ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

A Comparative Study of Issues in Big Data Clustering Algorithm with Constraint Based Genetic Algorithm for Associative Clustering

B.Kranthi Kiran[1], Dr.A Vinaya Babu[2]

Assistant Professor, Department of Computer Science and Engineering, JNTUHCEJ, Karimnagar, Telangana, India[1]

Professor, Department of Computer Science and Engineering, JNT University Hyderabad, Telangana, India[2]

ABSTRACT: Clustering can be defined as the process of partitioning a set of patterns into disjoint and homogeneous meaningful groups, called clusters. The growing need for distributed clustering algorithms is attributed to the huge size of databases that is common nowadays. The task of extracting knowledge from large databases, in the form of clustering rules, has attracted considerable attention. Distributed clustering algorithms embrace this trend of merging computations with communication and explore all the facets of the distributed computing environments. Ensemble learning is the process by which multiple models, such as classifiers or experts, are strategically generated and combined to solve a particular computational intelligence problem. An important feature of the proposed technique is that it is able to automatically find the optimal number of clusters (i.e., the number of clusters does not have to be known in advance) even for very high dimensional data sets, where tracking of the number of clusters may be highly impossible. The proposed Optimal Associative Clustering algorithm using genetic algorithm and bayes factor for precision is able to outperform two other state-of-the-art clustering algorithms in a statistically meaningful way over a majority of the benchmark data sets. The result of the proposed optimal associative clustering algorithm is compared with one existing algorithm on two multi dimensional datasets. Experimental result demonstrates that the proposed method is able to achieve a better clustering solution when compared with existing algorithms.

KEYWORDS: Distributed Clustering, Ensemble Learning, Associative clustering, genetic algorithm, multi-dimensional data, bays factor, contingency table.

I. INTRODUCTION

Distributed computing and data mining are nowadays almost ubiquitous[1,2]. Authors propose methodology of distributed data mining by combining local analytical models (built in parallel in nodes of a distributed computer system) into a global one without necessity to construct distributed version of data mining algorithm. Different combining strategies for clustering and classification are proposed and their verification methods as well. Data mining means extracting hidden, previous unknown knowledge and rules with potential value to decision from mass data in database [1,2]. Association rule mining is a main researching area of data mining area, which is widely used in practice. With the development of network technology and the improvement of level of IT application, distributed database is commonly used. Distributed data mining is mining overall knowledge which is useful for management and decision from database distributed in geography. It has become an important issue in data mining analysis. Distributed data mining can achieve a mining task with computers in different site on the internet. It can not only improve the mining efficiency, reduce the transmitting amount of network data, but is also good for security and privacy of data. Based on related theories and current research situation of data mining and distributed data mining, this thesis will focus on analysis on the structure of distributed mining system and distributed association rule mining algorithm[25,6].

Clustering can be defined as the process of partitioning a set of patterns into disjoint and homogeneous meaningful groups, called clusters[1,2]. The growing need for distributed clustering algorithms is attributed to the huge size of databases that is common nowadays[6,26]. Clustering is a “grouping a collection of objects into subsets or clusters, such that those within one cluster are more closely related to one another than objects assigned to different clusters”, is a fundamental process of Data Mining. In particular, clustering is fundamental in knowledge acquisition. It is applied in

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

various fields including data mining, statistical data analysis, compression and vector quantization. Clustering is, also, extensively applied in social sciences.

The task of extracting knowledge from large databases, in the form of clustering rules, has attracted considerable attention. The ability of various organizations to collect, store and retrieve huge amounts of data has rendered the development of algorithms that can extract knowledge in the form of clustering rules, a necessity. Distributed clustering algorithms embrace this trend of merging computations with communication and explore all the facets of the distributed computing environments. Thus a distributed algorithm must take under consideration that the data may be inherently distributed to different loosely cou-pled sites connected through a network.

