ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 8, October 2012

Data Clustering Using Data Mining Techniques

S.R.Pande [1], Ms. S.S.Sambare [2], V.M.Thakre [3]

Department of Computer Science, SSES Amti's Science College, Congressnagar, Nagpur, India[1]

Department of Computer Applications, Dhanwate NationalCollege, Congressnagar, Nagpur, India[2]

Department of Computer Science, S.G.B. Amarvati University, Amarvati, India[3]

ABSTRACT: Cluster analysis divides data into meaningful or useful groups (clusters). If meaningful clusters are the goal, then the resulting clusters should capture the "natural" structure of the data. For example, cluster analysis has been used to group related documents for browsing, to find genes and proteins that have similar functionality, and to provide a grouping of spatial locations prone to earthquakes. However, in other cases, cluster analysis is only a useful starting point for other purposes, e.g., data compression or efficiently finding the nearest neighbors of points. Whether for understanding or utility, cluster analysis has long been used in a wide variety of fields: psychology and other social sciences, biology, statistics, pattern recognition, information retrieval, machine learning, and data mining. In this paper, a survey of several clustering techniques that are being used in Data Mining is presented. Data mining adds to clustering the complications of very large datasets with very many attributes of different types. This imposes unique computational requirements on relevant clustering algorithms. A variety of algorithms have recently emerged that meet these requirements and were successfully applied to real-life data mining problems.

Keywords: Clustering, partitioning, data mining, hierarchical clustering, k-means, density-based, grid-based

I. INTRODUCTION

Clustering is the process of grouping a collection of objects (usually represented as points in a multidimensional space) into classes of similar objects. Cluster analysis is a very important tool in data analysis. It is a set of methodologies for automatic classification of a collection of patterns into clusters based on similarity. Intuitively, patterns within the same cluster are more similar to each other than patterns belonging to a different cluster. It is important to understand the difference between clustering (unsupervised classification) and supervised classification.

Cluster analysis has wide applications in data mining, information retrieval, biology, medicine, marketing, and image segmentation. With the help of clustering algorithms, a user is able to understand natural clusters or structures underlying a data set. For example, clustering can help marketers discover distinct groups and characterize customer groups based on purchasing patterns in business. In biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionality, and gain insight into structures inherent in populations. Typical pattern clustering activity involves the following steps:

- pattern representation (including feature extraction and/or selection),
- definition of a pattern proximity measure appropriate to the data domain,
- clustering,
- data abstraction, and
- assessment of output.

Cluster analysis is an exploratory discovery process. It can be used to discover structures in data without providing an explanation/interpretation [13]. Cluster analysis includes two major aspects: clustering and cluster validation. Clustering aims at partitioning objects into groups according to a certain criteria. To achieve different application purposes, a large number of clustering algorithms have been developed[13][14][3]. While due to there are no general purpose clustering algorithms to fit all kinds of applications, thus, it is required an evaluation mechanism to assess the quality of clustering results that produced by different clustering algorithms or a clustering algorithm with different parameters, so that the user may find a fit cluster scheme for a specific application. The quality assessment process of clustering results is regarded as cluster validation. Cluster analysis is an iterative process of clustering and cluster verification by the user facilitated with clustering algorithms, cluster validation methods, visualization and domain knowledge to databases.

In this paper, we give a review of cluster analysis. First we introduce clustering, clustering algorithms and their features, and also the drawbacks of these algorithms. This is followed by the introduction of cluster validation, existing the cluster validation methods, and the problems with the existing cluster validation approaches.

II. CLUSTERING ALGORITHMS

Clustering is considered as an unsupervised classification process [14]. The clustering problem is to partition a dataset into groups (clusters) so that the data elements within a cluster are more similar to each other than data elements in different clusters by given criteria. A large number of clustering algorithms have been developed for different purposes [13][14][3]. Based on the strategy of how data objects are distinguished, clustering techniques can be broadly divided in two classes: hierarchical clustering techniques and partitioning clustering techniques [3]. However there is no clear

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 8, October 2012

boundary between these two classes. Some efforts have been done on the combination of different clustering methods for dealing with specific applications. Beyond the two traditional hierarchical and partitioning classes, there are several clustering techniques that are categorized into independent classes, for example, density-based methods, Grid-based methods and Model based clustering methods [7][3]. A short review of these methods is described below.

