International Journal of
Science and Engineering Investigations vol. 2, issue 20, September 2013
ISSN: 2251-8843

A Clustering Algorithm for Data Mining with Prediction Technique

Akbar Momeni[1], Karim Khazaei[2]
[1,2]Islamic Azad University, Takestan Branch
([1]momeni_akbar@yahoo.com, [2]k_khazaei@yahoo.com)

Abstract-Data mining is a process that uses a collection of techniques and methods to recognizing patterns and relation of data in a database or a collection of centralized data's. Data mining can be used in variety of applications such as finance, commercials, statistical, weather forecasting and other sciences Because data's continuously are increasing in the form of size and dimensions, we need high speed techniques and algorithms for categorizing and processing of information. Applying the parallel processing in data mining is one of the methods to increasing the speed. *K-means* is one of the algorithms that are used in data mining process that clusters the data. We propose the K-means clustering algorithm in a parallel manner. Because this algorithm divides the clusters in two parts using an estimation function in each step, we called it Parallel Bisection K-MEANS Prediction (PBKP) that is done in the form of message passing multi processor systems. Two part K-means generate the clusters with the same size. Simulation results by running the algorithms on a Linux work-stations shows that the speed of PBKP have linear relation with number of processors and data's. Also PBPK operate better than parallel K-means respect to dimensions , number of data items and number of clusters.

Keywords- *datamining, dataclearance, online data anlysing, parallel processing and machine learning.*

I. INTRODUCTION

This Today with extenuation of database systems and high volume stored data in this systems we need a tool so we can process the stored data and then deliver them to users. Using simple queries in SQL and various ordinary reporting tools we can deliver some information to users such that they can conclude the data and logical relations between them. But when the volume of data is high, users though industrious and thoroughbred can't recognize useful patterns among the high volume of data and if they can the cost of operation in term of man power and financial is high. Moreover, users usually pose a hypothesis and then prove or disprove the hypothesis that the observed pay reports. But today, the need for knowledge discovery methods that Idiomatically find the Knowledge, internet explorer with minimal user intervention and automatic models are able to express logical relationships. Data mining is one of the methods by which useful patterns in the data with minimum user intervention are identified, and the information can be provided to users and analysts. This information can be provided to users and analysts according to their critical decisions are made in organizations. Data mining in the exploratory analysis of the statistics of the data used in the discovery of hidden information unknown to the massive volume of data should be. Furthermore, data mining and artificial intelligence, machine learning is also closely related. Therefore data mining combines database, artificial intelligence, machine learning and statistical theories to providing functional areas. It should be noted that the term data mining is used when a large volume of data, terabytes or peta bytes in our faces. Much greater volume of data and the relationships between them is complicated by difficult access to information is embedded in the data and role of data mining as one of the knowledge discovery is more clear.

II. DATA MINING, BASIC AND CONCEPT

For understanding concept of datamining, first we should have a correct definition of some words,i.e data, information and knowledge.

- Data: any symbol, digit, character, string or signal that hasn't a special context in our mind.
- Information: if beside the data there is a string in order to describing its concept, the data is converted to information.
- Knowledge: existence of a relation between two information element shows a knoledge in that context. In a simple definition, knowledge can demonstrate the relation between informational elements.
- intellect: the upmost level of insight that is showed with signs and symbols.

A. Ddata mining definition

First, there are many definitions in doccuments for datamining. Most of them say that datamining is equivalent to mining i.e we search a valuable subject among a mass of data. In a simple definition datamining is exploiting knowledge among the mass of raw data. This naming is a bit suitable, because for example we name the mining operation for gold extracting as gold exploiting rather than sand exploiting, hence

it was better to give it a name like knowledge mining. Now we transfer to say some definition of datamining:

- Data mining is the process of extracting valid, underestandale and reliable information from large databases and using it in making decisions on important business activities.
- data mining is a semi-automated process to analyze large databases to find useful patterns is defined.
- Data mining i.e searching in a data bases to find patterns among data.
- Data mining is extracting macro knowledge, referancable and modern from large databases.
- Data mining is analyzing of visible data sets to find reliable relations among data.