Lately, big data clustering has been widely studied in many areas, together with statistics, machine learning, pattern recognition, and image processing [13-15]. In the regions, the scalability of clustering techniques and the methods for big data clustering much vigorous research has been dedicated. Different techniques has been introduced to overcome the problems happened in large database clustering, including initialization by clustering a model of the data and by means of an initial crude partitioning of the complete data set [7]. On the other hand, the most well-known representatives are partitioning clustering techniques such as CLARANS [11]; hierarchical clustering techniques such as BIRCH [10]; grid clustering techniques such as STING [8] and WAVECLUSTER [9]. Each technique has its benefits and shortcomings. For processing very large databases they are not appropriate. It is hard to obtain both high precision and competence in a clustering algorithm of large data. The two targets the entire time clash with each other. The power of a single computer is not sufficient in order to process massive data sets. Parallel and allocated clustering is the key method. In a distributed environment, it will be extremely scalable and low cost to do clustering.

In this paper, we propose an associative constraint based data clustering using multi-dimensional data. Bayes factor is employed to constraint based data clustering process in this paper. Additionally, genetic algorithm is used to obtain optimal clustering results. We evaluate the proposed algorithm on two real-world multi-dimensional data provided by UCI Machine Learning Repository. The remainder of this paper is organized as follows. Section 2 provides a review of related works. In section 3 explains basic concept of associative clustering. In section 4 focus on efficient implementation of proposed Associative Constraints Based Optimal Clustering algorithm. We devote section 5 to the experimental evaluation of our algorithm. Finally, we conclude in section 6.

II. REVIEW OF RELATED WORKS

In past, the cluster analysis has been widely applied to many areas such as medicine, chemistry, social studies, and so on. The main goal is to identify structures or clusters present in the data. Existing clustering algorithms can be classified into two main categories: hierarchical method and partitioning methods. Hierarchical methods are either agglomerative or divisive. Given n objects to be clustered, agglomerative methods begin with n clusters. In each step, two clusters are chosen and merged. This process continuous until n clusters is generated. While hierarchical methods have been successfully applied to many biological applications, they are well known to suffer from the weakness that they can never undo what was done previously. Once an agglomerative method merges two objects these objects will never be re-grouped into the same cluster.

In contrast, given the number k of partitions to be found, a partitioning method tries to find the best k partitions of the n objects. It is very often case that the clusters found by a partitioning method are of higher quality than the k clusters produced by a hierarchical method. Because of this property developing partitioning methods has been one of the main focuses of cluster analysis research. Indeed, many partitioning method has been developed, some of them based on k-means, some on k-mediod, some on fuzzy analysis, etc. Among them, we have chosen the k-medoid methods as the basis of our algorithm for the following reasons. First one unlike many other partitioning methods, the k-medoid methods are very robust to the outliers. Second, clusters found by k-medoid methods do not depend on the order in which the objects are examined. Furthermore, they are invariant with respect to the translations and orthogonal transformations of the data points. Last but not the least, experiments have shown that the k-medoid methods described below can handle very large data sets quite efficiently. Now we represent the two best known k-medoid methods on which the proposed algorithm is based.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

A. *PAM*

Partitioning around methods was introduced by the Kaufman and Rousseeuw. PAM's approached is to conclude a representative object for each cluster to find the k clusters. This representative object is called as medoid, is meant to the most centrally located object within the cluster. Once medoid have been selected, each non-selected is grouped with the medoid to which it is the most similar. More precisely, if Oj is a non-selected object, and Om is a selected medoid, we say that Oj belongs to the cluster represented by Om, if d (Oj, Om) = minO d (Oj, Om), where the notation minO denotes the minimum over all medoids Oe, and the notation d (O1, O2) denotes the dissimilarity or the distance between objects O1 and O2. All the dissimilarity values are given as inputs to PAM. Finally, the quality of clustering is measured by the average dissimilarity between an object and the medoid of its cluster. To find the k medoids, PAM begins with an arbitrary selection of k objects. Then in each steps, a swap between a selected object Om and non-selected object Op is made, as long as such a swap would result in as improvement of the quality of the clustering.

Algorithm PAM

i. Select k representative objects arbitrarily.
ii. Compute TCmp for all pairs of objects Om, Op where Om is currently selected and Op is not.
iii. Select the pair Om, Op which corresponds to minOm, Op TCmp. If the minimum TCmp is negative, replace Om and Op, and goes back to step ii.
iv. Otherwise, for each non-selected object, find the most similar representative object.

