International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-2, Issue-2, February 2016
ISSN: 2395-3470
www.ijseas.com

Recent Trends in Correlation Clustering

Maheshwari Neelam Harikishan, Dr. Bala Buksh
R.N. Modi Engineering College, Kota, India

Abstract
The existing numeric data clustering algorithms are found limited to clustering based on the representation of data objects to be clustered. The pair wise relationships between any two data points were found to have existed in what is called Categorical data. This brought to a variety of categorical data clustering techniques. The issues left unresolved then included handling high dimensionality of data and the multiple relations data points can have with each other instead of only pair wise relationships. This brought to the Correlation clustering technique with different correlations in different clusters. This paper discusses the same and surveys the recent developments in the Correlation clustering.
Keywords: *Clustering, Categorical relationships, Correlation Clustering, Edge labeled graphs, Recent Trends.*

1. Introduction

Clustering in machine learning refers to grouping identical objects among a set of objects into a single structure, specifically called a cluster. This highly depends on the nature of data objects to be clustered and the method of calculating similarity between them. The data objects 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. Some typical examples for the understanding of the same can be the blood type of a person: A, AB, B or O; categories of the available rocks: igneous, sedimentary or metamorphic. A data object belongs to either one of the possible categories hence, defining a relation with the other similar objects belonging to that only category. Mixed data contains both of the two worlds, that is numeric and categorical data. A lot of research is headed in this direction for years with some of the most popular data clustering methods discussed in [1,2,3,4,5,6].
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 [7]. 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.
Correlation Clustering has observed a widespread research interest and is being used by many real world applications like pattern recognition, image segmentation, parallel and distributed systems etc. There have been proposed extensions of the same, the very popular of which is the work by Bonchi et al [8] who introduced the chromatic version of Correlation Clustering by the name of Chromatic Correlation Clustering categorizing data points with the help of colors of the joining edges. This paper lists the details of the Correlation clustering methodology with a brief survey about the latest developments in this domain.

2. Correlation Clustering

The concept of Correlation Clustering was introduced by Bansal et al in 2004[7]. They considered an edge labeled graph for the purpose of clustering. 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. Authors took help of a similarity function (f) derived from past data and the resulting clusters relied on this similarity function. Deriving the function (f) from past data reduced the complexity of the function. 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. A graph contains an approximately equal mix of both the negatively and positively labeled edges out of which determination of a perfect clustering is difficult. In such cases, a clear approach of deleting all the negatively labeled edges and connecting the remaining components of the graph returns the desired clusters.

3. Recent Trends

3.1 Chromatic Correlation Clustering

Bonchi et al [8] proposed a simplified approach of handling correlated data by eliminating the requirements of minimizing disagreements and maximum agreements. Rather the relations between objects are considered based on the similarity in the color of edges. The vertices joining similarly colored edges belong to the same cluster. Therefore, the desired clustering just revolves around a set of different colored edges and an objective function used for clustering the similarly colored edges into a single cluster. 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 pairwise 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.

3.2 Correlation Clustering for Hyperspectral Imagery

Mehta and Dikshit [9] use Correlation Clustering for unsupervised classification in hyperspectral imagery because of the ability of Correlation Clustering in performing feature reduction and clustering simultaneously. Plus different set of features can be selected for ORCLUS [10], a correlation clustering algorithm has been modified. Principal Component Analysis (PCA) feature reduction tool has been replaced with Segmented Principal Component Analysis (SPCA). The other modification included using eigenvectors corresponding to maximum eigen values as used in PCA as opposed to the smallest eigen values as done in the original ORCLUS algorithm. Results of the proposal are then tested on three real hyperspectral images.

3.3 Correlation Clustering for Attribute Weighting

Carbonera and Abel in 2014 [11] proposed Correlation Clustering for attribute weighing for the first time. Unlike the traditional attribute weighing categorical clustering algorithms that compute the

weights either by the frequency of the mode category or the average distance a data object has from the cluster mode; this approach considers the correlations of an attribute with other attributes for measuring what contributions that attribute offers. The motivation behind their work lies in the cognitive studies that show that the learning in humans turns spontaneous when the correlations among the attributes of the perceived objects are explored. The other advantages of the proposal over the traditional algorithms of related type are its non parametric nature, that is no dependence on the user defined parameters, and no requirement of previously labeling of data.