B. Challenges in datamining

Maybe the most important blind spots of data mining are data existence, data healthy and data features sufficiency. The purpose of data existence is primarily there is a data for searching, but this is a broblem that there is in many real environment. Data healthy i.e real data are correct. For example the gender of e person name "Jack" hasn't entered as female. data features sufficiency i.e recorded features for each person or object is soficient for learning the model and finding the regulation of data. But our mean from challenges in data mining challenges are problems that datamining methods are faced with. They are:

- High volume of data: data mining algorithms works with many records that have collected in ;ong periods and they aren't analysable with calssic data processing methods.
- High dimention of data: The order of dimention are the fields or features. The more characteristics i.e more difficult to analyze the data.
- Distributed nature of data: target data in data mining are in various sources of data, and it cause that there are several methods that can be inefficient information processing in data mining.
- Heterogeneous nature of data:, Because the data are collected from different data sources, data that refer to the same concept, they may use different scales, so data should be homogen for correct analys.
- Quality of data: it is for times that quality of data is low, for example when data are noisy, deviated, losed or repeated.
- Has no ownership of data:for many reasons e.g. distribution we may can't own all of data together for datamining.

C. Data mining phases

Knowledge discovery of the abundance of information is a repetitive process. As indicated in Figure 2. data mining only is one of the steps of extracting information among the data collection.hence we first say a brief explanation of preparing phases name data baking and then we mention the data mining process in which the existing patterns in data collection methods using machine learning methods are primarily related to the extraction point can be.prepration steps are as follow:

1. Data cleanning(eliminating the nise and data incosistency)
2. Data integration(Several data sources are combined)
3. Data selection(retrieving the data associated with analysis form the database)
4. Data conversion(The data is converted to a form suitable for data mining e.g. matching process as outline)
5. Data mining(The main process of intelligent routines are used to extract patterns from data)
6. Pattern evaluation: (To determine the correct pattern by measuring tools)
7. Knowledge presentation(Visual representation, representation of knowledgecan be used to present the knowledge discovery techniques to users)

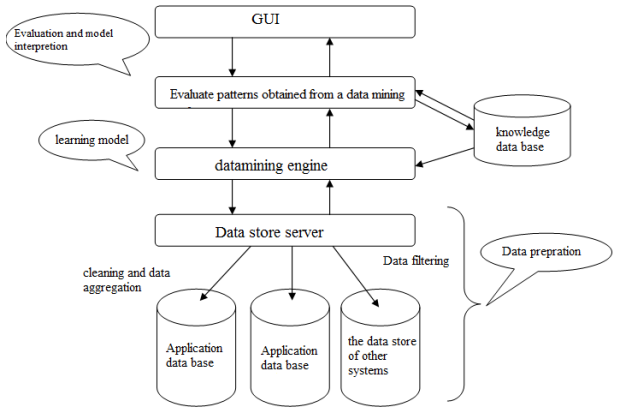

Figure 1. Architecture of a data mining system

D. Prepare and refining the data

Maybe As mentioned in the first paret of the paper data is any symbol, digit, character, string or signal that hasn't a special context in our mind. In this section we are going to study it as the most important input in data mining process. We study the data in the case of features, number of values and data sets.

1) Data dimensions

Feature dimension, data usually are kept as a set of records and each record is constructed of set of reatures and properties. In fact each property shows a reality about the record. Some concepts about the properties of a feature are:

- Character characteristic: only equal and unequal opretors can be used about this property, but greater and lower oprators can't be used to. Examples of this property are name,familiy, eye color , gender and etc.
- Sequential characteristic: equal, unequal, greater and lower, oprators can be used to this property. But arithmetic operation can't be used to. Examples of this property are date,birth date, age, weight and etc.

- Interval feature: equal, unequal, greater and lower, add and sub oprators can be used for this feature. But multiplication and division operation can't be used to. Some examples are varoius event date, environment temprature and etc.
- Rating feature: all of the operators can be used to this fearture, scores obtained by students, prcice and etc. the most important blind spots of data mining are data existence.

2) Data refinement

The operational process to produce a set of explorable refined data is prepration. This process consists of two main parts, data extraction and data preprocessing.

