A Review on Data Clustering Algorithms for Mixed Data

GJCST Computing Classification
H.3.3, H.2.8

D. Hari Prasad[1] Dr. M. Punithavalli[2]

***Abstract*-Clustering is the unsupervised classification of patterns into groups (clusters). The clustering problem has been addressed in many contexts and by researchers in many disciplines; this reflects its broad appeal and usefulness as one of the steps in exploratory data analysis. In general, clustering is a method of dividing the data into groups of similar objects. One of significant research areas in data mining is to develop methods to modernize knowledge by using the existing knowledge, since it can generally augment mining efficiency, especially for very bulky database. Data mining uncovers hidden, previously unknown, and potentially useful information from large amounts of data. This paper presents a general survey of various clustering algorithms. In addition, the paper also describes the efficiency of Self-Organized Map (SOM) algorithm in enhancing the mixed data clustering**

Keywords-Data Clustering, Data Mining, Mixed Data Clustering, Self-Organized Map algorithm.

I. Introduction

Clustering is one of the standard workhorse techniques in the field of data mining. Its intention is to systematize a dataset into a set of groups, or clusters, which contain “similar” data items, as measured by some distance function. The major applications of clustering include document categorization, scientific data analysis, and customer/market segmentation. Data clustering has been considered as a primary data mining method for knowledge discovery. Clustering using Gaussian mixture models is also extensively employed for exploratory data analysis. The six sequential, iterative steps of Data mining processes are: 1) problem definition; 2) data acquisition; 3) data preprocessing and survey; 4) data modeling; 5) evaluation; 6) knowledge deployment [1]. The purpose of survey before data preprocessing is to gain insight knowledge into the data possibilities and problems to determine whether the data are sufficient. Moreover the survey assists us to select the proper preprocessing and modeling tools. Typically, several different data sets and preprocessing strategies need to be considered. For this reason, efficient visualizations and summarizations are essential.

Primarily the focus must be on clustering since they are important characterizations of data. The clustering method implemented should be fast, robust, and visually efficient. In the case of clustering Q means, the foremost step is partitioning a data set into a set of clusters Q_i, where i = 1 C. Data clustering techniques are gaining escalating reputation over traditional central grouping techniques, which are centered on the conception of “feature” (see e.g. [2], [3]). Several data clustering techniques have been put forth by researchers to assist in the development of knowledge.

Fuzzy clustering [4] is a simplification of crisp clustering where each sample has a varying degree of membership in all clusters. In many real-world applications, in fact, a feasible feature-based description of objects might be difficult to obtain or inefficient for learning purposes while, on the other hand, it is often possible to obtain a measure of the similarity or dissimilarity between objects. Among the central algorithmic procedures for perceptual organization are clustering principles like generalized k-means methods or clustering methods for proximity data [15].

The remainder of this paper is organized as follows section II describes the background study that is related to clustering algorithms proposed earlier, section III explains the challenging problems and areas of research and section IV concludes the paper with fewer discussions.

II. Background Study

A wealth of clustering techniques had been described in the literature. This section of the paper presents an overview on these clustering algorithms put forth by various researchers. In general, major clustering methods can be classified into five categories: partitioning methods, hierarchical methods, density-based methods, grid-based methods and model-based methods.

A. *Clustering of the Self-Organizing Map*

A novel method [1] was put forth by Juha Vesanto and Esa Alhoniemi for clustering of Self-Organizing Map. According to the method proposed in this paper the clustering is carried out using a two-level approach, where the data set is first clustered using the SOM, and then, the SOM is clustered. The purpose of this paper was to evaluate if the data abstraction created by the SOM could be employed in clustering of data. The most imperative advantage of this procedure is that computational load decreases noticeably, making it possible to cluster large data sets and to consider several different preprocessing strategies in a restricted time. Obviously, the approach is applicable only if the clusters found using the SOM are analogous to those of the original data.

