International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 02 Issue: 09 | Dec-2015 www.irjet.net p-ISSN: 2395-0072

Clustering Data with Categorical Relationships

Jaishri Gothania [1], Dr. Bala Buksh [2]

[1] *Mtech Scholar, Department of Computer Engineering, R. N. Modi Engineering College, Kota, Rajasthan, India*

[2] *Director, Department of Computer Engineering, R. N. Modi Engineering College, Kota, Rajasthan, India*

Abstract - ***The research headed towards clustering data aims at considering numeric data, categorical data or a mixture of both. For such data, the concentration was finding a relationship between the data points to be clustered. The relationships were limited to being either binary or fuzzy. Both involved a numeric value called distance or any other similarity measure between two data points and cluster them together if found similar. With time, a new kind of relationship called categorical relationship was observed between data points, far different from the traditionally seen ones. This paper focuses on handling data points having categorical relationships and the techniques emerged till date in this direction.***

Key Words: *Clustering, Categorical Relationships, Edge- labeled graphs, Spectral Clustering, Correlation Clustering, Chromatic Correlation Clustering*

1. 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. Effective Clustering takes into account two aspects; the nature of the data points o be clustered and the relationships between these data points. The data can be numeric, categorical or mixed. The relationships between the data points are observed to be binary, fuzzy or the newly observed categorical. All the previous research works were limited to picking a numeric value called distance or any other similarity measure between data points to cluster them accordingly. This similarity can be perfectly deduced if we find what relationship two data points are holding for each other. Whereas binary relationship categorized the data points as similar or dissimilar with respect to any similarity measure used and clustering them accordingly, fuzzy relationship pointed out a percentage of similarity or dissimilarity between data points with the less similar ones more probable to lie in the same cluster. Both the binary and the fuzzy relationships involved computation on the actual representations of the objects.

With advancements in technology, focus shifted towards data points having categorical relationships or associated labels between them. For example, the people associated on social networking sites, are clustered on the basis of their relationships with each other, some being acquaintances, friends, close friends, family. Each relationship is provided with some privileges, priorities or permissions. Clustering such data cannot be efficiently done through the traditionally used similarity functions or through a numeric value. Spectral Clustering [1], Correlation Clustering [2] and Chromatic Correlation Clustering [3] are some newly proposed research works for handling data with categorical relationships. This paper surveys the research works and their extensions with the sole purpose of clustering such data objects.

2. CORRELATION CLUSTERING

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.

2.1 Correlation Clustering by Bansal et al (2004)

Bansal et al in 2004 [2] introduced Correlation Clustering for recognizing pair-wise relationships between data points through edge labeled graphs and cluster them accordingly. Unlike the conventional clustering algorithms, it introduced the notion of labeling graphs as positive or negative. The labeling was done to form clusters with maximized agreements and minimized disagreements. Agreement implies that the cluster should have maximum number of positive edges within clusters and negative edges between clusters. Disagreement, on the other hand,

means just the opposite of agreement, i.e. more negative edges within clusters. The similarity function in the problem could be taken by some past data and the resultant clustering should be related to the similarity function as much as possible. The authors also point out a limitation of the conventional clustering algorithms of needing to know the number of clusters a prior. This limitation is overlooked in their paper because the resultant clustering can any number of clusters based on the edge labels. For the objective of maximizing disagreements and minimizing agreements, the authors proposed a constant factor approximation and PTAS respectively.

2.2. Correlation Clustering with Noisy Partial Information

Correlation Clustering by Bansal et al [2] was encountered having issues in its average-case models, to which authors in [4] 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 [4] 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.

2.3 Correlation Clustering In Data Streams

Authors in [5] 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.