Data extraction: In the first phase of a comprehensive set of data collected and stored in a data warehouse. This data set can also be geographically distributed in different sources are collected. In the next step required data for exploration are collected from datastore, this increase the velocity of knowledge schooling from data. For examole in medical and health we should collect data from various resources in different cities. If we want to get a model of diabetic patientsonly thethe information of diabetic patients are extracted from this data store.

Data preprocessing: At this stage of the action takes place causing various problems have resolved the problem to be investigated. The data for the learning process of the model is refined and ready. This operation are data cleaning,filtering of the samples, sampling, data convertion, discretization and dimension decreasing. Data mining algorithms and learning after the data preprocessing step in data mining process, the data is ready to apply to the learning model. In the model learning phase step, the discipline of data pre-processing, respect to the data mining method that is chosenis identified and produced and is sent to the next step for evaluation. We examine three common method in data mining in this section, e.g. grouping, clusterin and Mining Association Rules, and check the algorithmes that are used in each of them.

3) Grouping method

In Grouping there is a label for each of the records of the data set, that indicates the fact in the the problem.this label causes the grouping algorthms are considered as supervides algorithms. In supervides algorithms the algorithm first learn a model in training phase and then the Model performance is evaluated in evaluation phase. In grouping algorithm the entire dataset is divided to training and test data sete. In the training phase of a learning algorithm can model a set of training data. The shape of the consructed model is depended to the used alogorithm. The next step was to apply the model on the test data set and its accuracy is calculated. This process do in repeatition and if accurace is improved caompared to previous model, the learning process continues. Obviously the model that learned a concept from the training data has more accuracy that the model that retains only the training data very well for the experimental data set .Types of classification algorithms can be outlined as follows:

- Methods based on decision tree
- Law based methods
- Memory-based reasoning
- Neural networks
- The methods that are based on Bayes theorem
- Support Vector Machines

4) Clustering

Clustering does not consider any record label, and records are classified to set of clusters only based on the similarity that is inroduced. No tags causes that clusterng algorithm classified as unsupervised.i.e there isn't any training and evaluation step. cluster refers to a set of data that are similar. Clustering is the most important method in unsupervised learning. There are two goals of clustering, preception and conclusion. Different clusters according to their final shape can be one of the following forms:

- well-separated clusters
- centralized clusters
- proximity clusters
- Condensation clusters
- Conceptual clusters
- Cluster-based objective function

Clustering algorithms can be studied from different aspects. Some of these measures include:

- exclusive versus Non-exclusive: Cluster non-exclusive categories for each point may belong to several clusters have, While the exclusive cluster point belongs to only one cluster.
- Fuzzy versus non-fuzzy: In fuzzy clustering, each point inside a weight will be between 1 and 0 Dashd. Clustering is the potential difference is the sum of the probabilities must equal one.
- Partial to complete: In some cases only part of the data we need to cluster
- heterogeneous versus homogeneous: cluster sizes, shapes and different densities is done.

However, the dimensions of the various species can be introduced into the most popular clustering methods as:

- Partitioning Clustering: divide the data set into non-overlapping subsets such that each data set is just a subset.
- Hierarchical clustering: a nested set of clusters into a hierarchical tree to witness the end of the clustering process.
- Density-based clustering: a data set into subsets that will be considered the density and distribution of their records.

5) Discovering the associative rules

The discovery of association rules, such as Unsupervised clustering is a descriptive technique. Used to explore the data set contains a large number of transactions that each transaction is included in several pieces. Association rules discovery algorithms to find a set of rules or associations

among these transactions are dependent. The rules say that you can discover what objects there is an impressive collection of other objects. An important output of the data mining approach, consisting of a set of rules, though - when events indicate the relationships among a set of objects together with one another. In this algorithm, a set of terms used to describe them first turn.

- Object set: the set of one or more object. K-member to the complex object is a collection object that contains K is.
- Total support: frequency of occurrence of objects in the current transaction.
- Multiple objects sets: a collection of objects that support the number is greater than or equal to a threshold.
- Associative Rules: express the relationship between a set of objects is frequent
- Support: the fraction of transactions that contain all items of a particular category.

