The International Journal Of Engineering And Science (IJES)
||Volume|| 3 ||Issue|| 2 ||Pages|| 01-05 || 2014 ||
ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805

A Relational Graph Based Approach using Multi-Attribute Closure Measure for Categorical Data Clustering

[1,] K.Nagarajan, [2,] Dr.M.Prabakaran,
[1,] *Research Scholar of Bharathidasan University, Assistant Programmer,*
Department of Computer Science, Government Arts College, Ariyalur, Tamilnadu, India.
[2,] *Research Advisor, Assistant Professor,*
Department of Computer Science, Government Arts College, Ariyalur, Tamilnadu, India.

ABSTRACT

Conventional clustering approaches suffer with the scalability of number of attributes based on which the clustering is performed. There are approaches to cluster data points with multiple attributes but suffers with overlapping and multiple iteration needed to perform clustering , also the measure computed for the variation of data points between cluster also will not be effective when doing with multiple attributes. To overcome the problem identified we propose a new graph based approach which represents the relation between the data points and clusters. The relational graph consists of various vertices and edges, each vertex represent a data point. There will be an edge between two different edges only if there is a multi-attribute closure between them. We compute the attribute closure, using all the attributes of the data points. We use threshold method to select the data point has closure to other one and the value of threshold is set based on number of attributes the data point has. The proposed method produces good results compare to other approaches discussed in this era and we have evaluated the proposed method with various data sets.

INDEX TERMS: *Categorical Data, Clustering, Data Mining, Multi-Attribute Closure.*

Date of Submission: 08 January 2014 Date of Acceptance: 15 February 2014

I. INTRODUCTION:-

As the technology development rising at each moment of life the people started using various ways to find information from the knowledge base. Clustering is one of the machine learning method to group similar information as a subset , so that to facilitate the retrieval of information as easier one and to provide good support for the learner. There has been various standard methods in the domain of information mining and grouping. K-Means a popular method to group numerical data points which takes set of parameters as an argument with the input data set and returns a number of subset of data points as mentioned in the parameter. The k-means algorithm takes argument as how many number of clusters has to be formed and it computes distance measure using Euclidean distance or any of the method to compute the distance between data points. The process computing distance and shifting the data point from one group to another will be iterated until there is no data point has to be shifted from a cluster to another. This process will be controlled by an input parameter specified by number of iteration. The quality of clustering is depend on the method of distance computing and number of iterations. Mainly k-means will be useful for single attribute based clustering approaches, while this will not be efficient and useful in case of multi attribute based clustering approach.

Categorical data is one which contains as many numbers of attributes and all are discrete without any ordering. For example the attributes of normal human being are HD={sex, marital} which has discrete values of sex={male, female} and marital={married, unmarried}. Like this any data point can have as many numbers of attributes but clustering these data points are difficult in nature.

There are many number of algorithms have been introduced for clustering categorical data, each 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. For example the feature-based approach transforms the problem of cluster ensembles to clustering categorical data (i.e., cluster labels), the direct approach that finds the final partition through relabeling the base clustering results, graph-based algorithms that employ a graph partitioning methodology and. the pair wise-similarity approach that makes use of co-occurrence relations between data points.

II. BACKGROUND:

There are many approaches has been discussed in the past for clustering of categorical datas, we explore few of them in this chapter to find the prose and corners of the methods proposed.

A Robust Clustering Algorithm for Categorical Attribute [1] makes use of a link graph, in which nodes and links represent data points (or tuples) and their similarity, respectively. Two tuples are similar if they shared a large number of attribute values. Note that the link connecting two nodes is included only when the corresponding similarity exceeds a user-defined threshold. With tuples being initially regarded as singleton clusters, ROCK merges clusters in an agglomerative hierarchical fashion, while optimizing a cluster quality that is defined in terms of the number of links across clusters. Note that the graph models used by ROCK and LCE are dissimilar—the graph of data points and that of attribute values (or clusters), respectively. Since the number of data points is normally greater than that of attribute values, ROCK is less efficient than LCE. As a result, it is unsuitable for large datasets [2]. Also, the selection of a "smooth function" that is used to estimate a cluster quality is a delicate and difficult task for average users [4].

CACTUS [3] also relies on the co-occurrence among attribute values. In essence, two attribute values are strongly connected if their support (i.e., the proportion of tuples in which the values co-occur) exceeds a prespecified value. By extending this concept to all attributes, CACTUS searches for the "distinguishing sets," which are attribute value sets that uniquely occur within only one cluster. These sets correspond to cluster projections that can be combined to formulate the final clusters.

Consensus Clustering [8], the cluster the data points using three consensus like Iterative Voting Consensus, Iterating Probability Consensus and Iterative pairwise consensus. This method perfoms the three steps to form the final cluster. This is and cluster ensample approach which is not using the underlying data features of the data points.