B. *Kernel-Based Clustering*

Mark Girolami presents a Mercer Kernel-Based Clustering [5] algorithm in Feature Space. This paper presents a method for both the unsupervised partitioning of a sample of data and the estimation of the possible number of inherent

About-[1]Senior Lecturer, Department of Computer Applications, Sri Ramakrishna Institute of Technology, Coimbatore, India.
About-[2]Director, Department of Computer Science, Sri Ramakrishna Arts College for Women, Coimbatore, India.

clusters which generate the data. This work utilizes the perception that performing a nonlinear data transformation into some high dimensional feature space increases the probability of the linear separability of the patterns within the transformed space and therefore simplifies the associated data structure. In this case, the eigenvectors of a kernel matrix which defines the implicit mapping provides a means to estimate the number of clusters inherent within the data and a computationally simple iterative procedure is presented for the subsequent feature space partitioning of the data.

C. Grouping of Smooth Curves and Texture Segmentation using path-based clustering

A Path-Based Clustering algorithm [6] was described by Fischer and Buhmann for grouping of smooth curves and texture segmentation. This paper proposed a new grouping approach referred to as Path-Based Clustering [7], which measures local homogeneity rather than global similarity of objects. The new Path-Based Clustering method defines a connectedness criterion, which groups objects together if they are connected by a sequence of intermediate objects. Moreover an efficient agglomerative algorithm is proposed to minimize the Path-Based Clustering cost function. This approach utilizes a bootstrap resampling scheme to measure the reliability of the grouping results.

D. Bagging for Path-Based Clustering

Fischer and Buhmann present bagging for path-based clustering [8]. A resampling scheme for clustering with similarity to bootstrap aggregation (bagging) is presented in this paper. This aggregation (Bagging) is used to develop the quality of path-based clustering, a data clustering method that can extort stretched out structures from data in a noise stout way. In order to increase the reliability of clustering solutions, a stochastic resampling method is developed to deduce accord clusters. Moreover this paper also evaluates the quality of path-based clustering with resampling on a large image dataset of human segmentations.

E. Isoperimetric Graph Partitioning for Data Clustering

Leo Grady and Eric L. Schwartz together proposed an approach known as Isoperimetric Graph Partitioning for Data Clustering and Image Segmentation [9]. This paper, adopts a different approach, based on finding partitions with a small isoperimetric constant in an image graph. The algorithm described in this paper generates high quality segmentations and data clusters of spectral methods, but with improved speed and stability. The term “partition” in this paper refers to the assignment of each node in the vertex set into two (not necessarily equal) parts. Graph partitioning has been strongly influenced by properties of a combinatorial formulation of the classic isoperimetric problem: For a fixed area, find the region with minimum perimeter.

F. Improving Classification Decisions by Multiple Knowledge

The new approach to combine multiple sets of rules for text categorization using Dempster’s rule of combination [10] was described by Yaxin Bi et al. A boosting-like technique for generating multiple sets of rules based on rough set theory and model classification decisions from multiple sets of rules as pieces of evidence which can be combined by Dempster’s rule of combination is developed in this approach. This approach is employed to set of benchmark data collection, both individually and in combination. The experimental results show that the performance of the best combination of the multiple sets of rules on the benchmark data is significantly better than that of the best single set of rules.

G. Clustering Algorithm for Data Mining

Zhijie Xu et al. expressed a Modified Clustering Algorithm for Data Mining [11]. This paper describes a clustering method for unsupervised classification of objects in large data sets. The new methodology particularly combines the simulating annealing algorithm with CLARANS (clustering Large Application based upon Randomized Search) in order to cluster large data sets efficiently. The parameter T is used to control the process of clustering. In every step of the search, if the cost of the neighbor is less than the current, set the current to the neighbor. Otherwise, accept the neighbor with the probability of exp (-(Scost-currentcost)/T).

