International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622
Internationalconferecne on Industrial Automation and Computing (ICIAC- 12th & 13th April 2014)

RESEARCH ARTICLE OPEN ACCESS

Survey on Cluster Ensemble Approach for Categorical Data Clustering

Mr. P. O. Chitnis*, Mr. A. M. Bainwad **

*(Department of Computer Science and Engineering, SGGS&IT, SRTMU, Nanded, Maharashtra, India
** (Department of Computer Science and Engineering, SGGS&IT, SRTMU, Nanded, Maharashtra, India

ABSTRACT-
Clustering is a powerful technique in data mining, it is useful to determine structure the original data. Many well-established clustering algorithms, have been designed for numerical data, still, these algorithms cannot be used for categorical data. Clustering aims to categorize data into groups or clusters such that data in the one cluster present related to each other than to those data in different cluster. Traditional Clustering algorithms based on distance function such as Manhattan, Euclidean, etc... These distance function is used for to find similarity between the databases such that data points in the same partition are more similar than points in different partitions. The traditional clustering algorithms cannot apply on for categorical attributes. The categorical data are clustered and represented using the cluster ensembles. Cluster ensembles are used as the best alternative to the standard cluster analysis. The data set has been clustered by using any of the well-known cluster algorithm and represented as a cluster ensemble. The cluster ensembles generated a final data partition based on information and the information is prefect to make use of it.
Keywords – Categorical Data, Cluster Ensemble, Clustering, Data mining

I. INTRODUCTION

Data mining corresponds to extracting or mining knowledge from large amounts of data. The most important feature of data mining is that it deals with high dimensional data set. Clustering aims to categorize data into groups or clusters such that the data in the one cluster is related to each other than to those data present in different cluster. This requires the algorithms used in data mining to be scalable. However, most algorithms presently used in data mining do not scale well when applied to high dimensional data set because they were initially developed for other applications which involve low dimensional data set.

Each algorithm has its own strengths and weaknesses. For a particular data set, different algorithms, or even the same algorithm with different parameters, usually provide distinct solutions. Therefore, it is difficult for users to decide which algorithm would be the proper alternative for a given set of data. The No Free Lunch theorem suggests

[11], "There is no single clustering algorithm that performs best for all data sets".

Many well-established clustering algorithms, have been designed for numerical data, still, these algorithms cannot be used for categorical data. Clustering aims to categorize data into groups or clusters such that data in the one cluster is related to each other than to those data present in different cluster Categorical attributes is colour = {green, white, black, red} or employee = {regular , contract, Ad Hoc}. The special properties of categorical attributes, the clustering of categorical data seems more complicated than that of numerical data. Many algorithms concentrated on numerical data whose characteristic properties can be used to define a distance function between data points. However much of the data existed in the database are categorical where attribute values cannot be naturally ordered as numerical values.

For example, consider a market basket database containing one transaction per customer, each transaction containing the set of items purchased by the customer. The transaction data can be used to cluster the customers such that customers with similar buying patterns are in a single cluster. For example, one cluster may consist of predominantly married customers with infants who buy diapers, baby food, toys etc. (In addition to necessities like milk, sugar and butter), while another may consist of high-income customers that buy imported products like French and Italian wine, Swiss cheese and Belgian chocolate. The clusters can then be used to characterize the different customer groups, and these characterizations can be used in targeted marketing and advertising such that special products are directed towards special customer groups. The characterizations can also be used to predict buying patterns of new customers based on their proles. For example, it may be possible to conclude that high-income customers buy imported Foods and then mail customized catalogs for imported goods to only these

high-income customers .Clustering is the dynamic field of research in data mining. The choice of clustering algorithm depends both on the type of data available and on the particular purpose and application. The major clustering methods can be categorized into

- Hierarchical Algorithms
- Partitional Algorithms
- Density Based Algorithms
- Grid Based Algorithms
- Model Based clustering Algorithms

Problem of clustering becomes more challenging when the data is categorical, that is, when there is no inherent distance measure between data values. Categorical data has a different structure than the numerical data. The distance functions in the numerical data might not be applicable to the categorical data. Algorithms for clustering numerical data cannot be applied to categorical data.