III. K-MEANS CLUSTERING ALGORITHM

In this section first of all we will represent sequential K-Means clustering Algorithm and then Bisection K-Means so Parallel K-means and at last Parallel K-Means with prediction.

A. Sequential K-means clustering algorithm(SK)

k-means is a popular algorithm to solve the problem of clustering a data set into k clusters. If the data set contains n data points, X1, X2, . . . , Xn, of dimension d, then the clustering is the optimization process of grouping n data points into k clusters so that the following global function is either minimized or maximized.

$$\sum_{j=1}^{k}\sum_{i=1}^{n} f(X_i, c_j) \qquad (1)$$

The goal of this function f(Xi,Cj) is to optimize different aspects of intra-cluster similarity, inter-cluster dissimilarity and their combinations. For example, the Euclidean distance function minimizes the intra-cluster dissimilarity. In that case, a data point Xi is assigned to the cluster with the closest centroid Cj and the global function is minimized as a result. When the cosine similarity function is used, a data point Xi is assigned to the cluster with the most similar centroid Cj and the global function is maximized as a result. This optimization process is known as a NP-complete problem [6], and the sequential k-means (SK) algorithm was proposed to provide an approximate solution [7]. The steps of SK are as follows:

1. Select k data points as initial cluster centroids.
2. For each data point of the whole data set, compute the clustering criterion function with each centroid. Assign the data point to its best choice.
3. Recalculate k centroids based on the data points assigned to them.
4. Repeat steps 2 and 3 until convergence.

The computation complexity of SK is determined by the number of data points (n), the dimension of the data point (d), the desired number of clusters (k), and the number of loops in SK(L). For a processor, we assume all the basic operations have the same unit operation time (t). At each loop, the computation complexity of the calculation step is dominated by the clustering criterion function, which has f(n,k,d) operations. For the update step, recalculating the centroids needs kd operations. Thus, the time complexity of the whole k-means algorithm is:

$$T_{SK} = (f(n,k,d) + kd)Lt \qquad (2)$$

If the clustering criterion function is the Euclidean distance function, f (n,k,d) = 3nkd+nk+nd; and for the cosine similarity function, f(n,k,d)=2nkd+nk+nd. Under the assumption that the number of data points (n) is much larger than d and k, Eq. (2) could be rewritten as:

$$T_{SK} = FnkdLt \qquad (3)$$

where F is a constant for the clustering criterion function used. SK is very sensitive to the selection of initial centroids, and different initial centroids could produce different clustering results. For the same initial centroids, each run of SK on the same data set always has the same L.

B. Bisection K-means clustering algorithm(BK)

The bisecting k-means (BK) [13] is a variant of the k-means algorithm. The key point of this algorithm is that only one cluster is split into two subclusters at each step. This algorithm starts with the whole data set as a single cluster, and its steps are:

1. Select a cluster Cj to split based on a rule.
2. Find 2 subclusters of Cj by using the k-means algorithm (bisecting step):
    a. Select 2 data points of Cj as initial cluster centroids.
    b. For each data point of Cj , compute the clustering criterion function with the 2 centroids and assign the data point to its best choice.
    c. Recalculate 2 centroides based on the data points assigned to them.
    d. Repeat steps 2.b and 2.c until convergence.
3. Repeat step 2, I times, and select the split that produces the clusters satisfying the global function.
4. Repeat steps 1, 2 and 3 until k clusters are obtained.

I is the number of iterations for each bisecting step, which is usually specified in advance. There are many different rules that we can use to determine which cluster to split, such as selecting the largest cluster at the end of the previous iteration or based on the clustering criterion function. However, it has been reported that the differences between them are small according to the final clustering result. In this paper, we always split the largest remaining cluster. The computation complexity

of BK is determined by the size of Cj at each bisecting step (nj), the dimension of the data point (d), the desired number of clusters (k), the number of loops of k-means in each bisecting step (L), and the number of iterations for each bisecting step (I). In the bisecting step, f(nj,2,d) operations are required for the calculation step and 2d operations for the centroids updating step. Since each bisecting step produces one more cluster, total k−1 bisecting steps are needed to produce k clusters. Thus, the time complexity of BK can be represented as:

$$T_{BK} = (f(\overline{n_j}, 2, d) + 2d)\overline{L}I(k-1)t \quad (4)$$

where nj is the average size of Cj of each bisecting step, and L is the average number of loops of k-means for each iteration of a bisecting step. Under the assumption that nj is such larger than d and k, Eq. (4) could be rewritten as:

$$T_{BK} = \frac{2F}{k}\overline{n_j}d\overline{L}I(k-1)t \quad \text{when } \overline{n_j} \le n \quad (5)$$