Experimental results show that PAM work satisfactorily for small data sets. But it is not efficient in dealing with medium and large data sets.

B. *CLARA*

Introduced by Kaufman and Rousseeuw to handle large data sets, CLARA relies on sampling. Instead of finding representative objects for the entire data set, CLARA draws a sample of the data set, applies PAM on the sample, and finds the medoids of the sample. The points are that if the sample is drawn in a sufficiently random way, the medoids of the sample would approximate the medoids of the entire data set. To come up with better approximations, CLARA draws multiple samples and gives the best clustering as the output. Here for accuracy the quality of a clustering is measured based on the average dissimilarity of all objects in the entire data set, and not only of those objects in the samples.

Algorithm CLARA

i. For i = 1 to 5 repeat the following steps:
ii. Draw a sample of 40+2k objects randomly from the entire data set, and call algorithm PAM to find the k medoids of the sample.
iii. For each object Oj in the entire data set, determine which of the k medoids is the most similar to Oj.
iv. Calculate the average dissimilarity of the clustering obtain in the previous steps. If this value is less than the current minimum, use this value as the current minimum and retain k medoids found in step ii as the best of the medoids obtained so far.
v. Return to step I to start the next iteration.

In some state of the art the following concepts are used:

i. The introduction and development of CLARANS, which aims to use randomized searched to facilitate the clustering of a large number of objects;
ii. A study on the efficiency and effectiveness of three different approaches to evaluate the exact separation distance between two polygons, the approach that over estimates the exact distance by using the minimum distance between the vertices, and the approach that under estimate the exact distance by using the separation distance between the isothetic rectangles of the polygons.

A. *Ensemble Learning*

An ensemble is itself a supervised learning algorithm, because it can be trained and then used to make predictions. The trained ensemble, therefore, represents a single hypothesis. This hypothesis, however, is not necessarily contained within the hypothesis space of the models from which it is built. Thus, ensembles can be shown to have more flexibility

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

in the functions they can represent. Ensembles combine multiple hypotheses to form a (hopefully) better hypothesis. In other words, an ensemble is a technique for combining many weak learners in an attempt to produce a strong learner. The term ensemble is usually reserved for methods that generate multiple hypotheses using the same base learner.Ensemble learning is primarily used to improve the (classification, prediction, function approximation, etc.) performance of a model, or reduce the likelihood of an unfortunate selection of a poor one. Other applications of ensemble learning include assigning a confidence to the decision made by the model, selecting optimal (or near optimal) features, data fusion, incremental learning, non-stationary learning and error-correcting.

B. *Distributed Data Mining Algorithms*

Most DDM algorithms are designed upon the potential parallelism they can apply over the given distributed data. Typically the same algorithm operates on each distributed data site concurrently, producing one local model per site. Subsequently all local models are aggregated to produce the final model. In essence, the success of DDM algorithms lies in the aggregation. Each local model represents locally coherent patterns, but lacks details that may be required to include globally meaningful knowledge. For this reason, many DDM algorithms require a centralization of a subset of a local data to compensate it. Therefore, minimum data transfer is another key attribute of the successful DDM algorithm. Data mining deals with the problem of analyzing data in scalable manner. DDM is a branch of data mining that defines a framework to mine distributed data paying careful attention to the distributed data and computing resources. The development of data mining algorithms that work well under the constraints imposed by distributed datasets has received significant attention from the data mining community in recent years. Most distributed clustering algorithms have their foundations in parallel computing, and are thus applicable in homogeneous scenarios. They focus on applying centre-based clustering algorithms, such as K-Means, K-Harmonic Means and EM in a parallel fashion. Here we are discussing clustering algorithm called K-Means and will also try to focus partially on its distributed version clustering algorithm called DK-means. It is proved that this algorithm achieves same result as of K-means. We are trying to focus on the areas like, how this algorithm works, what king of results it will produce, what will be the limitations, which environment it will require, what will be the future scope and many more.