H. Dominant Sets and Pairwise Clustering

A graph-theoretic approach [12] for Pairwise data clustering was developed by Massimiliano Pavan and Marcello Pelillo. A correspondence is established between dominant sets and the extrema of a quadratic form over the standard simplex, thereby allowing the use of straightforward and easily implementable continuous optimization techniques from evolutionary game theory. In order to study the robustness of the approach against random noise in the background, the level of clutter is allowed to vary, starting from 100 to 1,000 points. Extensions of the approach presented in this paper involving hierarchical data partitioning and out of-sample extensions of dominant-set clusters can be found in [13], and [14], respectively.

I. A Conceptual Clustering Algorithm

Biswas et al. in [17] put forth a conceptual clustering algorithm for data mining. Their paper described an unsupervised discovery method with biases geared toward partitioning objects into clusters that improve interpretability. Their algorithm, ITERATE, employs: (i) a data ordering scheme and (ii) an iterative redistribution operator to produce maximally cohesive and distinct clusters. The important task here is interpretation of the generated patterns, and this is best addressed by creating groups of data that demonstrate cohesiveness within but clear distinctions between the groups. In clustering schemes, data objects are represented as vectors of feature-value pairs.

Features represent properties of an object that are relevant to the problem-solving task. Distinctness or inter-class dissimilarity was measured by an average of the variance of the distribution match between clusters. Additionally, their empirical results demonstrated the properties of the discovery algorithm, and its applications to problem solving.

J. The New K-Windows Algorithm for Improving the K-Means Clustering Algorithm

The new K-windows algorithm for improving the K-means clustering algorithm was described by Vrahatis et al. in [18]. The process of partitioning a large set of patterns into disjoint and homogeneous clusters is fundamental in knowledge acquisition. It is called Clustering in the literature and it is applied in various fields including data mining, statistical data analysis, compression and vector quantization. The k-means is a very popular algorithm and one of the best for implementing the clustering process. The k-means has a time complexity that is dominated by the product of the number of patterns, the number of clusters, and the number of iterations. Also, it often converges to a local minimum. In their paper, they presented an improvement of the k-means clustering algorithm, aiming at a better time complexity and partitioning accuracy. Moreover, their approach reduces the number of patterns that are needed to be examined for similarity using a windowing technique. The latter is based on well known spatial data structures, namely the range tree, which allows fast range searches.

K. A Spectral-based Clustering Algorithm

Abdu et al. in [19] presented a novel spectral-based algorithm for clustering categorical data that combines attribute relationship and dimension reduction techniques found in Principal Component Analysis (PCA) and Latent Semantic Indexing (LSI). The new algorithm uses data summaries that consist of attribute occurrence and co-occurrence frequencies to create a set of vectors each of which represents a cluster. They referred to these vectors as “candidate cluster representatives.” The algorithm also uses spectral decomposition of the data summaries matrix to project and cluster the data objects in a reduced space. They referred to the algorithm as SCCADDS (Spectral-based Clustering algorithm for CAtegorical Data using Data Summaries). SCCADDS differs from other spectral clustering algorithms in several key respects. Initially, the algorithm uses the feature categories similarity matrix instead of the data object similarity matrix (as is the case with most spectral algorithms that find the normalized cut of a graph of nodes of data objects). SCCADDS scales well for large datasets. Second, non-recursive spectral-based clustering algorithms characteristically necessitate K-means or some other iterative clustering method after the data objects have been projected into a reduced space. SCCADDS clusters the data objects directly by comparing them to candidate cluster representatives without the need for an iterative clustering method. Third, unlike standard spectral-based algorithms, the complexity of SCCADDS is linear in terms of the number of data objects. Results on datasets widely used to test categorical clustering algorithms show that SCCADDS produces clusters that are consistent with those produced by existing algorithms, while avoiding the computation of the spectra of large matrices and problems inherent in methods that employ the K-means type algorithms

L. A New Supervised Clustering Algorithm

A new supervised clustering algorithm was projected by Li et al. in [20]. They suggested their algorithm for data set with mixed attributes. Because of the complexity of data set with mixed attributes, the conventional clustering algorithms appropriate for this kind of dataset are not many and the result of clustering is not good. K-prototype clustering is one of the most commonly used methods in data mining for this kind of data. They borrowed the ideas from the multiple classifiers combing technology, use k- prototype as the basis clustering algorithm in order to design a multi-level clustering ensemble algorithm in the paper, which adoptively selects attributes for re-clustering. Comparison experiments on Adult data set from UCI machine learning data repository show very competitive results and the proposed method is suitable for data editing.

