© 2016 IJEDR | Volume 4, Issue 2 | ISSN: 2321-9939

Advancement in Clustering with the Concept of Correlation Clustering:A Survey

[1]Aaditya Jain,[2]Pooja Mehar,[3]Dr. Bala Buksh
[1]M.Tech Scholar,[2]B.Tech Student,[3]Professor
[1]Department of Computer Science & Engineering,
[1]R. N. Modi Engineering College, Rajasthan Technical University, Kota, Rajasthan, India

Abstract–**Today many scientific, medical and business applications depend heavily on information derived through analysis of large data. Clustering plays an important role in the analysis of data points. A pair wise relationship between any two data points were found to have existed in what is called categorical data. But issue remains as it is so include multiple relations data points instead of pair wise relationships. Both traditional and newer methods of clustering are used widely to recognize structure of data. Correlation Clustering is a relatively newer method of viewing data and the relationships existing among data points. This paper provides a brief survey about the recent developments in the direction of Correlation Clustering.**

IndexTerms - Clustering, Categorical relationships, Correlation Clustering, Edge labeled graphs, Chromatic Correlation Clustering, Spectral Clustering.

I. INTRODUCTION

Clustering is an unsupervised machine learning approach aiming at categorizing similar data points among a set of data and grouping them together in a bundle, specifically called a cluster. It is of wide importance in areas of pattern analysis, statistical data analysis, image analysis, information retrieval, bioinformatics etc. The data points can be numeric, categorical or both. Numeric data is handled using distance metric as a similarity measure. The relationships between numeric data are observed to be either binary or fuzzy. Binary relationship checks whether two data points as similar or dissimilar as a whole. Fuzzy relationships on the other hand points out the percentage of similarity or dissimilarity between data points. Actual representation of objects is however the key requirement in determining both the binary or fuzzy relationships. For handling categorical data, the focus then shifted from the object representation to the relationship one object holds for the other. The term mixed data contains both of the things i.e. numeric and categorical data.

Categorical data clustering algorithms are limited to categorizing data on the basis of their relationship with one category among a given list of categories. A data object can be correlated to a variety of categories, like say; a person on a social media site is a brother to his sisters, a boss of some employees, a father of his children, all at once. His relationships to each one make him belong to a set of clusters, rather than one, if grouping of correlated objects is done. Such kind of relationships can be best described through graphs where the relationships between data objects are portrayed through specific labels, the edges show the interconnection between the points according to the labels and the vertices denote the data objects to be clustered. Clustering such correlated data objects is termed as Correlation Clustering [1]. An interesting point to note is that Correlated data can be categorical but certainly not the opposite of it is true. The reason is because categorical data limits the relationships to only one to many, or many to one whereas the correlation we are talking about includes many to many relationships.

II. BACKGROUND CONCEPTS

The traditional clustering algorithms relied on a real valued proximity function, f(.,.) to calculate distance or similarity between data points. This function f was either provided as input to the algorithm or computed through object representation. It also could be derived from some past training. However, it failed to recognize how two data points interact or communicate with each other and was applied multiple times in the algorithm to handle agreements. This led to recognizing pair-wise relationships among objects which was best identified by edge labeled graphs. In this direction Correlation Clustering [1], Chromatic Correlation Clustering [2] and Spectral Clustering [3], are some newly proposed research works for handling data with categorical relationships.

Correlation Clustering

Bansal et al in 2004 [1] introduced Correlation Clustering for recognizing pair wise relationships between data points through edge labeled graphs and cluster them accordingly. The edges of the graph are labeled as either positive or negative. The desired clustering is based on the notion of minimizing disagreements and maximizing agreements. Agreement here refers to the requirement that the positively labeled edges should be within clusters and the negatively labeled edges should be between clusters. Similarly, disagreement implies negatively labeled edges within clusters and the edges with positive labels between clusters. Therefore, the clustering process should continue

with the notion that the sum of number of disagreements should be minimum and the sum of number of agreements should be maximum. The other promising feature of this clustering approach that it does not require the prior knowledge of the number of clusters to be formed, unlike the conventional clustering algorithms. The need of knowing the number of clusters in advance is eliminated because the objective of the proposal of minimizing the sum of labels of the cut edges runs independently of the number of clusters.