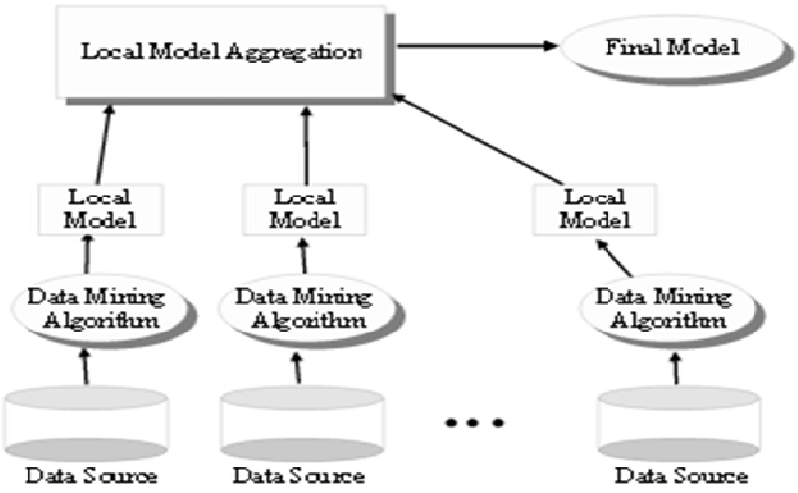

Fig.1. Distributed Data Mining Framework

A shown in Fig.1, the objective of DDM is to perform the data mining operations based on the type and availability of the distributed resources. It may choose to download the data sets to a single site and perform the data mining operations at a central location. However, that decision in DDM should be based on the properties of the computing, storage and communication capabilities. This is in contrast with the traditional centralized data mining methodology where collection of data at a single location prior to analysis is an invariant characteristic.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

In the DDM literature one or two assumptions is commonly adopted as to how data is distributed across sites: homogeneously (horizontally partitioned) and heterogeneously (vertically partitioned). Both viewpoints adopt the conceptual viewpoint that the data tables at each site are partitions of a single global table.

- In the homogeneous case, the global table is horizontally partitioned the tables at each site are subsets of the global table: they have exactly the same attributes.
- In the heterogeneous case the table is vertically partitioned. Each site contains a collection or columns (sites do not have the same attributes). However, each tuple each site is assumed to contain a unique identifier to facilitate matching. It is important to stress that the global table viewpoint is strictly conceptual. It is not necessarily assumed that such a table was physically realized and partitioned to form the tables at each site.

The development of data mining algorithms that work well under the constraints imposed by distributed datasets has received significant attention from the data mining community in recent years. The field of DDM has emerged as an active area of study. The bulk of DDM methods in the literature operate over an abstract architecture which includes multiple sites having independent computing power and storage capability. Local computation is done on each of the sites and either a central site communicates with each distributed site to compute the global models or a peer-to-peer architecture is used. In the latter case, individual nodes might communicate with a resource centralized node but they perform most of the tasks by communicating with neighbouring nodes by message passing over an asynchronous network.

For example. the sites may represent independent sensor nodes which connect to each other in an ad-hoc fashion. Some features of a distributed scenario where DDM is applicable are as follows.

1. The system consists of multiple independent sites of data and computation which communicate only through message passing.
2. Communication between the sites is expensive.
3. Sites have resource constraints e.g. battery power.
4. Sites have privacy concerns.

Typically communication is a bottleneck. Since communication is assumed to be carried out exclusively by message passing, a primary goal of many DDM methods in the literature is to minimize the number of messages sent.

Some methods also attempt to load-balance across sites to prevent performance from being dominated by the time and space usage of any individual site. “Building a monolithic database in order to perform non-distributed data mining, may be infeasible or simply impossible” in many applications. The cost of transferring large blocks of data may be prohibitive and result in very inefficient implementations. Privacy plays an important role in DDM as some participants may wish to not share their data but still participate in DDM.

Data clustering is a data exploration technique that allows objects with similar characteristics to be grouped together in order to facilitate their further processing. Data clustering has many engineering applications including the identification of part families for cellular manufacture. Clustering is the process of partitioning or grouping a given set of patterns into disjoint clusters .This is done such that patterns in the same cluster are alike and Patterns be-longing to two different clusters are different. Clustering has been a widely studied problem in a variety of application domains including neural networks, AI, and statistics.

