Journal of Information and Operations Management
ISSN: 0976–7754 & E-ISSN: 0976–7762 , Volume 3, Issue 1, 2012, pp-68-69
Available online at http://www.bioinfo.in/contents.php?id=55

AN EXPERIMENTAL STUDY OF TWO CONSENSUS CLUSTERING ALGORITHMS ON GRAPH DATA

RAO P.R.[1*] AND ANNIE RAJAN [2]

[1]Department of Computer Science and Technology, Goa University, Goa-403206. India.
[2]Department of I.T. and Computer Science, Dhempe College of Arts and Science, Panaji, Goa, India.
*Corresponding Author: Email– pralhadrrao@gmail.com

Received: December 12, 2011; Accepted: January 15, 2012

Abstract- In this paper a comparison of two consensus clustering methods is performed. The two methods studied in this paper are strong pattern method and bipartite graph method. The data is a graph on twenty vertices of densities of 20%, 40%, 70% and 90%, where each of them is a connected graph. For each input graph corresponding to a density value, we generated three similarity matrices namely, clustering coefficient based method, neighborhood based method and edge-betweeness based method (or shortest path method). Each similarity matrix is used to obtain the base clusters using CLUTO software. The methods used to obtain base clusters are, repeated bisection, direct K-way partitioning, agglomerative hierarchical algorithm and graph K-way partitioning. For each similarity matrix and for each partitioning method, partitions are generated in the range 9 to 11. These partitions are used to generate two consensus clusters by the two methods stated above. For each consensus cluster, we calculated accuracy and diversity using adjusted rand index. Our experimental study indicates that strong pattern based method of consensus clustering has higher accuracy. When compared with other three base algorithms, the highest accuracy value is observed for 20% density, for edge-betweeness similarity matrix, with base cluster obtained by graph K-way partitioning.
Keywords - *Clustering, Consensus cluster, Accuracy, Diversity, Graph data, Graph density.*

Citation: Rao P.R. and Annie Rajan (2012) An Experimental Study of Two Consensus Clustering Algorithms on Graph Data. Journal of Information and Operations Management ISSN: 0976–7754 & E-ISSN: 0976–7762, Volume 3, Issue 1, pp-68-69.

Intoduction

Clustering is an important issue in the exploration of data. There are wide areas of applications of clustering such as data mining, VLSI design and gene analysis. Clustering consists of discovering natural groups of similar elements in data sets. The various methods of graph clustering used in this paper are repeated bisection, direct K-way partitioning, agglomerative hierarchical clustering, and graph K-way partitioning. Diversity measure are used as a parameter to determine best consensus cluster. To summarize, the main contributions of this paper include the following; (1) used three topological distance matrices (2) four base clustering methods used in CLUTO software to generate base clusters. (3) two consensus clustering algorithms to find consensus clusters (4) used diversity measure as parameter for determining accuracy. (5) input graphs of varying densities have been generated for experimental study.

Related Work

Three graph clustering methods namely Markov clustering, iterative conductance cutting, and geometric MST clustering are presented in [7]. They have conducted experimental study, to study the performance of these algorithms. For a survey on graph clustering refer to [8]. To generate base clusters from all the four methods, we use CLUTO software. The problem of designing a consensus cluster using bipartite graph partitions is studied in [9]. A bipartite graph is constructed from a given cluster ensemble; this bipartite graph is partitioned; the resulting partition is the consensus cluster. The methods of determining consensus cluster using spanning tree and strong patterns is studied in [10]. They have determined which distance matrix performs well for the given set of partitions. Three techniques to determine cluster ensembles of weighted clusters have been developed in [11].

Algorithm

The general framework of our approach is given below.

a. Input is a graph G= (V, E), where V is set of vertices and E is set of edges of G.
b. Density of graph is obtained (from sparse to dense graph i.e 20%, 40%, 70%, 90%).
c. For each density, calculate the similarity matrix.
d. Each similarity matrix is applied to each base algorithm.
e. Each base algorithm partitions are generated.
f. For the partitions obtained from each of the base algorithm, consensus clusters are obtained.
g. Diversity measure is calculated by comparing individual clustering results with each consensus result.