II. RELATED WORK

Squeezer [1], well-organized algorithm for clustering categorical data, Squeezer, which gives the worth clustering results and at scalability. The Squeezer is a one-pass algorithm. Squeezer process the one data point in a cluster and then the succeeding data point is either put into an existing cluster or rejected to form a new cluster based on a given similarity function. Due to its characteristics, the suggested algorithm is fit for clustering data streams, where given a structure of points, the aim is to keep good clustering of the structure , using a small amount of memory and time. Outliers can also be controlled directly and efficiently in Squeezer. Squeezer is a clustering algorithm for categorical data .The basic idea of squeezer is simple. Squeezer repeatedly reads object from a dataset one by one. When the first object arrives, it forms a cluster. The next object are either put into exiting clusters or rejected by all existing clusters to from a new cluster by given a similarity function defined between an object and a cluster. Squeezer algorithm does not require the number of desired cluster as an input parameter. The parameter to be pre-specified is the value of similarity between the object and the cluster .Outliers can be handled efficiently and directly.

The time and space complexities of the Squeezer algorithm depend on the size of dataset. To simplify the analysis, we assume that the final number of clusters is k, and every attribute has the same number of distinct attribute values. Typical categorical attribute data considered for clustering consist of less than a hundred attribute values.

LIMBO [2] which is a hierarchical clustering algorithm that uses the Information Bottleneck (IB)

Framework to define a distance measure for categorical tuples. LIMBO has the advantage that it can produce clustering of different sizes in a single execution. The IB framework is used to define a distance measure for categorical tuples. LIMBO handles large data sets by producing a memory bounded summary model for the data. Clustering is a problem of major practical importance in numerous applications. The problem of clustering becomes more challenging when the data is categorical, that is, when there is no inherent distance measure between data values. As a result of its hierarchical approach, LIMBO allows us in a single execution to consider clustering of various sizes. Depending on the requirements of the user, LIMBO can control either the size, or the accuracy of the model it builds to summarize the data. LIMBO define a novel distance between attribute values that allows us to quantify the degree of interchangeability of attribute values within a single attribute.

Table 1 Comparison of Various Clustering Algorithms

Clustering Algorithms	Examples	Data Type	Cluster Shape	Data Sets	Function
Hierarchical Algorithms	ROCK	Mixed	Graph	Small size	Similarity Measure
	BIRICH	Numerical	Spherical	Large	Feature Tree
	CURE	Numerical	Arbitrary	Large	Similarity Measure
	LIMBO	Categorical	Spherical	Large	Distance
Partitional Algorithms	K-means	Numerical	Spherical	Large	Mean
	CLARA	Numerical	Arbitrary	Sample	Medoid
	CLARANS	Numerical	Arbitrary	Sample	Medoid
Density Based Algorithms	DENCLUE	Numerical	Arbitrary	Low Dimensional	Density Based
	DBSCAN	Numerical	Arbitrary	High Dimensional	Density Based
	OPTICS	Numerical	Arbitrary	Low Dimensional	Density Based
Grid Based Algorithms	CLIQUE	Mixed	Arbitrary	High Dimensional	Jacquard distance
Model Based clustering	STIRR	Categorical	Spherical	High Dimensional	Non-linear dynamical systems
	ROCK	Categorical	Spherical	High Dimensional	Hierarchical algorithm
	CLICK	Categorical	Spherical	High Dimensional	K-partite graphs

Genetic algorithm has also been adopted by a partitioning method for categorical data, i.e., GAClust [3]. The cobweb [4] is Model-based clustering methods are the bridge between the given data and some mathematical model. Such methods are often based on the assumption that the data are generated by a mixture of the underlying probability distributions.