3. APPLICATIONS OF CORRELATION CLUSTERING

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

3. 1 Phase Transition

Neda at al [6] observed the Phase Transition in a complete signed graph to be occurring with function *r(q)* at q=1/2; q being the relative size of the maximal cluster. Seeing the complexity associated with clustering random graphs, Monte Carlo simulations were done instead of a mathematically strict analysis. Testing was done both on Erdos-Renyi graphs [7, 8] in [6] and Barabasi-Albert type [9] scale free graphs in [10]. Aszalos [11], inspired by the work of Neda et al, then replaced the simulation tools of Neda et al in [10]. His other contributions included a new proposed storage method for graphs and increment in the number of nodes from 100-500. His method made it able for a desktop computer to generate needing thousands of clustering using this method. All the three mentioned algorithms worked knowing only some of the parameters of a random graph and no prior knowledge about its structure. For experimentation purpose, Azsalos [11] used Contraction, a simple greedy algorithm, involving partitioning of singletons, followed by selecting pairs of clusters and then joining them till no pairs can be further joined. The partitioning done is further restricted to following a rule, that is, a function $f_G^s(p)$ should be maximal; $f_G^s(p)$ denoting the cost value of partition for a signed graph adhering to the minimizing disagreement criterion. Several conjectures depicting the behavior of the deduced curves were formulated.

3.2 Two-edge connected Augmentation for Planar Graphs

For a given graph S containing weights of edges and a subset of the edges R, if the weight of R is minimum and each edge of R has a two-edge connected endpoint in $R \cup S$, then the given graph is said to have two-edge connected augmentation. The Correlation clustering problem was found to be a NP-hard [12] and the only possible improvement in this direction was a constant-factor approximation scheme by [13]. Authors in [14] propose to reduce correlation clustering to two-edge-connected augmentation for planar graphs and addressed the problem of Correlation Clustering being NP-Hard by giving a polynomial-time approximation scheme.

3.3 Hyperspectral Imagery

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 [15]. 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.

3.4 Image Partitioning Using Multicuts

Authors in [16] addressed the shortcoming of Maximum-A-Priori (MAP) point of having less scope of probabilistic inference and took correlation clustering as a multicut problem. One possible way of overcoming the addressed problem was by estimating 'error bars' to access sensitivities and uncertainties for future data analysis. Presenting a probabilistic approach to Correlation Clustering using Perturbed MAP, estimates for image processing, open contour parts due to imperfect local detection can be closed thereby reducing the local artefacts by topological priors. A significantly reduced computation time to seconds from minutes was observed using their proposed method.

4. CHROMATIC CORRELATION CLUSTERING

Bonchi et al [3] 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 a objective function to cluster the edges with the same color. The contributions include

- The Chromatic Clustering problem, like Correlation Clustering, image partitioning and other traditional clustering is NP-hard; therefore, random algorithm was designed which guaranteed approximation till the maximum degree of the input graph. The steps of the proposed algorithm were as follows
    - Taking a random edge of a graph as a pivot
    - Start building a cluster around it
    - Remove that cluster from the graph.
    - Repeating the same steps iteratively till all the edges are clustered
- Following were the advantages of the proposed algorithm.
    - Faster execution
    - Easy implementation
    - Parameter free clustering
- The only drawback encountered was that a large number of clusters can be formed, a solution to which was a proposed variant algorithm aiming at checking the choosing mechanism of the pivot and the building of cluster around it.
- The clustering problem can also limit the number of clusters as required by further optimizing the objective function according to the alternating-minimization paradigm.
- For the pairwise relations described through labels, authors proposed a yet another extension to the proposed randomized approximation algorithm by introducing some modifications to it. The modifications, in no way, disturbed its guarantee of approximation till the maximum degree of the graph
- Testing of the proposal on both synthetic and real life datasets in terms of the constructed ground-truth clustering and objective function showed outperformed results.

Authors in [17] generalized the correlation clustering problem. Like Correlation Clustering, the pairwise relations were categorical instead of the conventional binary relations. Linear approximation was achieved almost in all cases for Correlation Clustering. The authors improved the current knowledge state and theoretical understanding of the clustering problem through its designed constant approximation problem. Another contribution of their research work included the improvement of the approximation ratio of 4 through a deterministic linear programming-based algorithm. The proposed clustering algorithm was fast for the ground-truth clustering that is mostly hidden by the noisy observations and testing on synthetic and real life datasets with outperformed results for practical analysis.

5. 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. Even if they are squeezed somehow to being pair-wise, loss of information can possibly occur. 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. This concept was introduced by authors in [1]. They presented a framework for the same. 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 [18]. Concept of a natural random walk over hypergraphs is also introduced according to which the cut criterion and the regularization framework were interpreted. The cut criterion used is the Normalized- Cut approach of [18]. Analogous to the normalized cut criterion for simple graphs, a real-valued relaxed criterion for the hypergraphs was suggested called the Hypergraph Laplacian. This work was limited to classification and clustering using hypergraphs. The authors further extended their work in [19] for hypergraph embedding and transductive inference. The results of clustering using hypergraphs when compared with those of clustering using simple graphs were found significantly better.

