Survey on Efficient Algorithms to Improve the Clustering Performance in Partition and Grid Based Clustered Environment

Mr.D.Rajesh Kumar[1], Ms.S.Malathy[2], Ms. D. Saranya[3], Ms. S. Naveena[4]

Assist. Professor[1,2,3,4]

Department of Information Technology,

Bannari Amman Institute of Technology, Sathyamangalam -638401, India.

E-mail: **sangeraje@gmail.com[1]**, ksmalathy@gmail. com[2], saranya.d@butsathy.ac.in[3], naveena@bitsathy.ac.in

Abstract—The Clustering is an essential task in data mining to discover and explore groupings within data or entities. Clustering methods allows entities to be partitioned into distinct groups called segments. The learning is the process of generating useful data from a huge volume of data which can be classified as supervised learning and unsupervised learning. Clustering technique is a kind of supervised learning. A pattern representing a common behavior or taxonomies that exist among each item can be generated. This paper gives an overview of different partition and grid based clustering algorithm.

Keywords— Clustered Grid computing, Partition Clustering Technique, Supervised Learning, Unsupervised Learning.

I. INTRODUCTION

THE 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. Cluster analysis is used in a number of applications such as data analysis, image processing, market analysis etc. Clustering helps in gaining, overall distribution of patterns and correlation among data objects. This paper gives good inputs about the general working behavior, the methodologies to be followed and the constraints which affect the performance of the partition and grid based clustering algorithms in Datamining.

II. CLUSTERING OVERVIEW

Clustering algorithm can be divided into the following categories:

    A. Hierarchical clustering algorithm
    B. Partition clustering algorithm
    C. Spectral clustering algorithm
    D. Grid based clustering algorithm
    E. Density based clustering algorithm

A. Hierarchical Clustering Algorithm

The Hierarchical clustering algorithm groups data sets and objects to form a tree shaped structure. It can be broadly classified into agglomerative hierarchical clustering and divisive hierarchical clustering. 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, centroid or wards method. In divisive approach all data points are considered as a single cluster and they are splited into number of clusters based on certain criteria, and this is called as top down approach. In each successive iteration, a cluster is split into smaller clusters, until each object is in one cluster. Examples for this algorithms are LEGCLUST, BRICH (Balance Iterative Reducing and Clustering using Hierarchies), CURE (Cluster Using Representatives) and Chameleon.

B. Spectral Clustering Algorithm

Spectral clustering refers to a class of techniques which relies on the Eigen structure of a similarity matrix. Clusters are formed by partition data points using the similarity matrix. Any spectral clustering algorithm will have three main stages. They are

 ➢ Preprocessing: Deals with the construction of similarity matrix.
 ➢ Spectral Mapping: Deals with the construction of eigen vectors for the similarity matrix
 ➢ Post Processing: Deals with the grouping data points

The following are advantages of Spectral clustering algorithm:

 ➢ Strong assumptions on cluster shape are not made.
 ➢ Simple to implement.
 ➢ Objective does not consider local optima.
 ➢ Statistically consistent.
 ➢ Works faster.

The major drawback of this approach is that it exhibits high computational complexity. For the larger dataset it requires $O(n3)$ where n is the number of data points.

C. Grid based Clustering Algorithm

Grid based algorithm quantize the object space into a finite number of cells that forms a grid structure [1]. Operations are done on these grids. The advantage of this method is lower processing time.

Clustering complexity is based on the number of populated

grid cells and does not depend on the number of objects in the dataset. The major features of this algorithm are:

1. No distance computations.
2. Clustering is performed on summarized data points.
3. Shapes are limited to union of grid-cells.
4. The complexity of the algorithm is usually O(Number of populated grid-cells)

STING [1] is an example for this algorithm.

The main advantage of this approach is its Fast processing time, which is typically independent of number of data objects and dependent only on number of cells.

DBSCAN, DENCLUE and OPTICS [1] are examples for this algorithm.

D. Density based Clustering Algorithm

Most partitioning objects cluster objects based on the distance between objects. Other Clustering methods have been developed based on the notion of density. Density based algorithm continue to grow the given cluster as long as the density in the neighborhood exceeds certain threshold. This algorithm is suitable for handling noise in the dataset. The following points are enumerated as the features of this algorithm.

1. Handles clusters of arbitrary shape
2. Handle noise
3. Needs only one scan of the input dataset.
4. Needs density parameters to be initialized.

DBSCAN, DENCLUE and OPTICS [1] are examples for this algorithm.