The task of extracting knowledge from large databases, in the form of clustering rules, has attracted considerable attention. The ability of various organizations to collect, store and retrieve huge amounts of data has rendered the development of algorithms that can extractknowledge in the form of clustering rules, a necessity. Distributed clustering algorithms embrace this trend of merging computations with communication and explore all the facets of the distributed computing environments. Thus a distributed algorithm must take under consideration that the data may be inherently distributed to different loosely coupled

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

sites connected through a network.In a distributed computing environment the dataset is spread over a number of different sites. Thus, let us assume that the entire dataset X is distributed among m sites each one storing Xi for i = 1; .....m,

so X = $\cup$ (i=1,.....,m) Xi . Furthermore let us assume that there is a central site O that will hold the final clustering results.

III. ASSOCIATIVE CONSTRAINTS BASED OPTIMAL CLUSTERING ALGORITHM

Our proposed algorithm for associative constraints based optimal clustering algorithm mainly concentrate on the constraints based multi-dimensional data clustering. The constraints helps in identifying the right data to be clustered and the knowledge regarding the data also considered as a constraint which improves the accuracy of clustering. The associative clustering method help to identify relationship between the two clusters based on the constraint values.

Input : Multidimensional Data D
Output : Clustered Output

Intermediate processes: Subset formation and optimal Associative Constraint based data Clustering

Parameters :

$D = \{d_{ij}; 0 \le i \le m \text{ and } 0 \le j \le n\}$ → Dataset having n attributes

D_d = Discretized data

F max →Maxumum value of every feature

F min → Minimum value of every feature

N → dimensions

P → mapping = $p:$ $R^N \rightarrow \bigcup_j R^{n_j}, \quad n_j \le N$

m, n → subsets

i, j → subspaces

I → Interval

d→ Euclidean Distance

l→ first l no of shortest distances chosen for m^{th} subset

p→ point in a high dimensional data space

$Dev(k) = \dfrac{Max(d_k) - Min(d_k)}{2}$ → Deviation for every k

Start

for all $d_{ij} \in D$

do
call *Discretization*
end

Discretization

Compute F max , F min

Normalize as:

$$\left.\begin{array}{ll} 0, & input < 1(Dev(k)) \\ 1, & input < 2(Dev(k)) \\ 2, & input < 3(Dev(k)) \\ 4, & input < 4(Dev(k)) \end{array}\right\}$$

$\Downarrow$

D_d

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

Call Subset_formation
End
Subset_formation
Select P in space randomly
for all p

compute d

select first l for m

repeat

End

In our GA based optimal associative clustering algorithm, a chromosome representing the associative clustering task or process by two set of functions (dimensional selection and optimal cluster generation) is used and each chromosome is individually evaluated by using the fitness function. In the evolutionary loop, a set of individuals is chosen for evolutionary cross over and mutation. The possibility of evolutionary operator is chosen adaptively. The crossover operator converts two individuals (parents) into two offspring by joining parts from each parent. Now, single point crossover is used to convert two individuals. The mutation operator works on a single individual and forms an offspring by mutating that individual. On the basis of the fitness function and form the novel generation the recently generated individuals are assessed. The chromosome with the best fitness value is selected in every generation. The process ends after some number of generations either by the user or vigorously by the program itself, where the best chromosome acquired will be taken as the best solution. The best string of the last generation provides the solution to our clustering problem.

IV. RESULT AND DISCUSSION

The proposed clustering algorithm is executed in a windows machine containing configurations Intel (R) Core i5 processor, 3.20 GHz, 4 GB RAM, and the operation system platform is Microsoft Wnidow7 Professional. Also, we have employed mat lab latest version (7.12) for implementation.

4.1 Dataset description

For the experimental results, two real-world datasets namely adult and census downloaded from the UCI Repository of Machine Learning Databases [29].

UCI Adult data: This is the annual income data consisting of 48842 instances (mix of continuous and discrete) or 45222 instances (if instances with unknown values are removed). Also, it contains 6 continuous, 8 nominal attributes and 1 class attribute. This is extracted from the census dataset.

UCI Census data: The census data has 2,458,284 records with 68 categorical attributes, about 352 Mbytes in total. It was derived from the USCensus1990raw data set which was obtained from the (U.S. Department of Commerce) Census Bureau website using the Data Extraction System.

