumer click streams and telephone records, bulky sets of web pages, multimedia data, and financial transactions and so on. It raises new problems for the data stream community in terms of how to mine continuous arrival of high speed data items. Many researchers have focused on mining data streams and they proposed many techniques for data stream classification, data stream clustering and finding frequent items from data streams. Data stream clustering techniques are highly helpful to cluster the similar data items in data streams and also to detect the outliers, so they come under cluster based outlier detection. For this purpose we propose a new algorithm which is a combination of both hierarchical and partitioning clustering algorithms. A probabilistic hierarchical clustering algorithm in combination with K-Means and this yields more accurate results than that of those algorithms performed individually. The hybrid approach records good clustering accuracy.

Keywords— ***Data stream, Data stream clustering, hierarchical clustering algorithm, K-Means.***

I. INTRODUCTION

Data mining is an interdisciplinary subfield of computer science is the computational process of discovering pattern in large datasets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of data mining process is to extract information from a data set and transform it into an understandable structure for further use. A data stream is an ordered sequence of points that can be read only once or a small number of times. The performance of an algorithm that operates on data streams is measured by the number of passes the algorithm must make over the stream, when available memory is taken as the constrained. The data stream model is motivated by emerging application involving massive data sets, e.g., customer click streams, telephone records, large sets of web pages, multimedia data, and sets of retail chain transactions can be modeled as data streams [8]. Data streams can be solved using the methodologies of data stream clustering, data stream classification, frequent pattern mining, sliding window, Association technique and so on. Clustering data streams requires a process able to continuously cluster objects within memory and time restrictions. Clustering is a process of grouping objects with similar properties. Any cluster should exhibit two main properties; low inter-class similarity and high intra-class similarity. Clustering is an unsupervised learning i.e. it learns by observation rather than examples. There are no predefined class label exists for the data points. Clustering algorithm can be divided into five categories namely Partition clustering algorithm, Hierarchical clustering algorithm, Spectral clustering algorithm, Grid based clustering algorithm, Density based clustering algorithm [3].

A partitioning method constructs k partitions of the data, where each partition represents a cluster and k<=n. Most applications adopt popular heuristic methods such as greedy approaches like the k-means and the k-medoids algorithm, which progressively improve the clustering quality and approach a local optimum. A hierarchical clustering method works by grouping data objects into a hierarchy or "tree" of clusters. There are several orthogonal ways to categorize hierarchical clustering methods. They can be categorized into algorithmic methods, probabilistic methods and Bayesian methods. An agglomerative clustering starts with one-point (singleton) clusters and recursively merges two or more most appropriate clusters. A divisive clustering starts with one cluster of all data points and recursively splits the most appropriate cluster. The process continues until a stopping criterion (frequently, the requested number k of clusters) is achieved. The remaining paper includes the review of literature in Section 2. Section 3 explains about the clustering algorithms, the probabilistic hierarchical clustering and k-means. Experimental results are shown in Section 4 and Conclusions are given in Section 5.

II. LITERATURE SURVEY

A comprehensive classification of data stream mining challenges and then evaluates new methods of data stream mining is

proposed [1]. In this evaluation they consider the relation of those methods with different data mining techniques in an analytical manner. Results show that it is necessary to adopt many summarization and approximation techniques from the fields of statistics and computational theory, besides crucial changes that are needed in common data mining techniques.
Some of the improved hierarchical clustering algorithms were discussed [2]. The quality of a pure hierarchical clustering method suffers from its inability to perform adjustment, once a merge or split decision has been executed. This merge or split decision, if not well chosen at some step, may lead to some-what low quality clusters. One promising direction for improving the clustering quality of hierarchical methods is to integrate hierarchical clustering with other techniques for multiple phase clustering.

Different types of partition clustering algorithm were discussed [3]. It describes about the general working behavior, the methodologies followed on these approaches and the parameters which affects the performance of these algorithms. The drawback of k-means algorithm is to find the optimal k value and initial centroid for each cluster. This is overcome by applying the concepts such as genetic algorithm, simulated annealing, harmony search techniques and ant colony optimization.