In [10] an weighted ensample approach is discussed , it proposes two algorithms namely WSPA and WBPA provide as output a partition of the data into k clusters, with no information regarding feature relevance for each of the clusters. Clustering ensemble algorithm (WSBPA) that provides weighted clusters in output. This technique advances the WBPA method by adding to the final partition weighted features associated with each cluster. By assigning a value to each dimension, WSBPA captures the local relevance of features within each cluster. Thus, the structure of the output provided by a single run of LAC is preserved.

Combining multiple clustering using evidence accumulation [11], produces a clustering ensemble - a set of object partitions from a data set (n objects or patterns in d dimensions). To produce the cluster ensample different ways are used as:

(1)- applying different clustering algorithms, and (2)- applying the same clustering algorithm with different values of parameters or initializations. Further, combinations of different data representations (feature spaces) and clustering algorithms can also provide a multitude of signicantly different data partitionings. They propose a simple framework for extracting a consistent clustering, given the various partitions in a clustering ensemble. According to the EAC concept, each partition is viewed as an independent evidence of data organization, individual data partitions being combined, based on a voting mechanism, to generate a new n £ n similarity matrix between the n patterns. The final data partition of the n patterns is obtained by applying a hierarchical agglomerative clustering algorithm on this matrix.

Refining Pairwise Similarity Matrix for Cluster Ensemble Problem with Cluster Relations[12], used two new similarity matrices, which are empirically evaluated and compared against the standard co-association matrix on six datasets (both artificial and real data) using four different combination methods and six clustering validity criteria. In all cases, the results suggest the new link-based similarity matrices are able to extract efficiently the information embedded in the input clusterings, and regularly suggest higher clustering quality in comparison to their competitor.

Link-Based Cluster Ensemble Approach [13], proposed a novel technique for Categorical Data Clustering. It is more efficient than the former model, where a BM-like matrix is used to represent the ensemble information. The focus has shifted from revealing the similarity among data points to estimating those between clusters. A new link-based algorithm has been specifically proposed to generate such measures in an accurate, inexpensive manner. It uses weighted graph to identify the links between the clusters.

Most of the proposed methodologies are prevalent in nature suffers with initial set of clustering ie identifying the ensamples. The ensamples are generated based on the method of existing algorithm like k-means or some other algorithm. They perform clustering iteratively in the given ensamples. But all the methods are suffers with identifying the number of clusters has to be formed. We propose a new graph based method for clustering of categorical data which could identify and form number of clusters dynamically.

We are given a set of N data points X ={x1; x2; .. x_n} and a set of Clustering C={C_1, $C_{2\,..}$ C_n} of the data points in X. Each clustering Ci is a mapping from X to {1…, nxi} where nx_i is the number of clusters in C.

III. PROPOSED SYSTEM

The relational graph based categorical data clustering consists of following steps: 1. Initial Clustering 2. Relational Graph Based Clustering 3. Cluster Validation.

3.1 Initial Clustering

The proposed system produces initial clustering based on multi attribute variance method, generated variance and number cluster will be given and could be modified according to the requirement of the user. Given N data samples, we compute attribute based variance Dev_a for each attribute of $A\{a1,a2,\ldots a_n\}$. once we compute the attribute variance then number of possible clustering is computed using the cumulative attribute variance and that many number of cluster will be formed. The same variance will be used to identify the cluster of data point $D\{d_i\}$.

Algorithm:

```
Step1: Read input data set D.
Step2: identify number of data samples N.
Step3: initialize sd , AF, AT, csd, psd.
Step3: for each attribute $A_i$ of $d_i$ from D
        SA = sort(A(D)).
        For each $SA_i$
            cumulative variance Sdn = $\mu$ ($\sum\sigma^2$ ($A_0$ $A_{1,}$ An).)
            Csd=sdn; // current variance.
            Psd = sdn. // previous variance.
            If csd>psd then
                sd(i) = sdn .
                AF(i)=0;
                AT(i)= n.
                Sdn=0;
                I=n.
            End
    End
    Step4: Nc=$\Delta$(sd)
    Step5: create cluster C={c1,c2..$c_{Nc}$};
    Step6: assign labels to data points.
    Step7: stop
```