STIRR [5] is used for clustering data sets, and its use to the data mining and data analysis of categorical data. "Categorical data is nothing but tables with fields that cannot by numerical value, the names of company, name of fruit, name of flower, and name of employee. STIRR, an iterative algorithm based on non-linear dynamical systems. They represent each attribute value as a weighted vertex in a graph. Edges between vertices derived from tuples in the dataset are not explicitly maintained.
Facilitates a type of similarity measure arising from the co-occurrence of values in the dataset.

Different graph models have also been investigated by the STIRR [5], ROCK [6], and CLICK [7] techniques. STIRR, an iterative algorithm based on non-linear dynamical systems. They represent each attribute value as a weighted vertex in a graph. Edges between vertices derived from tuples in the dataset are not explicitly maintained. ROCK (Robust Clustering using linKs) clustering algorithm which belongs to the class of agglomerative hierarchical clustering algorithms. The steps involved in clustering using ROCK are described as follows after drawing a random sample from the database, a hierarchical clustering algorithm that employs links is applied to the sampled points. Finally, the clusters involving only the sampled points are used to assign the still existing data points on disk to the appropriate clusters. CLICKS stands for the letters in Subspace CLusterIng of Categorical data via maximal K-partite cliques. ClICK uses different vertical method to search complete data sets. It is a most effective scalable technique used for the subspecies of high-dimensional datasets.

Several density-based algorithms have also been developed for clustering data by using distance function. CACTUS [8], COOLCAT [9], and CLOPE [10] are example of density based algorithms. A very fast summarization based algorithm called CACTUS that discovers exactly such clusters in the data. CACTUS has two important characteristics. First, the algorithm requires only two scans of the dataset, and hence is very fast and scalable. Second, CACTUS can find clusters in subsets of all attributes and can thus perform a subspace clustering of the data. CACTUS consists of three phases: summarization, clustering, and validation. In the summarization phase, we compute the summary information from

the dataset. In the clustering phase, we use the summary information to discover a set of candidate clusters. In the validation phase, we determine the actual set of clusters from the set of candidate clusters COOLCAT which is capable of efficiently clustering large data sets of records with categorical attributes, and data streams. COOLCAT is well equipped to deal with clustering of data streams nothing but continuously arriving streams of data point since it is an incremental algorithm capable of clustering new points without having to look at every point that has been clustered so far.

Although, a large number of algorithms have been introduced for clustering categorical data, the No Free Lunch theorem [11] proposes there is no single clustering algorithm that performs best for all data sets and can discover all types of cluster shapes and structures presented in data. Each algorithm has its own strengths and weaknesses. For a particular data set, different algorithms, or even the same algorithm with different parameters, usually provide distinct solutions. Therefore, it is difficult for users to decide which algorithm would be the proper alternative for a given set of data. Recently, cluster ensembles have emerged as an effective solution that is able to overcome these limitations, and improve the robustness as well as the quality of clustering results. The main objective of cluster ensembles is to combine different clustering decisions in such a way as to achieve accuracy superior to that of any individual clustering.

III. CLUSTER ENSEMBLES GENERATION METHODS

A set of clustering, find a single clustering that agrees as much as possible with the input clustering. An important issue in combining cluster is that this is particularly useful if they are different. This can be achieved by using different feature sets as well as by different training sets, randomly selected or based on a cluster analysis .Cluster ensemble is procedure gives more accurate result instead of those traditional clustering algorithms. In general, the result obtained from one clustering algorithm for a given data set after several runs give the same results. In such a situation where all ensemble cluster decide how a data set should be partitioned, aggregation of base clustering results without affecting the nature of a data set. The following ensemble generation methods yield different clustering of the same data, by exploiting different cluster models and different data partitions.