It seen that the proposed method is outperformed having the accuracy of 76.4%, which is high compared with existing algorithm only achieved 71.33% in subset 20. It is seen that using UCI census data, the proposed algorithm takes minimum time when compares with existing [23] for clustering process.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

- *Experimental results on UCI adult data*

TABLE I a
ACCURACY PERFORMANCE OF ADULT DATASET: INITIAL SUBSET VS. ACCURACY

	accuracy	
Initial subset	**Existing**	**Proposed**
20	71.33	76.4
25	75.01	75.08
30	75.02	75.6
35	75.01	76

TABLE II b
TIME PERFORMANCE OF ADULT DATASET: TIME VS. INITIAL SUBSETS

	Time performance(ms)	
Initial subset	**Existing**	**Proposed**
20	14800	10000
25	11000	9000
30	8000	7000
35	4500	3000

- *Experimental results on UCI census data*

TABLE II a
ACCURACY PERFORMANCE OF CENSUS DATASET: ACCURACY VS. INITIAL SUBSETS

	accuracy	
Initial subset	**Existing**	**Proposed**
20	71.11	76.44
25	72.11	72.24
30	72.12	72.34
35	72.11	75.89

TABLE II b
TIME PERFORMANCE OF CENSUS DATASET: TIME VS. INITIAL SUBSETS

	Time performance(ms)	
Initial subset	**Existing**	**Proposed**
20	18000	16000
25	14000	10000
30	10000	8000
35	6000	4000

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

V. CONCLUSION

In this article, we propose an efficient approach to high-dimensional clustering using genetic algorithm. Then, by bays factor computation process associative constraint based clustering process was executed. Also, genetic algorithm is applied to optimization process to discover the optimal cluster results. The constraints based proposed algorithm assists in recognizing the right data to be clustered and the knowledge considering the data regarded as a constraint which enhances the precision of clustering. The data constraints furthermore assist in indicating the data related to the clustering task. Our experimental evaluation demonstrated that the proposed algorithm compares favorably to one existing algorithm on two multi dimensional dataset. Experimental results showed that the performance of this clustering algorithm is high, effective, and flexible.