Chromatic Correlation Clustering

Bonchi et al [2] further extended the concept of Correlation Clustering to Chromatic Correlation Clustering having categorical pair wise relation between data objects. Instead of clustering the data points according to maximizing agreements and minimizing disagreements criteria, Chromatic Correlation Clustering requires that the cluster formations should involve the vertices having different colored edges and an objective function to cluster the edges with the same color. The other contributions by Bonchi et al. in this direction include

- A randomized algorithm guaranteeing approximation till the maximum degree of the input graph as the Chromatic Correlation Clustering problem otherwise is a NP-Hard problem.
- A variant algorithm to control the number of clusters formed that checks the choosing mechanism of the pivot and the cluster that builds around it.
- Optimizations in the proposed objective function as per the alternating minimization paradigm to further limit the number of clusters.
- Extension of the randomized algorithm for describing the pair wise relations between a set of labels rather than a single label without hampering its approximation till the maximum degree of the graph.

The results of the proposed work when tested for the various real life and synthetic datasets verified the effectiveness of the proposal.

Spectral Clustering

Correlation Clustering and Chromatic Correlation Clustering were limited to work using simple graphs. But, the pair wise relationships can turn complex too. The edges joining the vertices representing the data objects can also possess some relation between them. Simple graphs fall weak at determining such relations. Based on the graphs being directed or undirected, the relationships are further categorized into asymmetric and symmetric. Pair-wise relationships cannot handle the complexities associated with such relations. These problems led to the notion of clustering data through hypergraphs which can connect more than two vertices. Partitioning hypergraphs for clustering is termed as Spectral Clustering [3]. Vertices in a weighted graph can be labeled or unlabelled or a mix of both. In the mixed case, where some vertices are labeled and the others are not, the labels can be assigned seeing either the similarity between vertices to the same class or the most common label in the classified neighbors of that vertex. For the unlabelled weighted graph, the same framework is used by generalizing partitioning methodology for undirected graphs as in [4]. The authors further extended their work in [5] for hypergraph embedding and transductive inference. The results of clustering using this approach with hypergraph compared with those of clustering using simple graphs were found significantly better.

III. Literature Search

A lot of research is headed in this direction for years, some of them discussed below:

Joachims and Hopcroft [6] in 2005 proposed a model to derive error-bounds for the Correlation Clustering problemrecovering which correct partition in the case of planted partition model is achieved. The authors study next the asymptotic behavior of the problem with respect to the graph density and the sparsity of the formed clusters. The significance of opting for Correlation Clustering is also analyzed statistically by the authors. Shortcomings of the problem are addressed by Achtert et al[11] and Ailon et al [8] who then proposed convincing solutions to the problems.

Giotis and Guruswami [7] in 2006 focused on the effect of keeping the number of clusters, k, fixed for the Correlation Clustering problem. The authors achieved a Polynomial Time Approximation Scheme (PTAS) for k>2 for both the maximizing agreements and minimizing disagreements case. Achieving PTAS was a trivial task for the minimizing disagreements problem which otherwise was observed to be an APX-hard problem when the constant k is not specified.

Ailon and Liberty [9] in 2009 focused on the agnostic nature of the Bansal et al's work that assumes the non- existence of any ground clustering with the solution cost computed against the input similarity function. Opposed to the assumption, the authors assumed that an unknown ground truth clustering exists and that the accuracy of the resultant clustering should be measured against the ground clustering. Provable approximation guarantees are provided by the authors.

Gu and Wang [10] provided a study of hierarchical clustering of volumetric data having correlation relations. Not much work in this direction has been proposed earlier. The authors proposed three clustering algorithms which on the basis of quality threshold, k-means and random walks investigate the correlation relations of the data in a climate dataset. Evaluation and qualitative and quantitative comparison of the algorithms concludes the efficacy of the proposal.

Ailon et al [12] in 2011 studied the problem of Bipartite Correlation Clustering problem followed by the only work in this direction by Amit[13]. As opposed to the 11- approximation guarantee, the authors find a 4-approximation guarantee in their first proposed algorithm which is an extension of the Ailon et al's work [8]. The algorithm is derandomized using the arguments in [8]. The second and the main algorithm is the proposed PivotBiCluster algorithm with straight forward implementation, 4-

approximation ratio and convergence in 0(|E|) steps where E corresponds to the number of edges in the bipartite graph. The disadvantages include the randomness in the algorithm and approximation achieved inexpecatation.