1.1 Homogeneous ensembles.

Cluster ensemble is a technique for improving stability and robustness of unsupervised classification [12]. Cluster ensemble follow a divided and combine strategy. First, the data is fragmented into a large number of small, spherical clusters, using algorithm with an arbitrary initialization of cluster centres. Multiple runs of the algorithm lead to different data partitions. Combination methods were applied to clustering ensembles obtained by K-means clustering [13], with random initialization of cluster centres, and a random selection of k in large intervals [14]. Clustering ensembles can go beyond what is typically achieved by a single clustering algorithm in several respects:

- **Robustness**. Better average performance across the domains and datasets.
- **Novelty.** Finding a combined solution unattainable by any single clustering algorithm.
- **Stability and confidence estimation**. Clustering solutions with lower sensitivity to noise, outliers or sampling variations. Clustering uncertainty can be assessed from ensemble distributions.
- **Parallelization and Scalability.** Parallel clustering of data subsets with subsequent combination of results.

1.2 Random-K

One of the most successful techniques is randomly selecting the number of clusters (k) for each ensemble member [16].Cluster ensemble or a single cluster will be able to determine the true structure in the data. Therefore, there might be an optimal number of clusters for the considered algorithm which is not necessarily the true number of clusters. The most successful heuristics has been randomly choosing the number of clusters assigned to each clustered in the ensemble [15].

1.3 Data subspace/sampling

Cluster analysis is very hard task for High dimensional data. Different clustering algorithms are established for low dimensional data, but as the dimensionality of the data increases, these algorithms have a tendency to failure [17]. The base clustering algorithm is applied to high dimensional data to get the single cluster ensemble. To get cluster ensemble of high dimensional data is divided into different data partitions by applying subspace [18]. Subspace is nothing but a different subset of the data or different features of a data set. In high dimensional data cluster, set of a given pair of data points is present in one cluster by applying same base clustering method. In other clusters the same data point is may be or may not be present. A common condition with high dimensional data is that some clusters may obtained from different subspaces of different combinations of features or attribute [19] [20]. In many practical approach, the combination of a different points as input subspace may obtained different cluster along

with set of dimensions, while points located in another cluster may form a tight group with respect to different dimensionsor by using different subspace. Each dimension could be relevant to at least one of the clusters. Common global dimensionality reduction techniques are unable to capture such local structure of the data [21]. Thus, a proper feature selection procedure should operate locally in input space. Local feature selection allows one to estimate to which degree features participate to the discovery of clusters. Such estimation is carried out using points within local neighborhoods, and it allows the embedding of adaptive distance measures in different regions of the input space.

1.4 Heterogeneous ensembles.

Heterogeneous ensembles are nothing but its combination of different clustering algorithms and base clustering is the combination of heterogeneous combination [23]. The different result of Base clustering can be generated by using different clustering algorithm with its different parameter setting such as the number of clusters, resampling, and randomization of the sample data. Many clustering algorithms are available, and different clustering algorithms may produce different clustering results due to data characteristics and experimental assumptions. A heterogeneous clustering ensemble method that uses a genetic algorithm [22] to generate good quality of cluster and clustering results with characteristics of data.

IV. CONCLUSION

There is no single clustering algorithm that gives worthy results for all types of data sets. Each algorithm has its own strengths and weaknesses. For a particular data set, different algorithms, or even the same algorithm with different parameters, usually provide distinct solutions. Therefore, it is difficult for users to decide which algorithm would be the proper alternative for a given set of data. Cluster ensemble has proved to be a decent alternative when facing cluster analysis problems. To overcome such type of a problem we use the cluster ensemble method. It proved solution when facing such problems. Cluster ensemble is a process combining base clustering result into one final cluster. The aim of this method to improve robustness and quality of data clustering.

REFERENCES

[1] Z. He, X. Xu, and S. Deng, "Squeezer: An Efficient Algorithm for Clustering Categorical Data," J. Computer Science and Technology, vol. 17, no. 5, pp. 611-624, 2002.

[2] P.Andritsos and V.Tzerpos, "Information-Theoretic Software Clustering," IEEE Trans. Software Eng., vol. 31, no. 2, pp. 150-165, Feb. 2005.