M. An Efficient Clustering Algorithm for mixed type attributes in Large Dataset

Jian et al. in [21] proposed an efficient algorithm for clustering mixed type attributes in large dataset. Clustering is a extensively used technique in data mining. At present there exist many clustering algorithms, but most existing clustering algorithms either are restricted to handle the single attribute or can handle both data types but are not competent when clustering large data sets. Few algorithms can do both well. In this article, they proposed a clustering algorithm that can handle large datasets with mixed type of attributes. They first used CF*tree (just like CF-tree in BIRCH) to pre-cluster datasets. After that the dense regions are stored in leaf nodes, and then they looked every dense region as a single point and used the ameliorated k-prototype to cluster such dense regions. Experimental results showed that this algorithm is very efficient in clustering large datasets with mixed type of attributes.

N. A Robust and Scalable Clustering Algorithm

A robust and scalable clustering algorithm was put forth by Chiu et al. in [22]. They employed this clustering algorithm for mixed type attributes in large database environment. In their paper, they proposed a distance measure that enables clustering data with both continuous and categorical attributes. This distance measure is derived from a probabilistic model that the distance between two clusters is equivalent to the decrease in log-likelihood function as a result of merging. Calculation of this measure is memory efficient as it depends only on the merging cluster pair and not on all the other clusters. The algorithm is implemented in the commercial data mining tool Clementine 6.0 which supports the PMML standard of data mining model deployment. For data with mixed type of attributes, their experimental results confirmed that the algorithm not only generates better quality clusters than the traditional k-means algorithms, but also exhibits good scalability properties and is able to identify the underlying number of clusters in the data correctly

O. Clustering Algorithm for Network Intrusion Detection system

Panda et al. in [23] described some clustering algorithms such as K-Means and Fuzzy c-Means for network intrusion detection. The objective of intrusion detection is to construct a system which would automatically scan network activity and detect such intrusion attacks. They built a system which created clusters from its input data, then automatically labeled clusters as containing either normal or anomalous data instances, and finally used these clusters to classify network data instances as either normal or anomalous. In their paper, they intended to propose a fuzzy c-means clustering technique which is capable of clustering the most suitable number of clusters based on objective function. Both the training and testing was done using 10% KDDCup'99 data, which is a very well-liked and broadly used intrusion attack dataset.

P. Clustering Algorithm-based on Quantum Games

A new clustering algorithm based on quantum games was projected by Li et al. in [24]. Mammoth successes have been made by quantum algorithms during the last decade. In their paper, they combined the quantum game with the problem of data clustering, and then they developed a quantum-game-based clustering algorithm, in which data points in a dataset are considered as players who can make decisions and implement quantum strategies in quantum games. After each round of a quantum game, each player's expected payoff is calculated. Soon after, he uses a link-removing-and-rewiring (LRR) function to change his neighbors and regulate the strength of links connecting to them in order to maximize his payoff. Further, algorithms are discussed and analyzed in two cases of strategies, two payoff matrixes and two LRR functions. Accordingly, the simulation results have demonstrated that data points in datasets are clustered reasonably and efficiently, and the clustering algorithms have fast rates of convergence. Furthermore, the comparison with other algorithms also provides an indication of the effectiveness of the proposed approach

Q. A GA-based Clustering Algorithm

Jie Li et al. in [25] proposed a GA-based clustering algorithm for large data sets with mixed and numeric and categorical values. In the field of data mining, it is frequently encountered to execute cluster analysis on large data sets with mixed numeric and categorical values. However, most existing clustering algorithms are only competent for the numeric data rather than the mixed data set. For this reason, their paper presented a novel clustering algorithm for these mixed data sets by modifying the common cost function, trace of the within cluster dispersion matrix. The genetic algorithm (GA) is used to optimize the new cost function to obtain valid clustering result. Experimental result illustrates that the GA-based new clustering algorithm is reasonable for the large data sets with mixed numeric and categorical values.