Bonchi et al [14] proposed the overlapping correlation clustering problem by relaxing the problem of Correlation Clustering a bit and allowing data points to be a part of more than one cluster. The weights on labels of an edge-labeled graph are a number in [0, 1]. The data points can be related to each other through a set of labels. The distance between these set of labels can be measured though Jaccard Coefficient and set intersection, and the non-trivial optimization problems of both the measures are solved through non-negative least squares and a greedy strategy respectively. A distributed version of the proposed algorithm has also been designed by the authors. The work can even be used to cluster objects with no available feature vectors.

Anava et al [18] in 2015 provided provable improved theoretical and practical guarantees for the Chromatic Correlation Clustering problem. A constant approximation framework is proposed which gives an approximation ratio of 4, is fast and easy in implementation but not practical. Authors also propose a fast heuristic algorithm with ground truth clustering obscured due to noise and applicable to real life scenarios.

IV. APPLICABILITY OF CORRELATION CLUSTERING

Applicability of Correlation Clustering in various applications has been found promising. Few of the related research works have been discussed below.

Parallel Correlation Clustering

Pan et al [15] in 2015 addressed the problem of Kwik cluster, a popular serial clustering algorithm, of involving a large number of clustering rounds and proposed two parallel correlation clustering algorithms, C4 and ClusterWild as solutions to the problem, the first being a parallel version of Kwik and achieving a 3-approximation ratio like Kwik but through concurrency control and the other being an easy to implement, coordination free and scalable algorithm and a serial variant to Kwik with not much loss to the 3- approximation ratio though abandoning consistency for scaling purpose. C4 and ClusterWild converge in polylogorithmic number of rounds. Both the algorithms outperform the previous algorithms in terms of running time and accuracy and therefore are effective algorithms for clustering.

Correlation Clustering in Data Stream

Ahn et al.[16] in 2015 extended the concept of correlation Clustering to be used in a data stream which not only consists of a sequence of edges with their weights but also updates like insertions and deletions of edges. Instead of putting maximum number of positively labeled edges in a cluster and negatively labeled edges between the clusters, it aimed to form separate clusters of positive edges and negative edges. A space approximation algorithm was also proposed yielding a polynomial time of $O(n.polylog\, n)$. The other contributions of their proposed work included developing linear sketch based data structures to measure the quality of a given node partition followed by combining these data structures to convex programming and sampling techniques for the approximation problem to be solved. Authors further extended their work to designing efficient algorithms for convex programming and to reduce the adaptivity of the sampling.

Correlation Clustering with Noisy Partial Information

Correlation Clustering by Bansal et al [1] was encountered having issues in its average-case models, to which Makarychev et al. in [17] proposed a semi-random model of Correlation Clustering. The average case models were found realistically impossible. Also, each pair of vertices had the same amount of similarity or dissimilarity which made clustering difficult. Two approximation algorithms were also proposed by authors in [18] in their semi-random model. The first algorithm had a Polynomial-Time Approximation Scheme (PTAS) for the instances and the second algorithm was a recovery algorithm for the planted partition giving a small classification error η.

Correlation Clustering for Hyperspectral Imaginary

Correlation Clustering can also be used in Hyperspectral imagery because of its ability to perform different feature selection on different clusters along with clustering data objects. ORCLUS, a correlation clustering algorithm was tested for the same by authors in [19]. They basically enhanced the Correlation Clustering problem in the following ways. Traditionally, Principle Component Analysis (PCA) was used for optimization of ORCLUS for feature selection but the authors used Segmented Principle Component Analysis (SPCA) instead. Another modification proposed in the paper was that the eigen vectors corresponding to smallest eigen values as used by PCA conventionally was changed to maximum values. After the required enhancements, the resultant ORCLUS algorithm was tested on three hyperspectral images.

V. CONCLUSION

The objective of this paper is to shift the focus of the clustering of data points from the traditional binary and fuzzy relationships to newly identified categorical relationships. For this, the similarity measures have been replaced with edge labeled graphs as in Correlation Clustering. Correlation Clustering deduced a new methodology of labeling the data points as positive or negative edges in an edge labeled graph with the notion of maximizing agreements and minimizing disagreements. The extended Correlation Clustering, called the Chromatic Correlation Clustering further introduced colors in the labeled edges with a cluster containing similarly colored edges and separating the differently colored ones. Spectral Clustering emphasized on the limitation of detecting pair-wise relationships between data objects through undirected or directed graphs and proposed hypergraphs in this context. These research works are discussed in detail along with their usability in different applications.