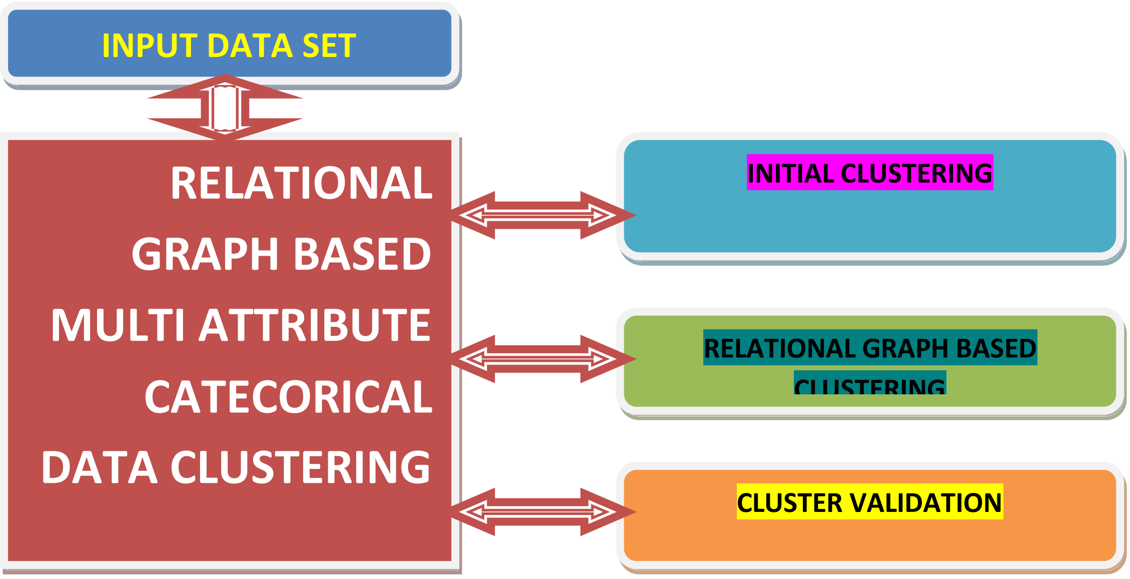

INPUT DATA SET
RELATIONAL GRAPH BASED MULTI ATTRIBUTE CATECORICAL DATA CLUSTERING
INITIAL CLUSTERING
RELATIONAL GRAPH BASED CLUSTERING
CLUSTER VALIDATION

Fig1: Proposed System Architecture.

3.2 Relational Graph Clustering:

The relational graph is constructed using the multi attribute closure (MAC) measure. The MAC is computed as the sum of number of attribute values and the average of them. This represents the overall similarity of the data points and clusters. With the computed MAC the graph is constructed. Given D data points, a relation graph G=(V,MAC) can be constructed where v is the set of vertices which denotes the set of clusters and MAC denotes the multiple attribute closure which represents the closeness of the clusters.

Algorithm:

```
Step1: start
Step2: read initial clusters C.
Step3: create graph G.
Step3: for each cluster $C_i$ from C
            Construct a G(V) vertex in G.
            Extract data points D from $C_i$
            Extract data points D1 from $C_{i+1}$
            Compute MAC = $\sum\phi(D(A_i)-D1(A_I))/\Omega(D)$.
            $\Phi$- sum of all equivalence of attributes of data points of clusters.
            $\Omega$ - total number of data points from both the clusters.
    End
Step4: if MAC>Cth
            Perform clustering between two clusters.
            End
Step5: stop.
```

3.3 Cluster Validation

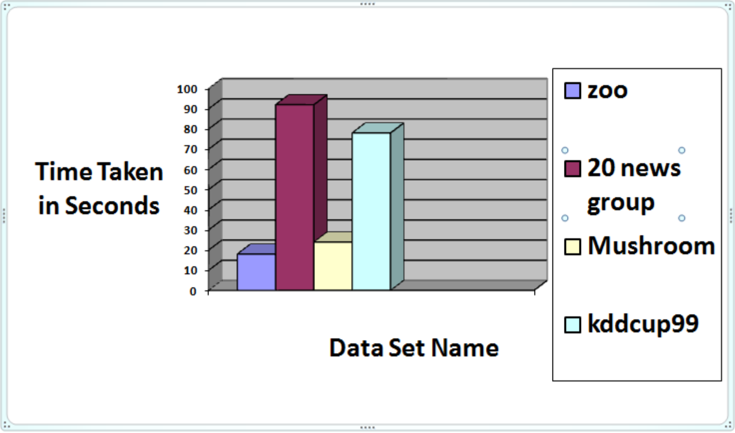
The validation of the cluster is performed by giving new inputs to the cluster algorithm. The algorithm is evaluated by the cluster label assigned to the data point and verified with the feature of the data points. This procedure is performed iteratively to identify the overlapping members of the cluster.

1. Results and Discussion:

The proposed method produced efficient results with various data sets. We have evaluated the algorithm with the following data sets.

