IJDCST @October Issue- V-1, I-6, SW-42
ISSN-2320-7884 (Online)
ISSN-2321-0257 (Print)

Optimized Clustering Process for Dynamic Data Documents

P.Raja Sekhar[1], Dr J.Srinivas Rao[2]

[1]Student, Nova College of Engineering and Technology, Ibrahimpatnam, Krishna Dist, Andhra Pradesh, India

[2] Professor, Nova College of Engineering and Technology, Ibrahimpatnam, Krishna Dist, Andhra Pradesh, India

Abstract: Data clustering is a primary tool for understanding the structure of data sets. It application domain includes machine learning, data mining, information retrieval, and pattern recognition etc. Clustering aims to categorize data into groups or clusters such that the data in the same cluster are more similar to each other than to those in different clusters. Although conventional algorithms includes k-means clustering and expectation maximization (EM) clustering, PAM etc and different clustering ensemble approaches were used for clustering process they have limitations in handling unrelated entries in dataset resulting in a detrimental performance. So previously a link-based algorithm which is a two stage process involving generation of a conventional matrix by discovering unknown entries through similarity between clusters in an ensemble, and then obtaining a weighted bipartite graph from this refined matrix. We observed the construction of the weighted bipartite graph generation is irrespective of the size of the matrix. For an optimized performance we propose to use ACO (ant colony optimization) Algorithm to Solve Minimum-Weighted Bipartite Matching for a smaller refined matrix and Metropolis Algorithm for Maximum-Weighted Bipartite Matching for a larger refined matrix. This kind of an adaptive approach to varying matrix sizes rather than a single static approach to all matrix sizes determines the optimization parameter such as timescales involved in data clustering process. An implementation of the proposed system validates our claim.

Index Terms: Data clustering, Cluster ensemble, Link-based similarity measure, Data sets, Minimum weighted bipartite matching, and Maximum weighted bipartite matching.

I. INTRODUCTION

Data clustering is the one fundamental concept for understanding representation and structure of the data sets. It plays main role in representation process for arranging records in sequential order

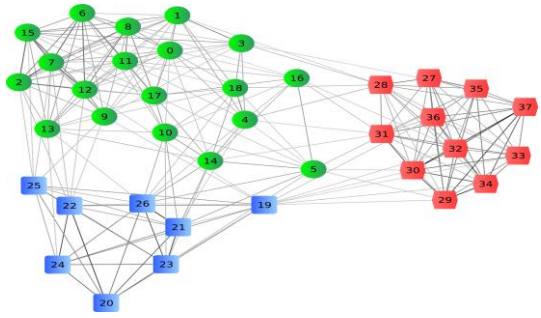

Figure1: Clustering Process Architecture.

For this representation of data concepts present in the data base we are most number of data clustering algorithms for arranging those datasets in equal centroid points with each portioning. k-means algorithms were used for data clustering process in data clustering. In data clustering using k-means working procedure can be calculated using inherit the properties from datasets. In this region after using k-means clustering algorithm traditionally we are developed a Link based clustering algorithm for arranging data clustering datasets on categorical data. 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.

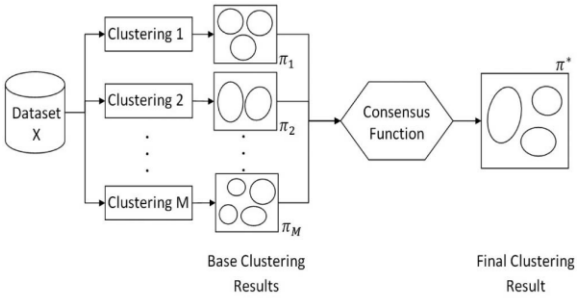

Figure 2: Data Clustering in link based clustering algorithm based on cluster ensemble.

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. A link based clustering process has a problem on data in context based clustering applications. In this paper we propose to use Ant Colony Optimization and Metropolis algorithm to solve the matching problem in context based clustering applications. We randomly select the matching edge in each iteration to generate an approximate solution.

II. BACK GROUND WORK

Clustering is a data mining technique used to place data elements into related groups without advance knowledge of the group definitions. Popular conventional algorithms include k-means clustering and expectation maximization (EM) clustering, PAM etc. However, these cannot be directly applied for clustering of categorical data, where domain values are discrete and have no ordering defined. Although, a large number of algorithms have been introduced for clustering categorical data, the "No Free Lunch" theorem suggests 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. Due to their inefficiency different clustering ensemble approaches (Homogeneous ensembles, Random-k, Data subspace/sampling, Heterogeneous ensembles, Mixed heuristics) to obtain data clusters were developed and used. Clustering ensembles combine multiple partitions of the given data into a single clustering solution of better quality. Works well for all datasets. Users need not choose the clustering filtration manually.