III. Challenging Problems And Areas Of Research

The algorithms proposed by researchers discussed in section II of this paper have their own advantages and limitations. The main requirements that a clustering algorithm should satisfy are: scalability, dealing with different types of attributes, discovering clusters with arbitrary shape, minimal requirements for domain knowledge to determine input parameters, ability to deal with noise and outliers, insensitivity to order of input records, high dimensionality, interpretability and usability. A number of problems are associated with conventional clustering algorithms. A few among them are current clustering techniques do not address all the requirements adequately (and concurrently), dealing with large number of dimensions and large number of data items can be problematic because of time complexity, the effectiveness of the method depends on the definition of "distance" (for distance-based clustering), if an obvious distance measure doesn't exist, then one must "define" it, which is not always easy, especially in multi-dimensional spaces, the result of the clustering algorithm (that in many cases can be arbitrary itself) can be interpreted in different ways [16]. A lot of algorithms for clustering data have been developed in recent decades, nonetheless, they all visage a major challenge in scaling up to very large database sizes, an accelerating development brought on by advances in computer technology, the Internet, and electronic commerce. The mainly focused research area is Clustering of mixed data. A clustering Q means partitioning a data set into a set of clusters Qi, where i = 1… C. In crisp clustering, each data sample belongs to exactly one cluster. Clustering algorithms may be classified as Exclusive Clustering, Overlapping Clustering, Hierarchical Clustering, and Probabilistic Clustering. Clustering objects into separated groups is an important topic in exploratory data analysis and pattern recognition. Many clustering techniques group the data objects together to "compact" clusters with the explicit or implicit assumption that all objects within one group are either mutually similar to each other or they are similar with respect to a common representative or Centroid. Clustering can also be based on mixture models [1]. In this approach, the data are assumed to be generated by several parameterized distributions (typically Gaussians). Distribution parameters are estimated using, for example, the expectation-maximation algorithm. Data points are assigned to different clusters based on their probabilities in the distributions. The implementation of clustering algorithms to mixed data is one of the challenging issues

IV. Conclusion

This proposed paper describes various algorithms presented by researchers for data clustering. Most of the real time applications need clustering of data. This data clustering can be implemented to mixed data which is the combination of numeric and strings. The clustering algorithm proposed in literature may have its own advantages and limitations. Developing an algorithm that meets all the requirements of the system is tangible. Different clustering algorithms like k-means, path-based clustering, clustering of self organized map are used widely for real world applications. The future work mainly concentrates on developing a clustering algorithm that meets all the requirements. Moreover, the future enhancement vision to develop a clustering algorithm that performs significantly well for mixed data set