A. Partitioning methods

Partitioning clustering algorithms, such as K-means , K-medoids PAM , CLARA and CLARANS assign objects into k (predefined cluster number) clusters, and iteratively reallocate objects to improve the quality of clustering results. K-means is the most popular and easy-to understand clustering algorithm [15]. The main idea of K-means is summarized in the following steps:

- Arbitrarily choose k objects to be the initial cluster centers/centroids;
- Assign each object to the cluster associated with the closest centroid;
- Compute the new position of each centroid by the mean value of the objects in a cluster
- Repeat Steps 2 and 3 until the means are fixed.

Fig. 1 presents an example of the process of K-means clustering algorithm.

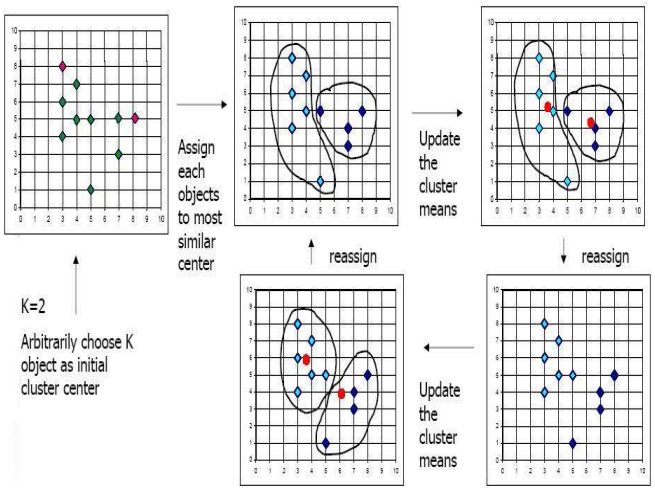

Fig. 1 An Example of clustering procure of K-means [7].

However, K-means algorithm is very sensitive to the selection of the initial centroids, in other words, the different centroids may produce significant differences of clustering results. Another drawback of K-means is that, there is no general theoretical solution to find the optimal number of clusters for any given data set. A simple solution would be to compare the results of multiple runs with different k numbers and choose the best one according to a given criterion, but when the data size is large, it would be very time consuming to have multiple runs of K-means and the comparison of clustering results after each run.

Instead of using the mean value of data objects in a cluster as the center of the cluster, a variation of K-means, K-medoids calculates the medoid of the objects in each cluster. The process of K-medoids algorithm is quite similar as K-means. Whereas, K-medoids clustering algorithm is very sensitive to outliers. Outliers could seriously influences clustering results.

To solve this problem, some efforts have been made based on K-medoids, for example PAM(Partitioning Around Medoids) was proposed by Kaufman and Rousseeuw [24]. PAM inherits the features of K-medoids clustering algorithm. Meanwhile, PAM equips a medoids swap mechanism to produce better clustering results. PAM is more robust than k-means in terms of handling noise and outliers, since the medoids in PAM are less influenced by outliers. With the O(k(n-k)2) computational cost for each iteration of swap (where k is the cluster number, n is the items of the data set), it is clear that PAM only performs well on small-sized datasets, but does not scale well to large datasets.

In practice, PAM is embedded in the statistical analysis systems, such as SAS, R, S+ and etc. to deal with the applications of large sized datasets, i.e., CLARA (Clustering Large Applications) [24]. By applying PAM to multiple sampled subsets of a dataset, for each sample, CLARA can produce the better clustering results than PAM in larger data sets. But the efficiency of CLARA depends on the sample size. On the other hand, a local optimum clustering of samples may not the global optimum of the whole data set. Ng and Han [25] abstracts the mediods searching in PAM or CLARA as searching k subgraphs from n points graph, and based on this understanding, they propose a PAM-like clustering algorithm called CLARANS (Clustering Large Applications based upon Randomized Search). While PAM searches the whole graph and CLARA searches some random sub-graphs, CLARANS randomly samples a set and selects k medoids in climbing sub-graph mountains. CLARANS selects the neighboring objects of medoids as candidates of new medoids. It samples subsets to verify medoids in multiple times to avoid bad samples. Obviously, multiple time sampling of medoids verification is time consuming. This limits CLARANS from clustering very large datasets in an acceptable time period.