The relations are also not limited to being pairwise or dyadic and can be triadic, tetradic or higher affinity relations. Clustering data objects having such relations was the objective of the research by authors in [20]. A two step algorithm was proposed for the same. The algorithm, with the use of any similarity measure, can be used effectively even for other types of clustering. The steps of the algorithm can be summarized as follows:

- First step uses a weighted graph to approximate the resulting hypergraph via a novel scheme.
- Second step involves portioning of the vertices of the graph through a spectral partitioning algorithm.

The proposed algorithm can be efficiently run on any order hyperedges including order two thereby simultaneously incorporating information related to each order. Performance analysis of the proposed algorithm proves its superiority when compared to any existing partitioning algorithm.

6. 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] 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¨ubingen, Germany, 2005.

[2] N. Bansal, A. Blum, and S. Chawla, “Correlation Clustering”, Machine Learning , 56, pp.89-113,2004.

[3] 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.

[4] K. Makarychev, Y. Makarychev and A. Vijayaraghavan, “Correlation Clustering with Noisy Partial Information” in JMLR: Workshop and Conference Proceedings , vol 40, pp.1–22, 2015.

[5] 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.

[6] Z. Neda, R. Florian, M. Ravasz, A. Libal and G. Gyorgyi, “Phase transition in an optimal clusterization model”, Physica A:Statistical Mechanics and its Applications, 362 (2) , pp.357-368, 2006.

[7] P. Erdos and A. Renyi, “On random graphs”, Publ. Math. Debre-cen, 6, pp. 290-297, 1959.

[8] P. Erdos and A. Renyi, “On the evolution of random graphs”, Magyar Tud. Akad. Mat. Kut. Int. K□özl., 5 ,17-61, 1960.

[9] A. L. Barabasi, and R. Albert, “Emergence of scaling in random networks”, Science, 286 (5439) , 509-512, 1999.

[10] Z. Neda, R. Sumi, M. Ercsey-Ravasz, M. Varga, B.Molnar and Gy. Cseh, “Correlation clustering on networks”, Journal of Physics A: Mathematical and Theoretical, 42 (34): 345003, 2009.

[11] L. Aszalos, J. Kormos and D. Nagy, “Conjectures On Phase Transition At Correlation Clustering Of Random Graphs” Annales Univ. Sci. Budapest., Sect. Comp. 42 (2014) 37-54.

[12] Y. Bachrach, P. Kohli, V. Kolmogorov, and M. Zadimoghaddam, “Optimal coalition structure generation in cooperative graph games”, In Conference on Artificial Intelligence, 2013.

[13] E. D. Demaine, D. Emanuel, A. Fiat, and N. Immorlica, “Correlation clustering in general weighted graphs”, Theoretical Computer Science, 361(2), pp.172–187, 2006.

[14] P. N. Klein, C. Mathieu and H. Zhou, “Correlation Clustering and Two-edge-connected Augmentation for Planar Graphs” in International Proceedings in Informatics, LIPIcs 02/2015; 30, pp. 554-567, 2015. DOI: 10.4230/LIPIcs.STACS.2015.554

[15] A. Mehta, O. Dikshit, "SPCA Assisted Correlation Clustering Of Hyperspectral Imagery " in ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume II-8, 2014

[16] J. Hendrik Kappes, P. Swoboda , B, Savchynskyy, T. Hazan and C. Schnorr, "Probabilistic Correlation Clustering and Image Partitioning Using Perturbed Multicuts" in 5th International Conference, SSVM 2015, Lège-Cap Ferret, France, May 31 - June 4, 2015, Proceedings, Scale Space and Variational Methods in Computer Vision, Volume 9087 of the series Lecture Notes in Computer Science pp 231-242, 2015.

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

[18] J. Shi and J. Malik, "Normalized cuts and image segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8):888{905, 2000.

[19] 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.

[20] S. Agarwal et al, "Beyond Pairwise Clustering" Proceedings of the IEEE CVPR 2005, San Diego, CA, 2005.