[3] D. Cristofor and D. Simovici, "Finding Median Partitions Using Information-Theoretical-Based Genetic Algorithms," J. Universal Computer Science, vol. 8, no. 2, pp. 153-172, 2002.

[4] D.H. Fisher, "Knowledge Acquisition via Incremental Conceptual Clustering," Machine Learning, vol. 2, pp. 139-172, 1987.

[5] D. Gibson, J. Kleinberg, and P. Raghavan, "Clustering Categorical Data: An Approach Based on Dynamical Systems," VLDB J., vol. 8, nos. 3-4, pp. 222-236, 2000.

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

[7] M.J. Zaki and M. Peters, "Clicks: Mining Subspace Clusters in Categorical Data via Kpartite Maximal Cliques," Proc. Int'l Conf. Data Eng. (ICDE), pp. 355-356, 2005.

[8] V. Ganti, J. Gehrke, and R. Ramakrishnan, "CACTUS: Clustering Categorical Data Using Summaries," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 73-83, 1999.

[9] D. Barbara, Y. Li, and J. Couto, "COOLCAT: An Entropy-Based Algorithm for Categorical Clustering," Proc. Int'l Conf. Information and Knowledge Management (CIKM), pp. 582-589, 2002.

[10] Y. Yang, S. Guan, and J. You, "CLOPE: A Fast and Effective Clustering Algorithm for Transactional Data," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 682-687, 2002.

[11] D.H. Wolpert and W.G. Macready, "No Free Lunch Theorems for Search," Technical Report SFI-TR-95-02-010, Santa Fe Inst., 1995.

[12] A. Gionis, H. Mannila, and P. Tsaparas, "Clustering Aggregation," Proc. Int'l Conf. Data Eng. (ICDE), pp. 341-352, 2005.

[13] A.P. Topchy, A.K. Jain, and W.F. Punch, "A Mixture Model for Clustering Ensembles," Proc. SIAM Int'l Conf. Data Mining, pp. 379-390, 2004.

[14] S. Monti, P. Tamayo, J.P. Mesirov, and T.R. Golub, "Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data," Machine Learning, vol. 52, nos. 1/2, pp. 91-118, 2003.

[15] A.L.N. Fred and A.K. Jain, "Combining Multiple Clusterings Using Evidence Accumulation," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 6, pp. 835-850, June 2005.

[16] L.I. Kuncheva and D. Vetrov, "Evaluation of Stability of K-Means Cluster Ensembles with Respect to Random Initialization," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 11, pp. 1798-1808, Nov. 2006.

[17] A.P. Topchy, A.K. Jain, and W.F. Punch, "Clustering Ensembles: Models of Consensus and Weak Partitions," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 12, pp. 1866-1881, Dec. 2005.

[18] X.Z. Fern and C.E. Brodley, "Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach," Proc. Int'l Conf. Machine Learning (ICML), pp. 186-193, 2003.

[19] A. Strehl and J. Ghosh, "Cluster Ensembles: A Knowledge Reuse Framework for Combining Multiple Partitions," J. Machine Learning Research, vol. 3, pp. 583-617, 2002.

[20] B. Fischer and J.M. Buhmann, "Bagging for Path-Based Clustering," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 25, no. 11, pp. 1411-1415, Nov. 2003.

[21] B. Minaei-Bidgoli, A. Topchy, and W. Punch, "A Comparison of Resampling Methods for Clustering Ensembles," Proc. Int'l Conf. Artificial Intelligence, pp. 939-945, 2004.

[22] X. Hu and I. Yoo, "Cluster Ensemble and Its Applications in Gene Expression Analysis," Proc. Asia-Pacific Bioinformatics Conf., pp. 297-302, 2004.

[23] M. Law, A. Topchy, and A.K. Jain, " Multiobjective Data Clustering," Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 424-430, 2004.