B. Hierarchical methods

Hierarchical clustering algorithms assign objects in tree-structured clusters, i.e., a cluster can have data points or representatives of low level clusters [7]. Hierarchical clustering algorithms can be classified into categories according their clustering process: agglomerative and divisive. The process of agglomerative and divisive clustering are exhibited in Fig. 2.

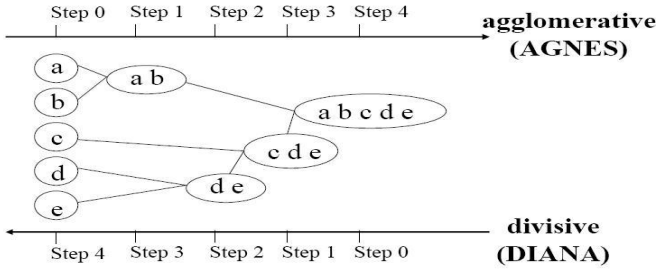

Fig. 2 Hierarchical Clustering Process [7]

- Agglomerative: one starts with each of the units in a separate cluster and ends up with a single cluster that contains all units.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 8, October 2012

- Divisive: to start with a single cluster of all units and then form new clusters by dividing those that had been determined at previous stages until one ends up with clusters containing individual units.

AGNES(Agglomerative Nesting) adopts agglomerative strategy to merge clusters. AGNET arranges each object as a cluster at the beginning, then merges them as upper level clusters by given agglomerative criteria step-by-step until all objects form a cluster, as shown in Figure 2. The similarity between two clusters is measured by the similarity function of the closest pair of data points in the two clusters, i.e., single link. DIANA (Divisive Analysis) adopts an opposite merging strategy, it initially puts all objects in one cluster, then splits them into several level clusters until each cluster contains only one object [24].

The merging/splitting decisions are critical in AGNES and DIANA. On the other hand, with O(n2) computational cost, their application is not scalable to very large datasets. Zhang et al [23] proposed an effective hierarchical clustering method to deal with the above problems, BIRCH (Balanced and Iterative Reducing and Clustering using Hierarchies). BIRCH summarizes an entire dataset into a CF-tree and then runs a hierarchical clustering algorithm on a multi-level compression technique, CF-tree, to get the clustering result. Its linear scalability is good at clustering with a single scan and its quality can be further improved by a few additional scans. It is an efficient clustering method on arbitrarily shaped clusters. But BIRCH is sensitive to the input order of data objects, and can also only deal with numeric data. This limits its stability of clustering and scalability in real world applications.

CURE uses a set of representative points to describe the boundary of a cluster in its hierarchical algorithm [6]. But with the increase of the complexity of cluster shapes, the number of representative points increases dramatically in order to maintain the precision.

CHAMELEON [26] employs a multilevel graph partitioning algorithm on the k-Nearest Neighbor graph, which may produce better results than CURE on complex cluster shapes for spatial datasets. But the high complexity of the algorithm prevents its application on higher dimensional datasets.

C. Density-based methods

The primary idea of density-based methods is that for each point of a cluster the neighborhood of a given unit distance contains at least a minimum number of points, i.e. the density in the neighborhood should reach some threshold [5]. However, this idea is based on the assumption of that the clusters are in the spherical or regular shapes.

DBSCAN(Density-Based Spatial Clustering of Applications with Noise) was proposed to adopt density-reachability and density-connectivity for handling the arbitrarily shaped clusters and noise [5]. But DBSCAN is very sensitive to the parameter Eps (unit distance or radius) and MinPts (threshold density), because before doing cluster exploration, the user is expected to estimate Eps and MinPts.

DENCLUE (Density-based Clustering) is a distribution-based algorithm [27], which performs well on clustering large datasets with high noise. Also, it is significantly faster than existing density-based algorithms, but DENCLUE needs a large number of parameters. OPTICS is good at investigating the arbitrarily shaped clusters, but its non-linear complexity often makes it only applicable to small or medium datasets [2].

D. Grid-based methods

The idea of grid-based clustering methods is based on the clustering oriented query answering in multilevel grid structures. The upper level stores the summary of the information of its next level, thus the grids make cells between the connected levels, as illustrated in Fig. 3.