E. Partition Clustering Algorithm

Partition clustering algorithm splits the data points into k partition, where each partition represents a cluster. The partition is done based on certain objective function. One such criterion functions is minimizing square error criterion which is computed as,

$$E = \Sigma \Sigma \ || \ p - mi \ // \ 2 \ (1)$$

Where p is the point in a cluster and mi is the mean of the cluster. The cluster should exhibit two properties, they are (1) each group must contain at least one object (2) each object must belong to exactly one group. The main drawback of this algorithm [3] is whenever a point is close to the center of another cluster, it gives poor result due to overlapping of data point

F. K-Means Algorithm

The algorithm classifies objects to a pre-defined number of clusters, which is given by the user (assume k clusters). The idea is to choose random cluster centers, one for each cluster. These centers are preferred to be as far as possible from each other. Starting points affect the clustering process and results. After that, each point will be taken into consideration to calculate similarity with all cluster centers through a distance measure, and it will be assigned to the most similar cluster, the nearest cluster center. When this assignment process is over, a new center will be calculated for each cluster using the points in it. For each cluster, the mean value will be calculated for the coordinates of all the points in that cluster and set as the coordinates of the new center. Once we have these k new centroids or center points, the assignment process must start over. Finally, this algorithm aims at minimizing an objective function, which is in this case a squared error function. One drawback of KM is that it is sensitive to the initially selected points, and so it does not always produce the same output. Furthermore, this algorithm does not guarantee to find the global optimum, although it will always terminate.

1. Choose random k points and set as cluster centers.
2. Assign each object to the closest centroid's cluster.
3. When all objects have been assigned, recalculate the positions of the centroids.
4. Go back to Steps 2 unless the centroids are not changing.

G. Genetic Algorithms

Genetic algorithms were inspired from Darwin's theory of evolution and were pioneered by John Holland. A genetic algorithm can be defined as a search algorithm based on the mechanics of natural selection and natural genetics [7], or as software and procedures modeled after genetics and evolution [1]. Genetic algorithms have at least the following elements in common: Populations of chromosomes, selection according to fitness, crossover to produce offspring, and random mutation of a new offspring [2]. In further details, the algorithm starts with a population of "individuals", each representing a possible solution to a given problem. Each possible solution within the population of a biological individual is coded in a so-called chromosome. Each chromosome (sequences of genes) is assigned a "fitness" according to how good a solution is to the problem based on a given fitness function. The solutions (individuals) are selected into the process according to their fitness, specifically those that follow the principles first laid down by Charles Darwin of the survival of the fittest for reproduction by "cross breeding" with other individuals in the population and used to construct new individuals as offspring with a hope that the offspring will be fit better than the old individuals and a generation is complete [12]. This process is repeated until certain criteria are met. Fig. 2 shows the basic steps for GAs [7]. In Fig. 2, t represents the generation number, and P stands.

    t = 0;
    Initialize P(t);
    Evaluate P(t);

IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 5, July 2014.
www.ijiset.com

ISSN 2348 – 7968

```
While not (termination condition)
    begin
    t=t+1;
    Select P(t) from P(t - 1);
    Recombine pairs in P(t);
    Mutate P(t);
    Evaluate P(t);
End
```

Goldberg's Pseudo-code of Genetic Algorithms, it represents the generation number, and P stands for population. The first population is initialized by coding it into a specific type of representation (i.e. binary, decimal, float, etc) then assigned to a cluster. Fitness is calculated in the evaluation step. While the termination condition is not met, which might be number of generations or a specific fitness threshold, the processes of selection, recombination, mutations and fitness calculations are done. Selection process chooses individuals from population for the process of crossover. Recombination (or crossover) is done by exchanging a part (or some parts) between the chosen individuals, which is dependent on the type of crossover (Single point, Two points, Uniform, etc). Mutation is done after that by replacing few points among randomly chosen individuals. Then fitness has to be recalculated to be the basis for the next cycle. This is the general form for GAs.

H. Genetic K-Means Clustering Algorithm for Mixed Numeric and Categorical Dataset

Genetic K-means algorithm (GKA) is the combination of genetic algorithm and K-means algorithm. GKA is suitable for numeric data set which can be overcome by Genetic K-means clustering algorithm for mixed dataset. It uses an enhanced cost function to handle the categorical data.

a. Methodology
  ➢ Objective function: Defines the objective function( ) for mixed data type (2) where distance of data d_i from the closest center c_j
  ➢ Selection: Calculates the probability distribution function (Pz) using the fitness value F(Sz). Pz = (3) where Sz denotes the solution.
  ➢ Mutation: Mutation is performed to achieve global optimum using the probability distribution function Pk. Pk = (4) where d(Xn,Ck) is the Euclidean distance between pattern Xn and the centroid Ck ,represents the maximum distance for the pattern Xn
  ➢ Finally convergence is obtained by applying K-means operator.