A query redirection method that improved K-means clustering algorithm performance and accuracy in distributed environment was proposed [4]. Analysis on k-mean and hierarchical algorithm by applying validation measures like entropy, f-measure, coefficient of variance and time. The experimental results show that k-mean algorithm performs better as compared to hierarchical algorithm and takes less time for execution.
A novel hybrid approach to combine merits of the hierarchical and k-means clustering algorithm and discard disadvantages has been presented [5]. It is different from existed method: carry out hierarchical clustering first to decide location and number of clusters in the first round and run the K-means clustering in another round.

III. METHODOLOGY

Clustering is one of the important tasks in data streams and outlier detection is based on the clustering methods that are used. In this research work the hierarchical clustering algorithm is combined with the partitioning clustering algorithm. The below figure shows the system architecture of the research work.

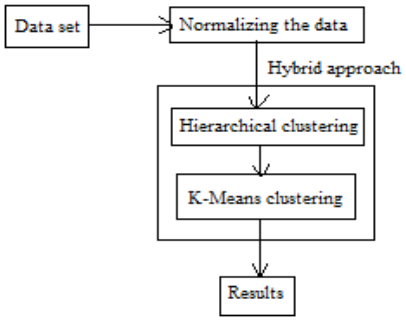

Fig. 1. System architecture of clustering algorithm

3.1 Dataset

In order to perform the data stream clustering, the NASA weather related data is the dataset with 18767 instances and 24 attributes. Data stream is an unbounded sequence of data as it is not possible to store complete data stream, for this purpose we divide the data into chunks of same size and each chunk is specified by the user which depends upon the nature of the data [6].

3.2 Distance Measurements

In clustering, we want to group data that are similar. Thus, we need to compute the distance between two data objects, which can be done in many forms. In this paper we are using the following measurements.

1. Euclidean distance

The most popular distance measure is Euclidean distance. Let i and j be two objects described by p numeric attributes. The Euclidean distance between two objects is defined as

$$d(i,j)= \sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+ \cdots +(x_{ip}-x_{jp})^2}. \quad (1)$$

2. Correlation Coefficient

For numeric attributes, we can evaluate the correlation between two attributes A and B, by computing the correlation coefficient. This is

$$r_{A,B=} \sum^{n}_{i=1}(a_i-A\square)\,(b_i-B\square)/n\sigma_A\,\sigma_B \quad (2)$$

3.3 Clustering

Clustering is the process of partitioning a set of data objects into subsets [7]. Each subset is a cluster, such that objects in a cluster are similar to one another, yet dissimilar to objects in other cluster. Clustering has been widely used in many applications such as business intelligence, image pattern recognition, web search, biology, and security. Clustering is also called as data segmentation and used for outlier detection.

Clustering is a challenging research field. There are different types of clustering algorithms namely partition clustering

algorithm, hierarchical clustering algorithm, spectral clustering algorithm, grid based clustering algorithm, density based clustering algorithm. Clustering is known as unsupervised learning because the class label information is not present.

3.4 Hierarchical clustering

A hierarchical clustering method works by grouping data objects into a hierarchy or "tree" of clusters [7]. Representing data objects in the form of a hierarchy is useful for data summarization and visualization. Hierarchical clustering methods can encounter difficulties regarding the selection of merge or split points. A promising direction for improving the clustering quality of hierarchical methods is to integrate hierarchical clustering with other clustering techniques, resulting in multiple-phase clustering. BRICH and Chameleon are the two such methods.

There are several orthogonal ways to categorize hierarchical clustering methods. They can be categorized into algorithmic methods, probabilistic methods and Bayesian methods. Agglomerative, divisive and multiphase methods are algorithmic meaning they consider data objects as deterministic and compute clusters according to the deterministic distances between objects. In agglomerative approach which is also called as bottom up approach, each data points are considered to be a separate cluster and on each iteration clusters are merged based on a criteria. The merging can be done by using single link, complete link, centroids or wards method. In divisive approach all data points are considered as a single cluster and they are divided into number of clusters based on certain criteria, and this is called as top down approach. Probabilistic methods use probabilistic models to capture clusters and measure the quality of clusters by the fitness of models. Bayesian methods compute a distribution of possible clustering, instead of outputting a single clustering over a dataset; they return a group of clustering structures and their probabilities.