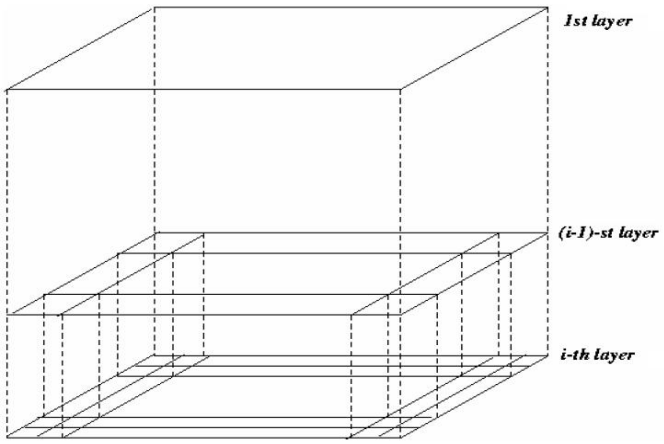

Fig.3 The grid-cell structure of gird-based clustering methods

Many grid-based methods have been proposed, such as STING (Statistical Information Grid Approach) [29], CLIQUE [28], and the combination of grid-density based technique WaveCluster [19]. The grid-based methods are efficient on clustering data with the complexity of O(N). However the primary issue of grid-based techniques is how to decide the size of grids. This quite depends on the user's experience.

E. Model-based clustering methods

Model-based clustering methods are based on the assumption that data are generated by a mixture of underlying probability distributions, and they optimize the fit between the data and some mathematical model, for example statistical approach, neural network approach and other AI approaches. When facing an unknown data distribution, choosing a suitable one from the model based candidates is still a major challenge. On the other hand, clustering based on probability suffers from high computational cost, especially when the scale of data is very large.

Based on the above review, we can conclude that, the application of clustering algorithms to detect grouping information in real world applications in data mining is still a challenge, primarily due to the inefficiency of most existing clustering algorithms on coping with arbitrarily shaped distribution of data of extremely large and high-dimensional datasets. Extensive survey papers on clustering techniques can be found in the literature [13][14][3].

III. CLUSTER VALIDATION

A large number of clustering algorithms have been developed to deal with specific applications [14]. Several questions arise: which clustering algorithm is best suitable for the application at hand? How many clusters are there in the studied data? Is there a better cluster

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 8, October 2012

scheme? These questions are related with evaluating the quality of clustering results, that is, cluster validation. Cluster validation is a procedure of assessing the quality of clustering results and finding a fit cluster strategy for a specific application. It aims at finding the optimal cluster scheme and interpreting the cluster patterns [9].

Cluster validation is an indispensable process of cluster analysis, because no clustering algorithm can guarantee the discovery of genuine clusters from real datasets and that different clustering algorithms often impose different cluster structures on a data set even if there is no cluster structure present in it [6]. Cluster validation is needed in data mining to solve the following problems [10]:

- To measure a partition of a real data set generated by a clustering algorithm.
- To identify the genuine clusters from the partition.
- To interpret the clusters.

Generally speaking, cluster validation approaches are classified into the following three categories Internal approaches, Relative approaches and External approaches [1].

We give a short introduction of cluster validation methods as follows.

A. Internal approaches

Internal cluster validation is a method of evaluating the quality of clusters when statistics are devised to capture the quality of the induced clusters using the available data objects only [21]. In other words, internal cluster validation excludes any information beyond the clustering data, and only focuses on assessing clusters' quality based on the clustering data themselves.

The statistical methods of quality assessment are employed in internal criteria, for example, root-mean-square standard deviation (RMSSTD) is used for compactness of clusters. R-squared (RS) for dissimilarity between clusters; and S_Dbw for compound evaluation of compactness and dissimilarity [7]. The formulas of RMSSTD, RS and S_Dbw are shown below.

$$RMSSTD = \sqrt{\frac{\sum_{\substack{i=1\ldots nc \\ j=1\ldots d}} \sum_{k=1}^{n_{ij}} \left(x_k - \overline{x_j}\right)^2}{\sum_{\substack{i=1\ldots nc \\ j=1\ldots d}} \left(n_{ij} - 1\right)}}$$

(1.1)

Where, x_j is the expected value in the j^{th} dimension; n_{ij} is the number of elements in the i^{th} cluster j^{th} dimension; n_j is the number of elements in the j^{th} dimension in the whole data set; n_c is the number of clusters.