Similarity Matrix

The similarity matrices are generated by calculating the similarity between two vertices. Methods to determine similarity between two vertices are explained below.

a. Clustering Coefficient based method.
b. Neighborhood based method.
c. Edge Betweenness-based

The three metrics capture different properties of the topology of the graph. The partitions from the three metrics are considered separately.

Base Algorithms

Base Graph clustering algorithms are used to obtain the base clusters. Here we use four *methods* for clustering. We have used the implementation available from CLUTO, which is a clustering package. Each of these algorithms takes an input similarity matrix and outputs partitions.

a. Repeated bisection (rbr)
b. Direct k - way partitioning (direct)
c. Agglomerative Hierarchical Algorithm (agglo)
d. Graph k-ways partitioning (graph)

Consensus Clustering Methods

In this section we are describing the consensus clustering methods used in this paper.

Strong pattern graph

Given K partitions of dataset E, a strong pattern is a maximal subset of elements of E that are clustered together in all of the K partitions.

Bipartite Graph.

Different clusters with instances are considered.

Experimental Study

We have tested the implementation on a Intel Core 2 Duo 1.86GHz machine. The memory capacity is 1 GB 533 MHz DDR2RAM. The operating system used is Windows 2003. The other programming language used is PHP and database used is MySQL. The online software used is CLUTO.

A graph of 20 points is considered for this experiment. Connected graphs are generated based on density D defined as,

$$D=2.m / (n(n-1)) \qquad (6)$$

Where m is the number of edges and n is the number of vertices in the graph. Graph with various densities i.e 20%, 40%, 70%, 90% are generated. The final partition obtained by the consensus algorithm is evaluated for the accuracy [13]. The results obtained are shown for 20% Dense Graph, Clustering Coefficient method.

Table-1 - 20% Dense Graph, Clustering Coefficient Method

rbr	direct	agglo	graph
0.64	0.49	0.64	0.71
0.47	0.31	0.47	0.55

Conclusion

We find that from experimental study that graph k-way partitions method gives more accuracy. The change in density does not affect this conclusion. It is found from this experiment, when the graph is 20% dense, using edge-betweenness similarity matrix, maximum accuracy is obtained for graph base algorithm. In future, we would like to perform experiments on real life data sets. We also intend to study the different consensus clustering algorithms. We plan to study different quality measures for consensus clustering methods.

Acknowledgment

The second author would like to thank the University Grants Commission for the financial assistance provided to her under Minor Research Project entitled “Design and Experimental study of Consensus Clustering”.

References

[1] Jain A.K., Murty M.N., Flynn P.J. (1999) *ACM Computing Surveys*, 31, 264-323.
[2] Gionis.A.., Mannila H. and Tsaparas P. (2005) *21st International conference on Data Engineering (ICDE'05)*, 341-352.
[3] Strehl A and Gosh J. (2002) *In proceedings of AAAI*, 93-98.
[4] Topchy A., Law M., Jain A.K and Punch W. (2004) *SIAM conference on Data Mining*, 379-390.
[5] Topchy A., Law M., .Jain A.K and Fred A. (2004) *IEEE International Conference on Data Mining, ICDM*, 225-232.
[6] Sitaram A., Srinivasan P., Duygu U. (2006) *ACM KDD, Philadelphia, USA*.
[7] Ulrik B., Marco G. and Dorothea W. (2007) *ACM journal of Experimental Algorithmics*, Vol 12, (article no. 1.1).
[8] Satu E. S. (2007) *Computer science. Review* 27-64.
[9] Xiaoli Z.F. and Carla E.B. (2004) *21st Intel conference on machine learning Banff, Canada,* 36.
[10] Joaquim F., Pinto Da Costa, Rao.P.R. (2004) *REVSTAT Statistical journal*, *vol. 2*, 127, 143.