The comparison of two Eqs. (3) and (5) shows that, when k is large, BK is even more efficient than sequential k-means. The computation of each loop in k-means involves n and k, while the computation of the iteration in each bisecting step of BK involves nj (nj ≤ n) and 2(k−1)/k ≈ 2 (when k is large).

C. Parallel K-means clustering algorithm(PK)

A parallel k-means (PK) algorithm is proposed in [5] based on the Single Program over Multiple Data streams (SPMD) parallel processing model and message-passing. PK is easy to implement and achieves a nearly linear speedup. The steps of PK are as follows:

1. Evenly distribute n data points to p processors so that each processor has n p data points in its disk, where np = n/p.
2. Select k data points as initial centroids, and broadcast them to the p processors
3. Each processor calculates the clustering criterion function for each of its n p data points with each centroid of k clusters, and assigns each data point to its best choice. (calculation step)
4. Collect all the information needed from the p processors to update the global cluster centroides, and broadcast then to the p processors. (update step)
5. Repeat steps 3 and 4 until convergence.

The strategy of this algorithm is to exploit the data-parallelism of the sequential k-means (SK) algorithm. Step 2 of SK shows that the calculation of the clustering criterion function for different data points could be done at the same time without affecting the final result. Thus, by distributing n data points to p processors, these processors can execute the calculation step on n/p data points at the same time. However, the trade-off is the communication time in step 4 of PK. At the end of each loop, the information of k centroids of dimension d is collected and then broadcast to p processors for the next loop's calculation step. The communication time is determined by k and d as:

$$T_{comm} = MdkL \quad (6)$$

where M is the unit time required to collect and broadcast from/to p processors for a floating point number. When the implementation language and the system are determined, M is a constant. It is reported that, for most architectures, M is associated with O(log p) [4]. The cost of broadcasting the initial centroids is ignored because it is constant for the same k and d, regardless of n, and it is relatively very small compared to the costs of other steps. Based on the time complexity of SK, the time complexity of PK could be represented as:

$$T_{PK} = \frac{Fnkd}{p}Lt + MdkL \quad (7)$$

Figure 2 shows how multiple processors work together to achieve the speedup. If the step 2 of SK takes t2 seconds, then the step 3 of PK takes t2/p seconds beacuse of parallization. Even though the step 2 of PK takes a little longer than the step 1 of SK because of the communication cost, the difference could be ignored if the time of the calculation step is relatively long. The difference between the step 4 of PK and the step 3 of SK can be explained in the same way.

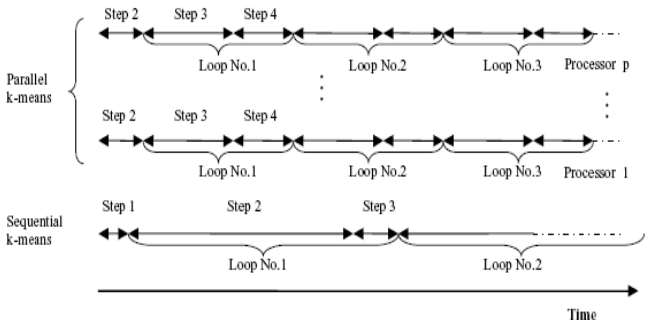

Figure 2. Comparison of sequential and parallel k-means algorithms

D. Parallel bisecto K-means clustering algorithm(PBK)

Like the parallel k-means (PK) algorithm, we can design a parallel bisecting k-means (PBK) algorithm based on the SPMD parallel processing model and message-passing. The steps of PBK are as follows:

1. Evenly distribute n data points to p processors.
2. Select a cluster Cj to split based on a rule, and broadcast this information to all processors.
3. Find 2 subclusters of Cj using the k-means algorithm (bisecting step):
   a. Select 2 data points of Cj as initial cluster centroids and broadcast them to the pj processors that have data members of Cj .
   b. Each processor calculates the clustering criterion function for its own data points of Cj with 2 centroids, and assigns each data point to its best choice. (calculation step)
   c. Collect all information needed to update 2 centroids, and broadcast them to the pj processors participating in the bisecting. (update step)

d. Repeat steps 3b and 3c until convergence.