b. SCALE Algorithm
SCALE[5] is a framework designed for the transactional dataset. The similarity measures used for the categorical data items for transaction dataset are Weighted Coverage Density (WCD). It is the measure of coverage density (CD). WCD preserves as many frequent items as possible within the cluster. It controls the overlapping of items among clusters. Data items are filled in a 2D grid. The problem of clustering is to minimize the unfilled number of cells with appropriate number of partition. Once the clusters are formed they are evaluated using two measures (i)LISR (Large Item Size ratio) and (ii) AMI (Average pair cluster Merging Index). LISR makes use of the percentage of large items in the clustering results to evaluate the clustering quality. AMI applies coverage density to indicate the structural difference between clusters.

1. Methodology
  ✓ Sampling: Huge volume of data in the transactional database are sampled into smaller groups
  ✓ Cluster structure assessment: Generate the candidate cluster based on the sample data set.
  ✓ Clustering: Done in three phases: cluster structure assessment phase, WCD based initial cluster assessment phase and iterative clustering refinement phase. Weighted Coverage Density is defined as:
$$WCD = (5)$$
where Occur(Ijk) represents the occurrence of the item Ijk, represents the number of transaction, represents the sum of occurrence of all items in cluster Ck.
  ✓ Evaluation: The algorithm is evaluated by two measures. LISR (Large Item Size ratio) and AMI(Average pair cluster Merging Index)
$$LISR = (6)$$
$$AMI = (7)$$
Where represents indicator function, represents minimum support, Mk represents number of items, and Di represents the dissimilarity measures. The author used two synthetic dataset and three real dataset (Zoo, mushroom and retail). The results of SCALE are compared with CLOP algorithm and found to be better with respect to the following features
  • Time spent on computing values for tuning parameter is reduced,
  • Handling larger dataset
  • Producing good quality cluster for domain specific measure.

I. Harmony K-Means algorithm

Harmony K-means [6] algorithm is based on the Harmony search optimization method which is proven using the finite Markov chain theory. It is a Meta heuristic algorithm and it achieves global optimum.
The advantages of this algorithm are,
  ➢ Needs only few mathematical prerequisite,
  ➢ Uses stochastic random searches,

IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 5, July 2014.
www.ijiset.com

ISSN 2348 – 7968

> Treats continuous variable without any loss of precision,
> Does not demand for initializing the decision variable
> Encoding or decoding is not performed for the decision variable.

a. Methodology
> Initialize the parameters Harmony Memory Size (HMS), Pitch Adjusting Rate (PAR), Number of Improvisation
> (NI), Harmony Memory considering Rate (HMCR).
> Generate a new harmony vector
> Update the harmony memory
> Check for the stopping criterion.
> Terminate the process when maximum numbers of improvisations are reached.
> The author experiment the algorithm using five dataset (three from TRECO, one from DMOZ and one from newsgroup) and compared the result with the K-means using the Euclidean and cosine correlation measures. He also compared with the GA, PSO, Mises-Fisher Generative Model based algorithm (GM).The main focus on the quality and speed of convergence

J. Initializing K–Means using Genetic Algorithm

Initializing K-means using Genetic algorithm [8] overcomes the initializing problem of the K-means algorithm by using the genetic algorithm. This approach solves the blind search problem of the K-means algorithm.

a. Methodology
> Initialization of initial population P0
> Repeat the step 3 through step 6 until termination
> condition is met
> Crossover or recombination is done
> Mutation is performed
> Evaluation is performed
> Apply K-means algorithm.

The author tested the algorithm using four different dataset chosen from MATLAB. The algorithms used for comparison were K-means, Genetic algorithm and Genetic algorithm initializing K-means (GAIK).

The trade off between average error rate and average time between these algorithms were listed by the author

K. Cluster Ensemble approach for mixed data

Dataset with mixed data type are common in real life. Cluster Ensemble [9] is a method to combine several runs of different clustering algorithm to get a common partition of the original dataset. In the paper divide and conquers technique is formulated. Existing algorithm use similarity measures like Euclidean distance which gives good result for the numeric attribute. This will not work well for categorical attribute. In the cluster ensemble approach numeric data are handled separately and categorical data are handled separately. Then both the

results are then treated in a categorical manner. Different types of algorithm used for categorical data are K-Modes, K-Prototype, ROCK [22] and squeezer algorithm. In K-Mode the total mismatch of categorical attributes of two data record is projected. The Squeezer algorithm yields good clustering result, good scalability and it handles high dimensional data set efficiently.