REFERENCES

1. Osmar R. Z., "Introduction to Data Mining", In: Principles of Knowledge Discovery in Databases. CMPUT690, University of Alberta, Canada, 1999.
2. Kantardzic, Mehmed. "Data Mining: Concepts, Models, Methods, and Algorithms", John Wiley and Sons, 2003.
3. E. Wainright Martin, Carol V. Brown, Daniel W. DeHayes, Jeffrey A. Hoffer and William C. Perkins, "Managing information technology", Pearson Prentice-Hall 2005.
4. Andrew Kusiak and Matthew Smith, "Data mining in design of products and production systems", in proceedings of Annual Reviews in control, vol. 31, no. 1, pp. 147- 156, 2007.
5. Mahesh Motwani, J.L. Rana and R.C Jain, "Use of Domain Knowledge for Fast Mining of Association Rules", in Proceedings of the International Multi-Conference of Engineers and Computer Scientists, 2009.
6. Souptik Datta Kanishka Bhaduri Chris Giannella Ran Wolff Hillol Kargupta "Distributed Data Mining in Peer-to-Peer Networks", Journal of internet computing, vol.10, no.4, pp.18-26. 2006.
7. Ron Wehrens and Lutgarde M.C. Buydens, "Model-Based Clustering for Image Segmentation and Large Datasets via Sampling", Journal of Classification, Vol. 21, pp.231-253, 2004.
8. W. Wang, J. Yang, R. Muntz, STING,"A Statistical Information Grid Approach to Spatial Data Mining", VLDB, 1997.
9. G. Sheikholeslami, S. Chatterjee, and A. Zhang, "WaveCluster: A multi-resolution clustering approach for very large spatial databases", VLDB, pp. 428-439, 1998.
10. T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: an efficient data clustering method for very large databases", In Proceedings of the ACM SIGMOD International Conference on Management of Data, pp.103-114, 1996.
11. Ng R. T., Han J.: "Efficient and Effective Clustering Methods for Spatial Data Mining", Proceedings 20th International Conference on Very Large Data Bases, pp.144-155, 1994.
12. Inderjit S. Dhillon and Dharmendra S. Modha, "A Data-Clustering Algorithm On Distributed Memory Multiprocessors", Proceedings of KDD Workshop High Performance Knowledge Discovery, pp. 245-260, 1999.
13. M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-bsed algorithm for discovering clusters in large spatial databases with noise", In SIGKDD, pp. 226–231, 1996.
14. R. T. Ng and J. Han, "Efficient and effective clustering methods for spatial data mining", In VLDB, 1994.
15. T. Zhang, R. Ramakrishnan, and M. Livny. Birch, "An efficient data clustering method for very large databases", In SIGMOD, pp. 103–114, 1996.
16. Jinchao Ji , Wei Pang, Chunguang Zhou, Xiao Han, Zhe Wang, "A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data", journal of Knowledge-Based Systems, vol. 30, pp. 129-135, 2012.
17. Chen L, Chen CL, Lu M., "A multiple-kernel fuzzy C-means algorithm for image segmentation", IEEE Transaction on System Man Cybernetics: Part B, vol. 41, no. 5, pp. 1263-74, 2011.
18. Wen-Yen Chen, Yangqiu Song, Hongjie Bai, Chih-Jen Lin and Edward Y. Chang, "Parallel Spectral Clustering in Distributed Systems", IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol.33, No.3, pp. 568 – 586, 2011.
19. Eshref Januzaj, Hans-Peter Kriegel and Martin Pfeifle, "Scalable Density-Based Distributed Clustering", Proceedings of 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), pp. 231-244, 2004.
20. Josenildo Costa da Silva and Matthias Klusch, "Inference in Distributed Data Clustering", Engineering Applications of Artificial Intelligence, Vol.19, No.4, pp.363-369, 2005.
21. Ruoming Jin, Anjan Goswami and Gagan Agrawal, "Fast and Exact Out-of-Core and Distributed K-Means Clustering", Journal of Knowledge and Information System, Vol. 10, No.1, pp. 17-40, 2006.
22. Genlin Ji and Xiaohan Ling, "Ensemble Learning Based Distributed Clustering", Emerging Technology in Knowledge Discovery and Data Mining, Vol. 4819, pp 312-321, 2007.
23. Olivier Beaumont, Nicolas Bonichon, Philippe Duchon, Lionel Eyraud-Dubois and Hubert Larcheveque, "A Distributed Algorithm for Resource Clustering in Large Scale Platforms", Principles of Distributed Systems, Vol.5401, pp.564-567, 2008.
24. Samuel Kaski, Janne Nikkila¨ , Janne Sinkkonen, Leo Lahti, Juha E.A. Knuuttila, and Christophe Roos," Associative Clustering for Exploring Dependencies between Functional Genomics Data Sets", IEEE/ACM Transactions On Computational Biology And Bioinformatics, Vol. 2, No. 3,pp: 203-216, 2005.
25. Yao Yuhui, Chen Lihui, Andrew Goh, Ankey Wong, " Clustering Gene Data Via Associative Clustering Neural Network", Proceedings of the 9th International Conference on Neural Information Processing, Vol.5, pp: 2228- 2232, 2002.
26. Hesam Izakian, Ajith Abraham, Vaclav Snasel, "Fuzzy Clustering Using Hybrid Fuzzy c-means and Fuzzy Particle Swarm Optimization", World Congress on Nature and Biologically Inspired Computing (NaBIC 2009), India, IEEE Press, pp. 1690-1694, 2009.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

27. Swagatam Das, Ajith Abraham, Amit Konar, "Automatic Clustering Using an Improved Differential Evolution Algorithm", IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems And Humans, Vol. 38, No. 1, 2008.
28. J.H. Holland, Adaptation in Natural and Artificial Systems, MIT Press, Cambridge, MA, 1992.
29. I.J. Good., "On the Application of Symmetric Dirichlet Distributions and Their Mixtures to Contingency Tables," Annals of Statistics, vol. 4, pp. 1159-1189, 1976.
30. UCI Repository of Machine Learning databases, University of California, Irvine, Department of Information and Computer Science, http://www.ics.uci.ed/~mlearn/MLRepository.html