4. Repeat steps 2 and 3 I times, and select the split that produces the clusters satisfying the global function.
5. Repeat steps 2, 3 and 4 until k clusters are obtained.

Let's discuss the computation complexity of PBK. Only the largest remaining cluster Cj is split at each bisecting step. Cj has nj data points which may not be evenly distributed over the p processors. For each bisecting step, each of the pj processors can execute the calculation step on njp data points at the same time, where njp represents the number of data members of Cj allocated to a processor. Obviously, the number of data points that each processor is working on may be different. And for each processor, the value of njp for each bisecting step may change, too. For each bisecting step, the largest njp determines the time of the calculation step because, in order to start the update step, every processor must wait until all the other processors complete the calculation step. The calculation time of each bisecting step is:

$$T_{comp} = \frac{2F}{k} \max(n_{jp}) d\overline{L} I t \tag{8}$$

At the end of each iteration, the information of 2 centroides of dimension d is collected and broadcast to the pj processors for the calculation of next iteration. Those processors that have no data point of the selected cluster Cj do not participate in this update step. The communication time of each bisecting step can be represented as:

$$T_{comm} = 2M_j d\overline{L} I \tag{9}$$

E. Parallel bisecting k-means with prediction (PBKP) alg.

In order to improve the processor utilization, we propose a new algorithm, named Parallel Bisecting k-means with Prediction (PBKP). PBKP tries to predict the future bisecting step. At each bisecting step, instead of splitting one largest cluster, it splits two largest clusters. In this way, the processor utilization is improved and the number of bisecting steps is reduced. As a result, the speedup is increased and the total execution time is shortened. It has been reported that bisecting k-means tends to produce the clusters of similar sizes [13], where the size of a cluster is the number of data points in the cluster. When the size of a selected cluster is S, the sizes of its two subclusters are usually around S/2. We can take advantage of this characteristic by modifying the simple PBK. First, let's look at the example of BK shown in Fig.4. An initial cluster A has 20 data points, which is represented as a box with the cluster name and size in it. Each arrow represents a bisecting step, and the associated label shows the order of the step. Figure 4.a shows the case of original BK algorithm. At the first bisecting step, cluster A is split into clusters B and C. Since BK tends to produce clusters of similar sizes, if C.size<B.size < 2C.size, then it is quite probable that, after cluster B is split at the second bisecting step, cluster C will be split at the third bisecting step by assuming two subclusters of B would not be larger than C.

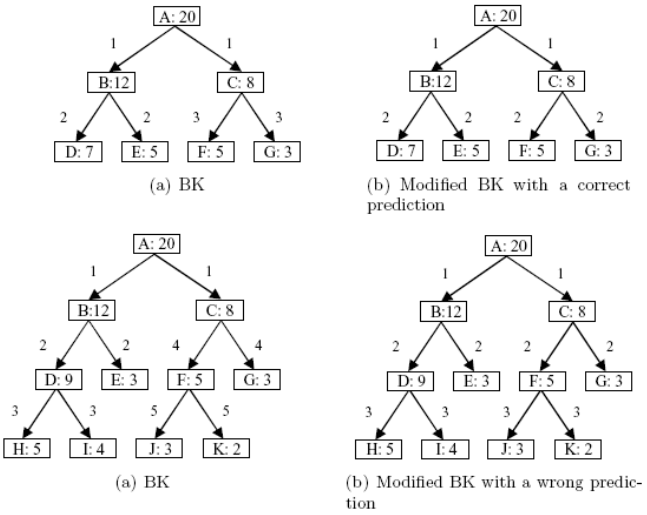

Figure 3. Bisecting k-means with correct and wrong prediction

Figure 3.b shows the case that we predict cluster C is the one to be split at the third bisecting step and do it one step ahead. So, clusters B and C are split into two subclusters, respectively, at the second bisecting step. Here, C.size < B.size < 2C.size is a prerequisite of our prediction step. In other words, the prediction step could be performed only under this condition. This example shows that splitting the largest cluster and the second largest cluster at one bisecting step does not change the final clustering result, whereas one bisecting step is reduced.