3.5 Probabilistic Hierarchical clustering

Probabilistic hierarchical clustering aims to overcome some of the disadvantages of other clustering algorithms by using probabilistic models to measure distances between clusters. The task of clustering is to estimate the generative models as accurately as possible using the observed data objects to be clustered. The data generative models adopt common distribution functions, such as Gaussian distribution which are governed by parameters. Probabilistic hierarchical clustering algorithm is as follows.

A probabilistic hierarchical clustering method can adopt the agglomerative clustering framework, but use probabilistic models to measure the distance between clusters. The clustering scheme can start with one cluster per object, and merge two

Input: A dataset containing n objects

Output: A hierarchy of clusters

Method:

Create a cluster for each object $C_i=\{o_i\}$, $1\leq i\leq n$;

for i=1 to n

Find pair of clusters C_i and C_j such that $C_i,C_j=\arg\max_{i\neq j} \log(P(C_i \cup C_j) / P(C_i)P(C_j))$;

if $\log(P(C_i \cup C_j) / P(C_i)P(C_j)) > 0$ then merge C_i and C_j;

else stop;

clusters, C_i and C_j, if the distance between them is negative. In each iteration, we try to find C_i and C_j so as to maximize $\log(P(C_i \cup C_j) / P(C_i)P(C_j))$.The iteration continues as long as $\log(P(C_i \cup C_j) / P(C_i)P(C_j)) > 0$, that is, as long as there is an improvement in clustering quality. These are more interpretable and can handle partially observed data.

3.6 Partitioning Clustering Algorithm

A partitioning method constructs k partitions of the data, where each partition represents a cluster and k<=n. In other words it is a one level partitioning on data sets and typically adopt exclusive cluster separation. These are distance based and use an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another.

Achieving global optimality in partitioning based clustering is often prohibitive, potentially requiring an exhaustive enumeration of all the possible partitions. Most applications adopt popular heuristic methods such as greedy approaches like the k-means and the k-medoids algorithm, which progressively improve the clustering quality and approach a local optimum.

1. K-Means Clustering

K-means clustering is a well known partitioning method. This is a centroids based partitioning technique which uses the centroids of a cluster. The centroids can be defined in various ways such as by the mean or medoids of the objects. The distance between the object of the cluster to the centroid is measured by the Euclidean distance. The quality of cluster can be measured by the within cluster variation, which is the sum of squared error

between all the objects in the cluster and the centroids is given as

$$E=\sum_{i=1}^{k}\sum_{p\epsilon Ci}dist(p,c_i)^2 \quad (3)$$

E is the sum of the squared error for all objects in the dataset. This objective function tries to make the resulting k clusters as compact and as separate as possible. There are no efficient solutions known to any of these problems and some formulations are NP-hard. The algorithm has a space complexity of O (I * K * m * n), where I is the number of iterations, K is the number of clusters, m is the number of dimensions and n is the number of points. The K-means algorithm is as follows.

Input:

k: the number of clusters,

D: a dataset containing n objects.

Output: a set of k clusters.

Method:

arbitrarily choose k objects from D as the initial cluster centers

Repeat

For each point, find its closes centroid and assign that point to the centroid. This results in the formation of K clusters.

Recompute centroid for each cluster.

Until the centroids do not change.

In the first step, points are assigned to the initial centroids, which are all in the larger group of points. After points are assigned to a centroid, the centroid is then updated. In the second step, points are assigned to the updated centroids, and the centroids are updated again. When the k-means algorithm terminate, the centroids would have identified the natural groupings of points. For some combinations of proximity functions and types of centroids, k-means always converge to a solution i.e., k-means reaches a state in which no points are shifting from one cluster to another and hence the centroids do not change

IV. EXPERIMENT RESULTS

We have implemented our algorithm in MATLAB 7.10 (R2010a). The NASA weather related data is the dataset with 18767 instances and 24 attributes used for clustering. The clustering quality is assessed by the silhouette coefficient. To measure a cluster's fitness within a clustering, we can compute the average silhouette coefficient value of all objects in the cluster.