a. Methodology
> Splitting of the given data set into two parts. One for numerical data and another for categorical data
> Applying any one of the existing clustering algorithms for numerical data set
> Applying any one of the existing clustering algorithms for categorical data set
> Combining the output of step 2 and step 3
> Clustering the results using squeezer algorithm

The credit approval and cleve (heart diseases) dataset are used and measures the cluster accuracy and cluster error rate. The cluster accuracy 'r'is defined by

$$r = (8)$$

where K represents number of clusters, ai represents number of instance occurring in both the cluster I and its class and n represents number of instance in the dataset . Finally cluster error rate 'e' defined by

$$e = 1 - r (9)$$

where r represents cluster accuracy. The algorithm is compared with k-prototype algorithm.

L. Hybrid Evolutionary algorithm

The Hybrid Evolutionary algorithm [13] is based on the combination of Ant colony Optimization and Simulated Annealing (ACO-SA). The one of the major draw backs of K-means algorithm is, the result depends on the initial choice of the cluster center. In order to achieve global optima the initial choice of cluster center can be chosen with the help of ACO and SA. Ants are insects which find the shortest path between its nest and food with the help of Pheromone. Pheromone is a chemical substance deposited by ant used for the communication between them. The two most important factors used to find the shortest path are (i) the intensity of Pheromone and (ii) the path length.

Annealing is a process by which the liquid freeze or metal recrystalize. Ant colony Optimization is used for finding the colony between the data points. Simulated Annealing is used as a good local search algorithm for finding the best global position by using the Cumulative Probability.

a. Methodology
> Generation of initial population and trial intensity for the given dataset
> Selection of best local position using simulated annealing algorithm for the ith

IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 5, July 2014.
www.ijiset.com

ISSN 2348 – 7968

colony
➢ Selection of best global position for the ith colony
➢ Repeat step 2 and step 3 for all the colonies.
➢ Checks for the convergence condition C, if convergence is satisfied, stop the process. Otherwise repeat the process from step 2. The formula for convergence is

$$C = 2 \ (10)$$

where Xi represents position of ith individual, k represents iteration and N represents number of objects. The author compares ACO-SA with ACO, SA and K-means algorithm using six real life datasets (Iris, Wine, Vowel, CMC, Wisconsin breast cancer and Ripley's glass). The algorithm provides global optimum with smaller standard deviation and few functional evaluations.

M. Modified global K-means algorithm

The goal of the modified global k-means MGKM) [10] algorithm is to obtain global minima and to work well with the larger gene expression dataset. The author proposed modified global K-means algorithm. The paper focuses on the hard unconstrained partition clustering problem and to improve global search properties. The Global K-means computes the cluster successfully. During its first iteration the centroid for a set is computed. To compute k-partition at the kth iteration it uses the centers of the k-1 cluster from the previous iteration. Global K-means can be applied for a smaller dataset but it is not suitable for average or larger dataset because of high computational time. The problem can be avoided by using the squared distance concept to find the closest cluster center among k-1 clusters.

   a. Methodology
➢ Initialization is done by selecting a tolerance value greater than zero and computing the center x1
➢ Computation of cluster center for k partition is done
➢ Refinement of cluster center by using the objective function

$$f(\ x1,...xk) = 2 \ (11)$$

where xj represents the data point, ai represents cluster centroid of ith cluster.
➢ Check for stopping criteria, if not repeat the process from step 2.

$$\text{Stopping criteria} = \ < \varepsilon \ (12)$$

The author compares the MGKM algorithm with the multi-start K-means (MSKM) and global K-means (GKM) on six different gene expression datasets. The parameters for comparison are, number of cluster (N), CPU time (t) and cluster function (f). The MGKM outperforms the other algorithm as the number of cluster increases.

N. FGKA

Fast Genetic K-means algorithm (FGKA) [10] is inspired from the Genetic K-means Algorithm (GKA). Both GKA and FGKA achieve global optimum, still FGKA runs much faster than the GKA. The algorithm starts with the initialization phase with an initial population P0. Evolution of next population will be done with the selection, mutation and K-means operators which are done in sequence until a termination condition is met. The objective of selection operation is to find the population having greatest fitness value and assign smaller fitness value for illegal strings. The objective of mutation operation is to achieve global optimum. It generates positive probability and makes the pattern to move closer to the cluster and provides a legal solution.