However, there is no guarantee that our prediction is always correct. The example in Fig.4 shows this situation. Fig.4.a shows that cluster C is split at the fourth bisecting step because one subcluster of cluster B is larger than C. In this case, our prediction that cluster C would be split at the third step is wrong. Even so, Fig.4.b shows that splitting C at the second step with B does not change the final clustering result, either, while the total number of bisecting steps is reduced from 5 to 3.

There may be a case that some descendant of cluster B is always larger than cluster C, so that C will not be selected to split, but the chance is very slim. Our experimental results show that the quality of the final clustering of PBKP is as good as that of BK. The key point of PBKP is to split two clusters, instead of one cluster, at each bisecting step. As more clusters are split at each bisecting step, more data points are involved in the calculation step, and the distribution of those data points over the processors usually becomes more uniform. Thus, the processor idle time is reduced. Moreover, the reduced number of bisecting steps helps to reduce the communication cost of the parallel algorithm. The steps of PBKP are as follows:

1. Evenly distribute n data points to p processors.
2. Select the largest cluster Cj and the second largest cluster C'j from the remaining clusters to split if Cj.size < 2C'j.size (prediction step). Otherwise, select only the largest cluster to split. Broadcast this information to all processors.

3. Find 2 subclusters of Cj and C'j , respectively, by using the k-means algorithm (bisecting step):
    a. Select 4 data points as initial cluster centroids and broadcast them to the pj processors that have data members of Cj and C'j .
    b. Each processor calculates the clustering criterion function for its own data points of Cj and C'j with 2 sets of 2 centroids, respectively, and assigns each data point to its best choice (calculation step).
    c. Collect all information needed to update 4 centroids, and broadcast them to the pj processors (update step) .
    d. Repeat steps 3.b and 3.c until convergence.
4. Repeat steps 2 and 3 I times, and select the split that produces the clusters satisfying the global function.
5. Repeat steps 2, 3 and 4 until k clusters are obtained.

Since the subcluster sizes of each bisecting step are very similar in most cases, usually Cj.size<2C'j.size holds, hence both Cj and C'j are bisected.

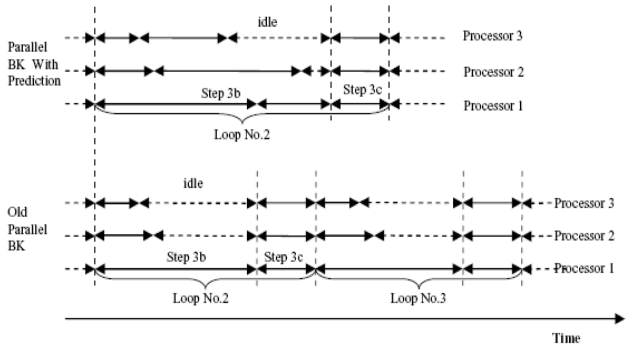

Figure 4. Time diagram of PBK and PBKP algorithms

Fig.4 compares the timing diagrams of PBK and PBKP. In order to simplify the comparison, we assumed I is 1. The prediction step reduces the total execution time in two ways. First, it reduces the processor idle time. Since two clusters' data points are usually more uniformly distributed over the processors than one cluster's, bisecting two clusters, one by one in two steps, may take longer time than splitting two in one step. Second, it reduces the total number of bisecting steps from k − 1 to k/2.