$$RS = \frac{SS_t - SS_w}{SS_t}$$

(1.2)

where,

$$SS_t = \sum_{j=1}^{d} \sum_{k=1}^{n_j} \left(x_k - \overline{x_j}\right)^2, \quad SS_w = \sum_{\substack{i=1\ldots nc \\ j=1\ldots d}} \sum_{k=1}^{n_{ij}} \left(x_k - \overline{x_j}\right)^2$$

(1.3)

The formula of S_Dbw is given as:

S_Dbw = Scat(c) + Dens_bw(c)

(1.4)

where Scat(c) is the average scattering within c clusters. The Scat(c) is defined as:

$$Scat(c) = \frac{\frac{1}{c} \sum_{i=1}^{c} \left\| \sigma(v_i) \right\|}{\left\| \sigma(X) \right\|}$$

(1.5)

The value of Scat(c) is the degree of the data points scattered within clusters. It reflects the compactness of clusters. The term is the variance of a data set; and the term is the variance of cluster ci. Dens_bw(c) indicates the average number of points between the c clusters (i.e., an indication of inter-cluster density) in relation with density within clusters. The formula of Dens_bw is given as:

$$Dens\_bw = \frac{1}{n_c\left(n_c - 1\right)} \sum_{i=1}^{n_c} \left(\sum_{\substack{j=1, \\ i \neq j}}^{n_c} \frac{density(u_{ij})}{\max\left\{density(v_i), density(v_j)\right\}} \right)$$

(1.6)

where uij is the middle point of the distance between the centres of the clusters vi and vj. The density function of a point is defined as the number of points around a specific point within the given radius.

B. Relative approaches

Relative assessment compares two structures and measures their relative merit. The idea is to run the clustering algorithm for a possible number of parameters (e.g., for each possible number of clusters) and identify the clustering scheme that best fits the dataset [1], i.e., they assess the clustering results by applying an algorithm with different parameters on a data set and finding the optimal solution. In practice, relative criteria methods also use RMSSTD, RS and S_Dbw to find the best cluster scheme in terms of compactness and dissimilarity from all the clustering results. Relative cluster validity is also called cluster stability, and the recent works on research of relative cluster validity are presented in [4].

C. External approaches

The results of a clustering algorithm are evaluated based on a pre-specified structure, which reflects the user's intuition about the clustering structure of the data set [11]. As a necessary post-processing step, external cluster validation is a procedure of hypothesis test, i.e., given a set of class labels produced by a cluster scheme, and compare it with the clustering results by applying the same cluster scheme to the other partitions of a database, as shown in the Fig. 4.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 8, October 2012

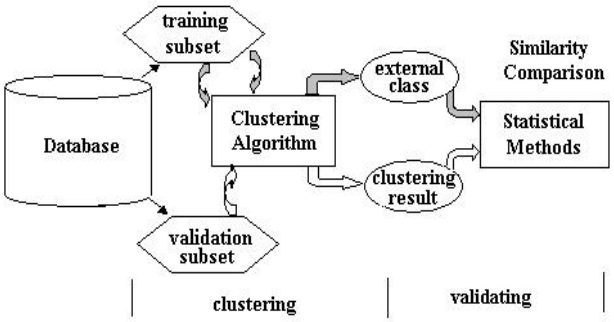

Fig. 4 External criteria based validation [22]

External cluster validation is based on the assumption that an understanding of the output of the clustering algorithm can be achieved by finding a resemblance of the clusters with existing classes [18],[17]. The statistical methods for quality assessment are employed in external cluster validation, such as Rand statistic [18], Jaccard Coefficient [12], Folkes and Mallows index [17], Huberts statistic and Normalized Γ statistic [21], and Monte Carlo method[16], to measure the similarity between the priori modelled partitions and clustering results of a dataset. Extensive surveys on cluster validation can be found in the literature [13],[14],[20],[8],[9],[11].

IV. THE PROBLEMS OF CLUSTER ANALYSIS

By the survey of cluster analysis above, it is clear that there are two major drawbacks that influence the feasibility of cluster analysis in real world applications in data mining. The first one is the weakness of most existing automated clustering algorithms on dealing with arbitrarily shaped data distribution of the datasets. The second issue is that, the evaluation of the quality of clustering results by statistics-based methods is time consuming when the database is large, primarily due to the drawback of very high computational cost of statistics-based methods for assessing the consistency of cluster structure between the sampling subsets. The implementation of statistics-based cluster validation methods does not scale well in very large datasets. On the other hand, arbitrarily shaped clusters also make the traditional statistical cluster validity indices ineffective, which leaves it difficult to determine the optimal cluster structure [9].