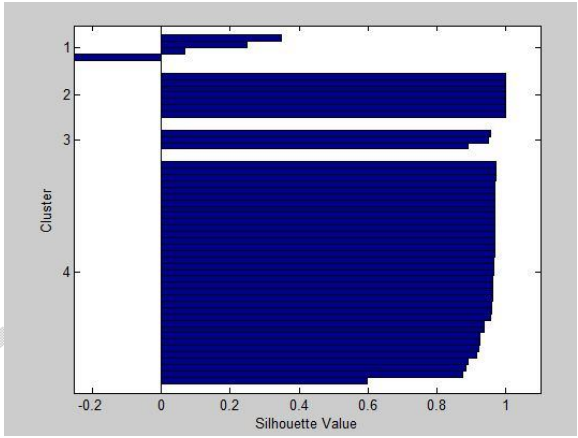

Fig. 2. Graph based on silhouette value after K-means.

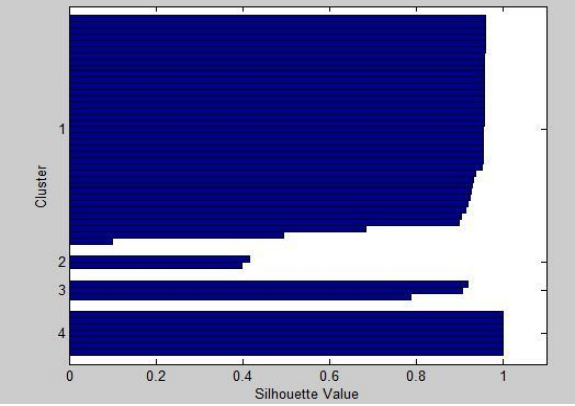

Fig. 3. Graph based on silhouette value after hierarchical with k-means.

Fig. 2. shows the cluster quality of K-means clustering for the given dataset by computing the Silhouette coefficient value for all the objects in the clusters. Fig. 3. shows the cluster quality of hierarchical with K-means clustering for the given dataset by computing the Silhouette coefficient value for all the objects in the clusters.

TABLE I. **RESULTS OBTAINED AFTER CLUSTERING**

	SSE	***Silhouette***	***Correlation***
k-means	0.7396	0.8806	0.4536
hierarchical	0.8621	0.8957	0.5895
Hierarchical+k-means	0	0.8957	0.7707

The above table shows the sum of squared errors (SSE), silhouette and correlation for different algorithms like k-means, hierarchical and hierarchical with k-means. The values differ for all the three algorithms and hierarchical with k-means yields better results than the other two algorithms.

V. CONCLUSIONS

A hybrid clustering algorithm for data streams is proposed. Work has been done on two types of clustering algorithms namely hierarchical and partitioning clustering algorithms. The proposed algorithm is a combination of the probabilistic hierarchical clustering with k-means. The experiment results have shown that Probabilistic Hierarchical with K-means yields more accurate results than K-Means clustering and probabilistic hierarchical clustering. Many more data stream clustering algorithms can be proposed. Along with hierarchical and partitioning clustering algorithms different types of clustering algorithm can be combined to form new hybrid algorithms.

References

[1] MahnooshKholghi, MohammadrezaKeyvanpour," An analytical framework for data stream mining techniques based on challenges and requirements" in International Journal of Engineering Science and Technology,2011.

[2] Yogita Rani and Dr. Harish Rohil, "A study of hierarchical clustering algorithm" in International Journal of Information and Computation Technology. ISSN 0974-2239 Volume 3, Number 11 (2013).

[3] S.Anitha Elavarasi and Dr. J. Akilandeswari, "A survey on partitioning clustering algorithms" in International Journal of Enterprise Computing and Business Systems Vol. 1 Issue 1 January 2011.

[4] Manpreet kaur and Usvir Kaur, " Comparison between k-mean and hierarchical algorithm using query redirection" in International Journal of Advanced Research in Computer Science and Software Engineering,volume 3, Issue 7, July 2013.

[5] Bernard Chen , Phang C. Tai , R. Harrison and Yi Pan, **"**Novel hybrid hierarchical-k-means clustering method (H-K-means) for microarray analysis".

[6] Jerusha Shalini Vaska and A. M. Sowjanya, "Clustering Diabetics Data Using M-CFICA", International Journal of Advanced Computer Research, Volume-5 Issue-20 September-2015

[7] Dr.S.Vijayarani Ms. P. Jothi , "Detecting outliers in data streams using clustering algorithms" in International Journal of Innovative Research in Computer and Communication Engineering Vol. 1, Issue 8, October 2013.