V. REFERENCES

1) Juha Vesanto and Esa Alhoniemi, “Clustering of Self-Organizing Map,” IEEE Transactions on Neural Networks, vol. 11, no. 3, May 2000, pp. 586-600.
2) J. Shi and J. Malik, “Normalized Cuts and Image Segmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pp. 888-905, Aug. 2000.
3) Y. Gdalyahu, D. Weinshall, and M. Werman, “Self-Organization in Vision: Stochastic Clustering for Image Segmentation, Perceptual Grouping, and Image Database Organization,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 10, pp. 1053-1074, Oct. 2001.
4) J. C. Bezdek and S. K. Pal, Eds., “Fuzzy Models for Pattern Recognition: Methods that Search for Structures in Data,” New York: IEEE, 1992.
5) Mark Girolami, “Mercer Kernel-based Clustering in Feature space,” IEEE Transactions on Neural Networks, vol. 13, no. 3, May 2002.
6) Bernd Fischer, and J. M. Buhmann, “Path-Based Clustering for Grouping of Smooth Curves and Texture Segmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 4, April 2003.
7) Fischer, T. Zoller, and J.M. Buhmann, “Path Based Pair wise Data Clustering with Application to Texture Segmentation,” Energy Minimization Methods in Computer Vision and Pattern Recognition, pp. 235-250, LNCS 2134, 2001.
8) Bernd Fischer, and J. M. Buhmann, “Bagging for Path-Based Clustering,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 11, November 2003.
9) Leo Grady and Eric L. Schwartz, “Isoperimetric Graph Partitioning for Data Clustering and Image Segmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004.
10) Yaxin Bi, Sally McClean and Terry Anderson, “Improving Classification Decisions by Multiple Knowledge,” Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence, 2005.
11) Zhijie Xu, Laisheng Wang, Jiancheng Luo and Jianqin Zhang, “A Modified Clustering Algorithm Data Mining,” IEEE 2005.
12) Massimiliano Pavan and Marcello Pelillo, “Dominant Sets and Pairwise Clustering,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 1, January 2007.
13) M. Pavan and M. Pelillo, “Dominant Sets and Hierarchical Clustering,” Proceedings of IEEE International Conference Computer Vision, vol. 1, pp. 362-369, 2003.
14) M. Pavan and M. Pelillo, “Efficient Out-of-Sample Extension of Dominant-Set Clusters,” Advances in Neural Information Processing Systems 17,L.K. Saul, Y. Weiss, and L. Bottou, eds., pp. 1057-1064, 2005.
15) J. M. Buhmann, “Data Clustering and Learning,” Handbook of Brain Theory and Neural Networks, M. Arbib, ed., pp. 308-312, Bradfort Books/MIT Press, second ed., 2002.
16) A Tutorial on Clustering Algorithms,http://home .dei.polimi.it/matteucc/Clustering/tutorial_html.
17) Gautam Biswas, Jerry B. Weinberg, and Douglas H. Fisher, “ITERATE: A Conceptual Clustering Algorithm for Data Mining,” IEEE Transactions on Systems, Man, and Cybernetics, vol. 28, part c, no. 2, pp. 100-111, 1998.
18) M. N. Vrahatis, B. Boutsinas, P. Alevizos, and G. Pavlides, “The New k-Windows Algorithm for Improving the k -Means Clustering Algorithm,” Journal of Complexity, Elsevier, vol. 18, no. 1, pp. 375-391, 2002.
19) Eman Abdu, and Douglas Salane, “A spectral-based clustering algorithm for categorical data using data summaries,” International Conference on Knowledge Discovery and Data Mining, ACM, Article no. 2, 2009.
20) Shijin Li, Jing Liu, Yuelong Zhu, and Xiaohua Zhang, “A New Supervised Clustering Algorithm for Data Set with Mixed Attributes,” Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, vol. 2, pp. 844-849, 2007.
21) Jian Yin, Zhi-Fang Tan, Jiang-Tao Ren, and Yi-Qun Chen, “An efficient clustering algorithm for mixed type attributes in large dataset,” Proceedings of 2005 International Conference on Machine Learning and Cybernetics, vol. 3, pp. 1611-1614, 2005.
22) Tom Chiu, DongPing Fang, John Chen, Yao Wang, and Christopher Jeris, “A robust and scalable clustering algorithm for mixed type attributes in large database environment,” International Conference on Knowledge Discovery and Data Mining, pp. 263-268, 2001.
23) Mrutyunjaya Panda, and Manas Ranjan Patra, “Some Clustering Algorithms to Enhance the Performance of the Network Intrusion Detection System,” Journal of Theoretical and Applied Information Technology, pp. 710-716, 2008.
24) Qiang Li, Yan He, and Jing-ping Jiang, “A novel clustering algorithm based on quantum games,”

Journal of Physics A: Mathematical and Theoritical, no. 44, 2009.

25) Jie Li, Xinbo Gao, and Li-cheng Jiao, "A GA-Based Clustering Algorithm for Large Data Sets with Mixed Numeric and Categorical Values," Proceedings of the 5th International Conference on Computational Intelligence and Multimedia Applications, IEEE Computer Society, p. 102, 2003