In addition, the inefficiency of clustering algorithms on handling arbitrarily shaped clusters in extremely large datasets directly impacts the effect of cluster validation, because cluster validation is based on the analysis of clustering results produced by clustering algorithms. Moreover, most of the existing clustering algorithms tend to deal with the entire clustering process automatically, i.e., once the user sets the parameters of algorithms, the clustering result is produced with no interruption, which excludes the user until the end. As a result, it is very hard to incorporate user domain knowledge into the clustering process. Cluster analysis is a multiple runs iterative process, without any user domain knowledge, it would be inefficient and unintuitive to satisfy specific requirements of application tasks in clustering.

V. CONCLUSIONS

Clustering lies at the heart of data analysis and data mining applications. The ability to discover highly correlated regions of objects when their number becomes very large is highly desirable, as data sets grow and their properties and data interrelationships change. At the same time, it is notable that any clustering "is a division of the objects into groups based on a set of rules – it is neither true or false". Some would argue that the wide range of subject matter, size and type of data, and differing user goals makes this inevitable, and that cluster analysis is really a collection of different problems that require a variety of techniques for their solution. The relationships between the different types of problems and solutions are often not clear. Every article that presents a new clustering technique shows its superiority to other techniques, it is hard to judge how well the technique will really do. In this paper we described the process of clustering from the data mining point of view. We gave the properties of a "good" clustering technique and the methods used to find meaningful partitioning.

REFERENCES

[1] L. Abul, R. Alhajj, F. Polat and K. Barker "Cluster Validity Analysis Using Subsampling," in *proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, Washington DC, Oct. 2003 Volume 2: pp. 1435-1440.
[2] M. Ankerst, M.M.Breunig, H.-P. Kriegel, J.Sander, "OPTICS: Ordering points to identify the clustering structure", in *proceedings of ACM SIGMOD Conference*, 1999 pp. 49-60.
[3] P. Berkhin, "A Survey of Clustering Data Mining Techniques" Kogan, Jacob; Nicholas, Charles; Teboulle, Marc (Eds) Grouping Multidimensional Data, Springer Press (2006) 25-72
[4] C. Baumgartner, C. Plant, K. Railing, H-P. Kriegel, P. Kroger, "Subspace Selection for Clustering High-Dimensional Data", Proc. of the Fourth IEEE International Conference on Data Mining (ICDM'04), 2004, pp.11-18.
[5] Ester M., Kriegel HP., Sander J., Xu X.: A density-based algorithm for discovering clusters in large spatial databases with noise. Second International Conference on Knowledge Discovery and Data Mining (1996)
[6] Guha S., Rastogi R., Shim K.: CURE: An efficient clustering algorithm for large databases. Proc. Of ACM SIGMOD Conference (1998)
[7] J. Han and M. Kamber, "*Data Mining: Concepts and Techniques*," Morgan Kaufmann Publishers, 2001.
[8] M. Halkidi, Y. Batistakis, M. Vazirgiannis, "On Clustering Validation Techniques" *Journal of Intelligent Information Systems*, Volume 17 (2/3), 2001, pp. 107–145.
[9] M. Halkidi, Y. Batistakis, M. Vazirgiannis, "Cluster validity methods: Part I and II", *SIGMOD* Record,31, 2002.
[10] Z. Huang, D. W. Cheung and M. K. Ng, "An Empirical Study on the Visual Cluster Validation Method with Fastmap", *Proceedings of DASFAA01*, Hong Kong, April 2001, pp.84-91.
[11] J . Handl, J. Knowles, and D. B. Kell, "Computational cluster validation in post-genomic data analysis", *Journal of Bioinformatics* Volume 21(15), 2005, pp. 3201-3212.
[12] Jaccard, S. (1908) Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.*, 44**,** 223–270.
[13] A. K. Jain and R. C. Dubes, "*Algorithms for Clustering Data*", Prentice Hall,1988.
[14] A. Jain, M. N. Murty and P. J. Flynn, "Data Clustering: A Review", *ACM Computing Surveys*, Volume 31(3), 1999, pp. 264-323.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 8, October 2012