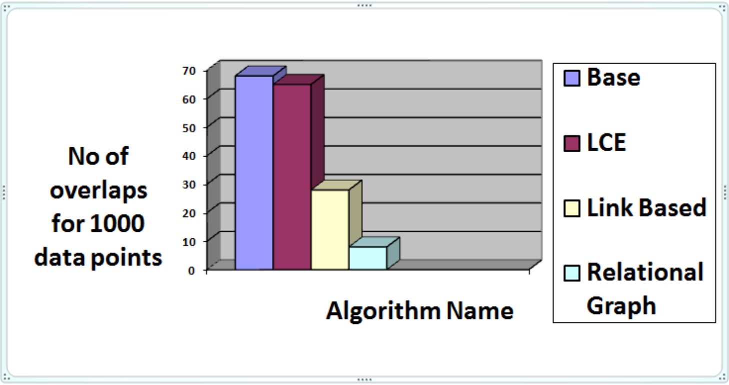
Dataset	Number of Data Points (N)	Attributes (d)	Attribute Values (AA)	Classes (K)
zoo	101	16	36	7
20 news group	1000	6084	12,168	2
Mushroom	8124	22	117	2
KddCup99	1,00,000	42	139	20

TABLE 1: Description of Data Sets



Graph 1: shows the time complexity of the proposed system.

The graph1 shows the time taken by the proposed method and each data set has different number of data points with different number of attributes. Even the kddcup99 has 1,00,000 data points it is still has very few ie only 42 attributes so that it takes less time than 20 news group with 1000 records with 6800 attributes. The proposed method takes less time than other algorithms.



Graph2: shows the data point overlap of different algorithm

From the graph 2, It is very clear that the number of overlaps produced by the proposed method is very less than other algorithms at the first iteration and will be removed at the next iteration.

CONCLUSION:

The relational graph based multi attribute closure based clustering produced good results. The multi attribute closure has affected the quality of clustering very effectively other than previous measures used by other algorithms. The attribute closure measure represent the closeness of the data points and their attribute values , so that the proposed method has done the clustering effectively and also the evaluation results shows the effectiveness of the proposed system. The proposed methodology can still improved with few other closure measures.

REFERENCES:

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] L.I. Kuncheva and S.T. Hadjitodorov, “Using Diversity in Cluster Ensembles,” Proc. IEEE Int’l Conf. Systems, Man and Cybernetics, pp. 1214-1219, 2004.

[6] H. Xue, S. Chen, and Q. Yang, “Discriminatively Regularized Least-Squares Classification,” Pattern Recognition, vol. 42, no. 1, pp. 93-104, 2009.

[7] A. Gionis, H. Mannila, and P. Tsaparas, “Clustering Aggregation,” Proc. Int’l Conf. Data Eng. (ICDE), pp. 341-352, 2005.

[8] N. Nguyen and R. Caruana, “Consensus Clusterings,” Proc. IEEE Int’l Conf. Data Mining (ICDM), pp. 607-612, 2007.

[9] 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.

[10] C. Domeniconi and M. Al-Razgan, “Weighted Cluster Ensembles: Methods and Analysis,” ACM Trans. Knowledge Discovery from Data, vol. 2, no. 4, pp. 1-40, 2009.

[11] 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.

[12] N. Iam-On, T. Boongoen, and S. Garrett, “Refining Pairwise Similarity Matrix for Cluster Ensemble Problem with Cluster Relations,” Proc. Int’l Conf. Discovery Science, pp. 222-233, 2008.

[13] T. Boongoen, Q. Shen, and C. Price, “Disclosing False Identity through Hybrid Link Analysis,” Artificial Intelligence and Law, vol. 18, no. 1, pp. 77-102, 2010.

[14] L. Getoor and C.P. Diehl, “Link Mining: A Survey,” ACM SIGKDD Explorations Newsletter, vol. 7, no. 2, pp. 3-12, 2005.

[15] D. Liben-Nowell and J. Kleinberg, “The Link-Prediction Problem for Social Networks,” J. Am. Soc. for Information Science and Technology, vol. 58, no. 7, pp. 1019-1031, 2007.

[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] Z. Yu, H.-S. Wong, and H. Wang, “Graph-Based Consensus Clustering for Class Discovery from Gene Expression Data,” Bioinformatics, vol. 23, no. 21, pp. 2888-2896, 2007.

[18] M. Al-Razgan, C. Domeniconi, and D. Barbara, “Random Subspace Ensembles for Clustering Categorical Data,” Supervised and Unsupervised Ensemble Methods and Their Applications, pp. 31-48, Springer, 2008.

[19] Z. He, X. Xu, and S. Deng, “A Cluster Ensemble Method for Clustering Categorical Data,” Information Fusion, vol. 6, no. 2, pp. 143-151, 2005.

[20] R. Agrawal, T. Imielinski, and A. Swami, “Mining Association Rules between Sets of Items in Large Databases,” Proc. ACM SIGMOD Int’l Conf. Management of Data, pp. 207-216, 1993.

[21] P.N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Addison Wesley, 2005.

[22] E. Minkov, W.W. Cohen, and A.Y. Ng, “Contextual Search and Name Disambiguation in Email Using Graphs,” Proc. Int’l Conf. Research and Development in IR, pp. 27-34, 2006.