III. CONCLUSION

In this survey we had projected various methodologies, terms, limitations, advantages and available recent innovation in with partition and grid based clustering algorithms. We hope, that the interested readers will have broad overview of this field and several starting point for further details. Also 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. This remains and will continue to be an active field that is incessantly rejuvenating itself to answer new challenges

IV. FUTURE ENHANCEMENT

Some interesting research topics in clustering of potential impact in the near future. Clustering for ultrahigh dimensional data selecting features on data sets with millions of features in Partition and Grid based clustering using concepts generalized from exiting algorithms and knowledge.

REFERENCES

[1] Jiawei Han, Micheline Kamber, "Data Mining Concepts and Techniques" Elsevier Publication.
[2] P.J. Flynn, A.K. Jain, M.N. Murty,1999. Data Clustering: A Review. ACM Computing Surveys, vol.31, no. 3: 264-323
[3] P. Berkhin,2002. Survey of Clustering Data Mining Techniques. Ttechnical report, Accrue Software, San Jose, Cailf.
[4] Dharmendra K Roy and Lokesh K Sharma, 2010. Genetic K-Means clustering Algorithm for mixed numeric and categorical data. International journal of Artificial Intelligence& ApplicationsVol1 No2.
[5] K. Chen, L. Liu, H. Yan, Z. Yi, 2010. SCALE: a scalable framework transactional data. Data Mining Knowledge Discovery. 20:1–27
[6] H. Abolhassani, M. Mahdavi, 2009. Harmony K-means algorithm for document clustering. Data Mining Knowledge Discovery.18:370–391.
[7] BB Firouzi, T Niknam, M Nayeripour, 2008. An efficient hybrid evolutionary algorithm for cluster analysis. World Applied Sciences Journal:300-30.
[8] B Al-Shboul, SH Myaeng, 2009. Initializing K-Means using Genetic Algorithms. World Academy of Science, Engineering and Technology.
[9] S Deng, Z He, X Xu, 2005. Clustering mixed numeric and categorical data: A cluster ensemble approach. Arxiv preprint cs/0509011.
[10] AM Bagirov, K Mardaneh, 2006. Modified global k-means algorithm for clustering in gene expression data sets. ACM International Conference Proceeding on Intelligent systems for bioinformatics - Volume 73: 23-27

IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 5, July 2014.
www.ijiset.com

ISSN 2348 – 7968

[11] Y Deng, Lu, F Fotouhi, S Lu, 2004. FGKA: a Fast Genetic K-means Clustering Algorithm. Proceedings of the ACM symposium on Applied computing: 622 - 623

[12] C Ding, X He, 2004. K-means clustering via principal component analysis. ACM twenty-first International Conference Proceeding on Machine learning : 29

[13] B.Bahmani Firouzi, T.Niknam, M.Nayeripour, 2008. A New Evolutionary Algorithm for Cluster Analysis. International Journal of Computer Science.

[14] Michael Laszlo Sumitra Mukherjee, 2007. A genetic algorithm that exchanges neighboring centers for k-means clustering. Pattern Recognition Letters, Volume 28: 2359-2366

[15] K. Krishna and M. Murty, 1999. Genetic K-Means Algorithm. IEEE Transactions on Systems, Man, and Cybernetics vol. 29, NO. 3: 433-439.

[16] Von Luxburg U, 2007. A Tutorial on Spectral Clustering, Statistics and Computing, 17(4):395-416.

[17] Cai X. Y. et al, 2008. Survey on Spectral Clustering Algorithms. Computer Science:14-18

[18] F. R. Bach and M. I. Jordan, 2006. Learning spectral clustering, with application to speech separation . Journal of Machine Learning Research : 1963–2001.

[19] L. Huang, M. I. Jordan, D. Yan, 2009. Fast approximate spectral clustering. Technical report, Department of Statistics, UC Berkeley.

[20] M. Livny, R.Ramakrishnan, T. Zhang, 1996. BIRCH: An Efficient Clustering Method for Very Large Databases. Proceeding ACMSIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery: 103-114.

[21] S. Guha, R. Rastogi, and K. Shim, 1998. CURE: An Efficient Clustering Algorithm for Large Databases. Proc. ACM Int'l Conf. Management of Data : 73-84.

[22] S. Guha, R. Rastogi, and K. Shim, 2000. ROCK: A Robust Clustering Algorithm for Categorical Attributes. Information Systems, vol. 25, no. 5 : 345-366.

[23] Santos, J.M, de Sa, J.M, Alexandre, L.A , 2008. LEGClust- A Clustering Algorithm based on Layered Entropic subgraph. Pattern Analysis and Machine Intelligence, IEEE Transactions : 62-75.

[24] M Meila, D Verma,2001. Comparison of spectral clustering algorithm. University of Washington,Technical report.