REFERENCES

[1] Nikhil Bansal, Avrim Blum, and Shuchi Chawla, “Correlation clustering”, In 2013 IEEE 54th Annual Symposium on Foundations of Computer Science, pages 238–238, IEEE Computer Society, 2004.

[2] F. Bonchi, A. Gionis, F. Gullo and A. Ukkonen, “Chromatic Correlation Clustering” in KDD '12, Proceedings of the 18th ACM SIGKDD International Conference On Knowledge Discovery And Data Mining, pp. 1321-1329, 2012.

[3] D. Zhou, J. Huang and B. Schölkopf, “Beyond Pairwise Classification and Clustering Using Hypergraphs”, Max Planck Institute Technical Report 143, Max Planck Institute for Biological Cybernetics, T¨¹bingen, Germany, 2005.

[4] J. Shi and J. Malik, “Normalized cuts and image segmentation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8):888-905, 2000.

[5] D. Zhou, J. Huang and B. Schölkopf, “Learning with Hypergraphs: Clustering, Classification, and Embedding”, Advances in Neural Information Processing Systems (NIPS), 19, pp. 1601-1608. (Eds.) B. Schölkopf, J.C. Platt and T. Hofmann, MIT Press, Cambridge, MA, 2007.

[6] Thorsten Joachims and John Hopcroft, “Error Bounds for Correlation Clustering”, Proceedings of the 22nd International Conference on Machine Learning, 2005.

[7] I. Giotis and V. Guruswami, “Correlation Clustering with a Fixed Number of Clusters”, Theory Of Computing, Vol. 2, pp. 249–266, 2006.

[8] N. Ailon, M. Charikar and A. Newman, “Aggregating inconsistent information: Ranking and clustering”, Journal of the ACM, Vol. 55, Issue 5, No. 23, 2008.

[9] N. Ailon and E. Liberty, “Correlation Clustering Revisited: The "True" Cost of Error Minimization Problems”, Proceedings of the 36th International Colloquium on Automata, Languages and Programming: Part I (ICALP '09), pp. 24-36, 2009.

[10] Yi Gu and Chaoli Wang, “A Study of Hierarchical Correlation Clustering for Scientific Volume Data”, Advances in Visual Computing, Volume 6455 of the series Lecture Notes in Computer Science, pp 437-446, 2010.

[11] E. Achtert, C. Bohm, H.P. Kriegel, P. Kroger and A. Zimek, “Robust, Complete, and Efficient Correlation Clustering”, Proceedings of SIAM International Conference on Data Mining (SDM), 2007.

[12] N. Ailon, N. Avigdor-Elgrabli, E. Liberty and A. van Zuylen, “Improved Approximation Algorithms for Bipartite Correlation Clustering”, Algorithms – ESA 2011, Proceedings of the 19th Annual European Symposium, Volume 6942 of the series Lecture Notes in Computer Science, pp 25-36, 2011.

[13] N. Amit, “The Bi-cluster graph editing problem”, Research Report, Tel Aviv University, 2004.

[14] F. Bonchi, A. Gionis and A. Ukkonen, “Overlapping correlation clustering”, Knowledge and Information Systems, Vol. 35, Issue 1, pp 1-32, April 2013.

[15] X. Pan, D. Papiliopoulos, S. Oymak, B. Recht, K. Ramachandran, and M. I. Jordan, “Parallel correlation clustering on big graphs”, Advances in Neural Information Processing Systems (NIPS), Vol. 28, December 2015.

[16] K. Ahn, G.Cormode, S.Guha, A. Mcgregor and A. Wirth, “Correlation Clustering in Data Streams” in Proceedings of the 32nd International Conference on Machine Learning (ICML-15), pp:2237-2246, 2015.

[17] Konstantin Makarychev, Yury Makarychev and Aravindan Vijayaraghavan, “Correlation Clustering with Noisy Partial Information” in JMLR: Workshop and Conference Proceedings vol 40:1–22, 2015.

[18] Y. Anava, N. Avigdor-Elgrabli and I. Gamzu, “Improved Theoretical and Practical Guarantees for Chromatic Correlation Clustering”, Proceedings of the 24th International Conference on World Wide Web (WWW '15), pp. 55-65, 2015.

[19] A. Mehta and O. Dikshit, “SPCA Assisted Correlation Clustering Of Hyperspectral Imagery”, ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume II-8, pp. 111-116, 2014.