[15] J. McQueen, "Some methods for classification and analysis of multivariate observations", *Proc. of 5th Berkeley Symposium on Mathematics, Statistics and Probability*, Volume 1, 1967, pp. 281-298.
[16] G. W. Milligan, "A Review Of Monte Carlo Tests Of Cluster Analysis", *Journal of* Multivariate Behavioral Research Vol. 16(3), 1981, pp. 379-407.
[17] G.W. Milligan, L.M. Sokol, & S.C. Soon "The effect of cluster size, dimensionality and the number of clusters on recovery of true cluster structure", IEEE Trans PAMI, 1983 5(1):40-47.
[18] Rand, W.M., Objective Criteria for the Evaluation of Clustering Methods. *J. Am. Stat. Assoc.*, 66:846-850, 1971.
[19] Sheikholeslami G., Chatterjee S., Zhang A.: Wavecluster: A multi-resolution clustering approach for very large spatial databases. Proc. of Very Large Databases Conference (1998)
[20] S. Theodoridis and K. Koutroubas, "*Pattern Recognition*", Academic Press. 1999.
[21] Vilalta R., Stepinski T., Achari M.: An Efficient Approach to External Cluster Assessment with an Application to Martian Topography, *Technical Report, No. UH-CS-05-08*, Department of Computer Science, University of Houston (2005).
[22] K-B. Zhang, M. A. Orgun, K. Zhang, "A Visual Approach for External Cluster Validation", Proc. of IEEE Symposium on Computational Intelligence and Data Mining (CIDM2007), Honolulu, Hawaii, USA, April 1-5, 2007, IEEE Press, 2007, pp576-582., Montreal, Canada (1996) 103-114.
[23] Zhang T., Ramakrishnan R. and Livny M.: BIRCH: An efficient data clustering method for very large databases. In Proc. of SIGMOD96
[24] KAUFMAN, L. and ROUSSEEUW, P. 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons, New York, NY.
[25] NG, R. and HAN, J. 1994. Efficient and effective clustering methods for spatial data mining. In Proceedings of the 20th Conference on VLDB, 144-155, Santiago, Chile.
[26] KARYPIS, G., HAN, E.-H., and KUMAR, V. 1999a. CHAMELEON: A hierarchical clustering algorithm using dynamic modeling, COMPUTER, 32, 68-75.
[27] HINNEBURG, A. and KEIM, D. 1998. An efficient approach to clustering large multimedia databases with noise. In Proceedings of the 4th ACM SIGKDD, 58-65, New York, NY.
[28] AGRAWAL, R., GEHRKE, J., GUNOPULOS, D., and RAGHAVAN, P. 1998. Automatic subspace clustering of high dimensional data for data mining applications. In Proceedings of the ACM SIGMOD Conference, 94-105, Seattle, WA.
[29] WANG, W., YANG, J., and MUNTZ, R. 1997. STING: a statistical information grid approach to spatial data mining. In Proceedings of the 23rd Conference on VLDB, 186-195, Athens, Greece.

Biography

S. R. Pande received his Ph.D. in Computer Science from RTM Nagpur University, Nagpur. He is currently Associate Professor of Computer Science and Head, Department of Computer Science, S.S.E.S. Amt's, Science College, Nagpur, Maharashtra state, India. He has about 21 years of experience in teaching and research. He has published several papers in national and international Journals and conferences. His areas of interest are Data Mining, Discrete mathematics, Theory of Computations, Neural Network, Fuzzy logic, Computer Graphics, Image Processing, Relational DBMS.

Ms S. S. Sambare received her M.Sc.(c/s), M.C.M., M.Phil. (I.T.) from RTM Nagpur University, Nagpur. She is currently Lecturer in Computer Application, Department of Computer Application, Dhanwate National College, Nagpur, Maharashtra state, India. Her research areas include Data Mining and Soft Computing.

V.M. Thakare is currently Professor and Head , Department of Computer Science and Engineering , SGB Amravati University, Amravati, Maharashtra state, India He has about 23 years of experience in teaching and research. He has published several papers in national and international Journals and conferences. His areas of interest are Robotics, Computer Architecture, Artificial Intelligence, Data Mining, and IT.