3.4 Parallel Correlation Clustering, C4 and ClusterWild!

Pan et al [12] modifies one of the popular Correlation Clustering algorithms, Kwik Cluster [13], a simple peeling scheme offering a 3- approximation ratio. The proposed modifications are done to eliminate the shortcomings of the sequential approach of the algorithm and the large number of peeling rounds involved in the algorithm. Since the mentioned shortcomings disturb the scaling of the clustering process for big graphs; authors propose a Parallel Correlation Clustering algorithm, C4. The effectiveness of the proposal is guaranteed by its 3-approximation ratio, limited overheads and scaling up to billion-edge graphs. The idea behind the proposal is running a number of peeling threads concurrently (therefore, the word parallel) and maintaining consistency in each of the threads such that there are no associated overheads in the algorithm. The consistency among the concurrently running threads is maintained through the Concurrency control paradigm of the Database Management research. This approach was able to eliminate the scaling challenges of the Kwik Cluster algorithm without losing the 3-approximation ratio of the original algorithm. Even a billion-edge graph can be clustered efficiently in a few seconds as per the results shown in the paper.

In their next paper, Pan et al [14] show another parallel correlation clustering algorithm by the name of *ClusterWild!*, a coordination free problem that, for achieving better scaling, abandons consistency. The cost of waiving consistency is a small loss encountered in the accuracy of the algorithm. The *ClusterWild!* Algorithm is a coordination free variant of the KwikCluster algorithm running on a “noisy” graph. The authors also show that the number of rounds in both the C4 and the *ClusterWild!* algorithms are polylogarithmic.

3.5 Correlation clustering in a dynamic data stream

Ahn et al [15] focused on clustering correlated objects in a dynamic data stream model. Unlike the simple data stream model consisting of sequenced edges with their labels(referred to as weights in the paper), the associated data stream updates the edge labels of the related edge labeled graph containing n nodes as it arrives. The updates include insertions and deletions of edges. Three types of weights are considered: unit weights containing a set of only unit positive and unit negative edges, bounded weights which should necessarily be non zero and bounded by some constant and lastly, arbitrary weights consisting of all weights of O(poly n). The objective behind the proposal is to find a node partition efficiently able to partition the negatively labeled edges in different cluster and the positively labeled edges in the same cluster. For ensuring the quality of the associated node partition, authors develop data structures based on linear sketches. To solve the space- approximation problem in O (n.polylog n) space, the developed data structures are then combined with convex programming and sampling techniques.

3.6 Scaling Correlation Clustering through fusion moves

Beier et al [16] worked on scaling the Correlation clustering problem and present novel scaling approaches for the same. Their other contributions are listed below.

- A novel energy based agglomerative clustering algorithm is proposed that gives monotonically increasing energy. With this algorithm, anytime performance of Cut, Clue and Cut [17] is increased.
- Efficient separation procedures are proposed that improve the anytime performance of polyhedral multicut problems [18].
- Introduction of cluster fusion moves as an unsupervised extension to the original fusion

moves [19] for the supervised segmentation thereby giving its polyhedral implementation. In other words, a current and a proposed partitioning method are fused to obtain monotonic improvement in partitioning and maintain a valid partitioning throughout.

- Implementation of the proposal is done through two proposed versatile proposal generators.

The advantages of the proposal can be listed as

- Scalable to large datasets
- Near optimum solution for the correlation clustering problem
- Good anytime performance

3.7 Correlation Clustering with Noisy Partial Information

Makarychev et al [20] aimed to study and propose a semi-random model for the general instances improve the realistic average case Correlation Clustering models by designing two approximation algorithms with better provable guarantees. The former algorithm gives a high probability solution of $(1+\delta)opt - cost + O_\delta(n\log^3 n)$ with $opt - cost$ being the value of the optimal solution for $\delta > 0$. The latter algorithm is effective enough to find an optimal ground truth clustering with a small classification error η.