The calculation time of each bisecting step is:

$$T_{comp} = \frac{2F}{k} \max(n_{jp} + n'_{jp}) d\overline{L} I t \qquad (10)$$

where n_{jp} and n'_{jp} represent the numbers of data members of C_j and C'_j, respectively, which are allocated to a processor. The range of $\max(n_{jp}+n'_{jp})$ is $[(n_j+n'_j)/p,(n_j+n'_j)]$. At the end of each iteration, the information of 4 centroids of dimension d is collected and broadcast to the pj processors for the next loop's calculation step. The communication time of each iteration is:

$$T_{comm} = 4M_j d\overline{L} I \qquad (11)$$

Thus, the time complexity of PBKP could be represented as:

$$T_{PBKP} = F\overline{\max(n_{jp} + n'_{jp})} d\overline{L} I t + 2\overline{M_j} d\overline{L} I k \qquad (12)$$

IV. CONCLUSION

As matter of fact in this paper, we introduce basic fundamentals aspect of data mining. So present three main categories of data mining techniques. Which there are Grouping, Clustering and Associative relationship between data items. This techniques can use with supervision and unsupervision. Although Data mining subject proposed in 80th but in 90th we had main progress in data mining. Despite of recent advancement in data mining we encounter with lots of problems in real environment. Because, the resources of data may be separated from each other. Also these resource may don't have uniform structure and be in deferent owned with deferent access roles. In other hand, privacy protection of data resource owners is one of main subjects in data mining. When the amount of data which stand in data where house growth dramatically, the early data mining techniques will not be useful and we need to use parallelism solutions. Data segmentation is one of key steps in this solutions.

REFERENCES

[1] Pang-Ning Tan "Book : Introduction to data mining" Pearson Addison Wesley ISBN 0321-42052-7, 2006.

[2] Jiawei Han , Micheline Kamber "Data Mining concept and technique" Morgan Kaufmann Elsevier Inc. ISBN 978-1-55860-901-3, 2006

[3] A. Hotho, S. Augustin A Brief Survey of Text Mining School of Computer Science Otto-von-Guericke-University Magdeburg May 13, 2005

[4] U. Nahm and R. Mooney. Text mining with information extraction. In Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases, 2002.

[5] A. Garg ,A. Mangala, Neelam Gupta, "PBIRCH:A Scalable Parallel Clustering Algorithm for Incremental data", 10th International Database Engineering and Applications Symposium (IDEAS'06), 0-7695-2577-6/06 IEEE, (2006).

[6] Kittisak Kerdprasop and Nittaya Kerdprasop, "A lightweight method to parallel k-means clustering," International Journal of Mathematics and Computers in Simulation,Volume 4, (2010).

[7] S.V. Adve, "Parallel Computing Research at Illinois: The UPCRC Agenda," 2008 .

[8] K.Kerdprasop and N.Kerdprasop, "A lightweight method to parallel k-Means clustering," International Journal of Mathematics and Computers in Simulation, Volume 4, pp 144-153, (2010).

[9] A. Raghuvira Pratap, J. Rama Devi, K. Nageswara Rao, "An Efficient Density based Improved K- Medoids Clustering algorithm," International Journal of Advanced Computer Science and Applications(IJACSA), Vol. 2, No. 6, ,pp 49-54, (2011).

[10] David Pettinger and Giuseppe Di Fatta "Space Partitioning for Scalable K-Means," Ninth International Conference on Machine Learning and Applications, pp 319-24, IEEE (2010).

Akbar Momeni received B.E. and M.E. degree in computer software engineering from Islamic Azad University (IAU), Iran in 2000 and 2007 respectively. Now he is PHD student in SRBIAU, Tehran, Iran. He also works as a university lecturer in Basic and Mathematics science at TIAU. He has worked for Ansar Bank (Credit and Finance Institute) as head if IT manager from 1999 to 2007. His research interests include Software engineering, Security assurance in SDLC, Wireless sensor network, Mobile ad-hoc network and security issues in these networks.

Karim Khazaei received the B.S. and M.S. degrees from Qazvin Islamic Azad University, Iran, in 2002 and 2007, respectively. he is currently pursuing her research in Cognitive Radios towards completion of her PhD from Science and Research Azad University, Tehran, Iran. He has a total academic experience of 6 years. At present he is ASSISTANT PROFESSOR (Senior Grade) in electronics and Computer department, Islamic Azad University, Takestan. Iran. His recent research interests include Sensor Networks, Ad-hoc Networks and Cognitive Radios. He has published over twelve journal and conference papers in these areas.