III. PROPOSED APPROACH

The underlying ensemble-information matrix presents only cluster-data point relations, with many entries being left unknown. Ignoring dataset unsolvable entries during clustering degrades the quality of the clustering result. Proposes a new link-based algorithm which is a two stage process. First it improves the conventional matrix by discovering unknown entries through similarity between clusters in an ensemble. Then to obtain the final clustering result, a graph partitioning technique is applied to a weighted bipartite graph that is formulated from the refined matrix. Obtained clustering results suggest that the proposed link-based method usually achieves superior clustering results compared to those of the traditional categorical data algorithms and prior cluster ensemble techniques. A weighted bipartite graph is formulated from the refined matrix obtained from link based cluster ensemble. The construction of the weighted bipartite graph is irrespective of the size of the matrix. For an optimized performance we propose to use. ACO (ant colony optimization) Algorithm to Solve Minimum-Weighted Bipartite Matching for a smaller refined matrix. Metropolis Algorithm for Maximum-Weighted Bipartite Matching for a larger refined matrix. This kind of an adaptive approach to varying matrix sizes rather than a single static approach to all matrix sizes determines the optimization parameter such as timescales involved in data clustering process.

IV. ANT COLONY OPTIMIZATION

In this algorithm we propose to observe the behaviour ants presented in dataset records.

Step1: While(terminate condition not met){

do

generate solutions

pheromone update

}end while

Figure 3: Algorithm for counting ants with similarity matching.

It simulates the ants and the pheromone evaporation on the paths. In each iteration, the ants select the paths according to the density of pheromone deposited. If the path has higher density, it will be selected with higher probability. After finding a path, the ants lay down pheromone on the path. In the ACO algorithm, each path represents a solution. Besides, the pheromone evaporates gradually. The ACO algorithms nearly used for accessing optimization solutions

V. PERFORMANCE ANALYSIS

In this section we are introducing two methods on solving weighing and bipartite matching. We find out the comparison results between every data set as a categorical data. We compare complexity of the of the existing link based algorithm process and our proposed algorithm process in dataset present data. Using our proposed approach we are solving a problem of TSP (travelling salesman problem). In the TSP problem, there are n cities. Each city has to be visited exactly once, and the tour ends at the starting city. The problem is to find a shortest tour to visit these n cities. Let dij be the distance between the city i and the city j and T_{ij} be the pheromone on the edge connects i and j. Each of the m ants decides independently on the city to be visited next.

They base their decision on the density of pheromone T_{ij} and a heuristic function N_{ij}. The probability of the above procedure can be calculated as follows:

$P^k_{ij} = [T_{ij}]^{\alpha}(N_{ij})^{\beta}/\sum l \epsilon N_l^k [T_{ij}]^{\alpha}(N_{ij})^{\beta}$.

In the above α, β are the parameters passing through the threshold value present in data set. By using this procedure can be introduced as follows:

Step1: Load Input files as data sets.

Step2: Apply data clustering process Ant Colony Optimization algorithm on categorical data.

Step3: Calculate the centroid using each record present data set.

Step4: Centroid formation for each node with weighted matrix present in data set.

Step 5: Display portioning details of clustering in categorical data.

Figure 4: Process of clustering.

As shown in the above figure formation of cluster ensemble with equal rotation between every node present in dataset. Those results are arranged in partitioning order by collecting all the cluster formed using data set.

VI. CONCLUSION

a link-based algorithm which is a two stage process involving generation of a conventional matrix by discovering unknown entries through similarity between clusters in an ensemble, and then obtaining a weighted bipartite graph from this refined matrix. We observed the construction of the weighted bipartite graph generation is irrespective of the size of the matrix. For an optimized performance we propose to use ACO (ant colony optimization) Algorithm to Solve Minimum-Weighted Bipartite Matching for a smaller refined matrix and Metropolis Algorithm for Maximum-Weighted Bipartite Matching for a larger refined matrix. This kind of an adaptive approach to varying matrix sizes rather than a single static approach to all matrix sizes determines the optimization parameter such as timescales. As further

work of our proposed work more context based clustering can be arranged in sequential order effectively to decrease the complexity present in the data set.

VII.REFERENCES

[1] Prof. Yuh-Dauh Lyuu Hung-Pin Shih," Two Algorithms for Maximum and Minimum Weighted Bipartite Matching", Proceedings of the Inaugural Workshop on Artificial Life (AL'01), pp. 70–78, Adelaide, Australia, December 2001.

[2] Natthakan Iam-On, Tossapon Boongoen, Simon Garrett, and Chris Price," A Link-Based Cluster Ensemble Approach for Categorical Data Clustering", IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 3, March 2012.

[3] T. St¨utzle, M. Dorigo. (1999) ACO Algorithms for the Traveling Salesman Problem. In K. Miettinen, M. Makela, P. Neittaanmaki, and J. Periaux, editors, Evolutionary Algorithms in Engineering and Computer Science. Wiley, 1999.

[4] N. Iam-On, T. Boongoen, and S. Garrett, "Refining Pair wise Similarity Matrix for Cluster Ensemble Problem with Cluster Relations," Proc. Int'l Conf. Discovery Science, pp. 222-233, 2008.

[5] 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.

[6] E. Abdu and D. Salane, "A Spectral-Based Clustering Algorithm for Categorical Data Using Data Summaries," Proc. Workshop Data Mining using Matrices and Tensors, pp. 1-8, 2009.

[7] 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.

[8] 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.