4. Conclusion

Consideration of real life data for clustering brings with it data of varying types related to each other. Handling such varying kind of data and clustering them effectively is now possible with the focus of clustering such data shifted from the conventional clustering algorithms to considering new structures like edge labeled graphs and new methods for clustering like Correlation Clustering. Correlation Clustering, through the use of edge labeled graphs, effectively clusters correlated data object through the notion of maximizing agreements and minimizing disagreements. Since the proposal of Correlation Clustering by Bansal et al in 2004, this field has become a hot research topic extending its applicability to almost every field. This paper is a brief surveys discussing some recent developments in this domain.

References

[1] E.W.Forgy, "Cluster analysis of multivariate data: efficiency v/s interpretability of classifications", Biometrics, 21, 768–769, 1965.

[2] G Tzortzis, A. Likas, "The MinMax k-Means clustering algorithm", Pattern Recognition, Volume 47, pp 2505–2516, Elsevier, 2014.

[3] S.Guha, R.Rastogi, and K.Shim Rock: "A robust clustering algorithm for categorical attributes". Proceedings of the 15th international conference on data engineering, IEEE Computer Society Sydney, Australia pp. 512–521, March 1999.

[4] V. Ganti, J. Gehrke and R. Ramakrishnan. CACTUS: Clustering categorical data using summaries" In Proceedings of ACM SIGKDD, International Conference on Knowledge Discovery & Data Mining, San Diego, CA, USA, 1999.

[5] Zhexue Huang, "Clustering large data sets with mixed numeric and categorical values", Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining, 1997.

[6] Chung-Chian Hsu, Chin-Long Chen and Yu-Wei Su, "Hierarchical clustering of mixed data based on distance hierarchy", Information Sciences, Volume 177, Issue 20, Pages 4474–4492, 2007.

[7] 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, 2002.

[8] Francesco Bonchi, Aristides Gionis, Francesco Gullo and Antti Ukkonen , "Chromatic Correlation Clustering" in KDD '12 Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, Pages 1321-1329, 2012.

[9] 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.

[10] C. C Aggarwal and P.S. Yu, "Finding generalized projected clusters in high dimensional spaces", Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 70-81, 2000.

[11] J.L. Carbonera and M. Abel, "Categorical data clustering: A correlation-based approach for

unsupervised attribute weighting, Proceedings of ICTAI, 2014.

[12] Xinghao Pan, Dimitris Papiliopoulos, Benjamin Recht, Kannan Ramachandran, and Michael I. Jordan, "Scaling up correlation clustering through parallelism and concurrency control", NIPS Workshop on Discrete and Combinatorial Problems in Machine Learning (DISCML), December 2014.

[13] Nir Ailon, Moses Charikar, and Alantha Newman, "Aggregating inconsistent information: ranking and clustering", Journal of the ACM (JACM), 55(5):23, 2008.

[14] Xinghao Pan, Dimitris Papiliopoulos, Sameet Oymak, Benjamin Recht, Kannan Ramachandran, and Michael I. Jordan, "Parallel correlation clustering on big graphs", Advances in Neural Information Processing Systems (NIPS) 28, December 2015.

[15] Kookjin Ahn, Graham Cormode, Sudipto Guha, Andrew Mcgregor and Anthony Wirth, "Correlation Clustering in Data Streams" in Proceedings of the 32nd International Conference on Machine Learning (ICML-15), pp:2237-2246, 2015.

[16] Thorsten Beier, Fred A. Hamprecht, and Jörg H. Kappes, "Fusion moves for Correlation Clustering", CVPR, page 3507-3516, IEEE, 2015.

[17] T. Beier, T. Kroeger, J. H. Kappes, U. Koethe, and F. A. Hamprecht, "Cut, Glue & Cut: A Fast, Approximate Solver for Multicut Partitioning", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2014, 2014.

[18] J. H. Kappes, M. Speth, G. Reinelt and C. Schnorr, "Higher-order segmentation via multicuts", CoRR, abs/1305.6387, 2013.

[19] V. Lempitsky, C. Rother, S. Roth, and A. Blake. Fusion moves for Markov random field optimization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(8):1392–1405, aug 2010.

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