Data Aggregation / Clustering Algorithm Evolutions: A review

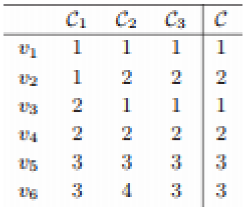
P. Pavankumar [1], Dr. G.Narender [2]
[1] Research Scholar
Associate Professor, KMIT, Hyderabad

Abstract*- Appropriated information conglomeration is an essential errand, permitting the decentralized assurance of significant worldwide properties, which would then be able to be utilized to coordinate the execution of different applications. The subsequent esteems result from the appropriated calculation of capacities like check, aggregate and normal. Some application cases can found to decide the system estimate, add up to capacity limit, normal load, larger parts and numerous others. In the most recent decade, a wide range of methodologies have been proposed, with various exchange offs as far as exactness, unwavering quality, message and time multifaceted nature. Because of the extensive sum and assortment of collection calculations, it can be troublesome and tedious to figure out which strategies will be more suitable this work surveys the cutting edge on circulated information accumulation calculations, giving three primary commitments. In the first place, it formally characterizes the idea of collection, describing the distinctive kinds of accumulation capacities. In this paper, we give a formal explanation of the grouping conglomeration issue, and we propose various calculations. Our calculations make utilization of the association between bunching collection and the issue of relationship grouping. In spite of the fact that the issues we consider are NP-hard, for a few of our techniques, we give hypothetical certifications on the nature of the arrangements. Our work gives the best deterministic estimation calculation for the variety of the relationship grouping issue we consider. We additionally indicate how testing can be utilized to scale the calculations for substantial datasets. We give a broad exact assessment exhibiting the helpfulness of the issue and of the arrangements.*

Keywords*-* Data clustering, clustering categorical data, clustering aggregation, correlation clustering

I. INTRODUCTION

Clustering is an important step in the process of data analysis and has applications to numerous fields. Informally, clustering is defined as the problem of partitioning data objects into groups (clusters) such that objects in the same group are similar, while objects in different groups are dissimilar. This definition assumes that there is some well-defined quality measure that captures intracluster similarity and/or intercluster dissimilarity. Clustering then becomes the problem of grouping together data objects so that the quality measure is optimized. There is an extensive body of literature on clustering methods, see, for instance, Jain and Dubes [1987]; Hand et al. [2001]; Han and Kamber [2001]. In this article, we consider an approach to clustering that is based on the concept of aggregation. We assume that given a set of data objects, we can obtain some information on how these objects should be clustered. This information comes in the form of m clusterings C1, ... , Cm. The objective is to produce a single clustering C that agrees as much as possible with the m input clusterings. We define a disagreement between two clusterings C and Cas a pair of objects (v, u) such that C places them in the same cluster, while Cplaces them in different clusters or vice versa. If d(C, C) denotes the number of disagreements between C and C , then the task is to find a clustering C that minimizes m i=1 d(Ci, C). As an example, consider the dataset V = {v1, v2, v3, v4, v5, v6} that consists of six objects, and let C1 = {{v1, v2}, {v3, v4}, {v5, v6}}, C2 = {{v1, v3}, {v2, v4}, {v5}, {v6}}, and C3 = {{v1, v3}, {v2, v4}, {v5, v6}} be three clusterings of V . Figure 1 shows the three clusterings where each column corresponds to a clustering, and a value i denotes that the tuple in that row belongs in the i-th cluster of the clustering in that column. The right-most column is the clustering C = {{v1, v3}, {v2, v4}, {v5, v6}}that minimizes the total number of disagreements with the clusterings C1, C2, C3. In this example, the total number of disagreements is 5 one with the clustering C2 for the pair (v5, v6), and four with the clustering C1 for the pairs (v1, v2), (v1, v3), (v2, v4), (v3, v4). It is not hard to see that this is the minimum number of disagreements possible for any partition of the dataset V . We define clustering aggregation as the optimization problem where, given a set of m clusterings, we want to find the clustering that minimizes the total number of disagreements with the m clusterings. Clustering aggregation provides a general framework for dealing with a variety of problems related to clustering: (i) it gives a natural clustering algorithm for categorical data; (ii) it handles heterogeneous data where tuples are defined over incomparable attributes; (iii) it determines the appropriate number of clusters and it detects outliers; (iv) it provides a method for improving the clustering robustness by combining the results of many clustering algorithms; and (v) it allows for clustering of data that is vertically partitioned in order to preserve privacy. We elaborate on the properties and the applications of clustering aggregation.



	C_1	C_2	C_3	C
v_1	1	1	1	1
v_2	1	2	2	2
v_3	2	1	1	1
v_4	2	2	2	2
v_5	3	3	3	3
v_6	3	4	3	3

Fig. 1. An example of clustering aggregation. C_1, C_2, and C_3 are th are the objects to be clustered. A value k in the entry (v_i, C_j) means of the clustering C_j. Column C is the clustering that minimizes th C_1, C_2, and C_3.

Clustering aggregation has been previously considered under a variety of names (consensus clustering, clustering ensemble, clustering combination) in a variety of different areas such as machine learning [Strehl and Ghosh 2002; Fern and Brodley 2003], pattern recognition [Fred and Jain 2002], bioinformatics [Filkov and Skiena 2004], and data mining [Topchy et al. 2004; Boulis and Ostendorf 2004]. The problem of correlation clustering is interesting in its own right, and it has recently attracted a lot of attention in the theoretical computer science community [Bansal et al. 2004; Charikar et al. 2003; Demaine et al. 2006; Swamy 2004]. We review some of the related literature on both clustering aggregation and correlation clustering in Section 3. Our contributions can be summarized as follows.

—We formally define the problem of clustering aggregation, and we demonstrate the connection between clustering aggregation and correlation clustering.

—We present a number of algorithms for clustering aggregation and correlation clustering. We also propose a sampling mechanism that allows our algorithms to handle large datasets. The problems we consider are NP-hard, yet we are

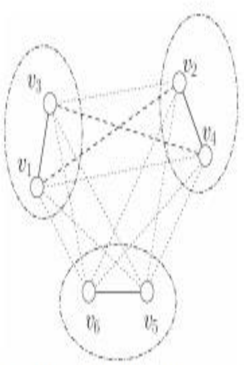


Fig. 2. A correlation clustering instance for the dataset in Figure 1. Solid edges indicate distances of 1/3, dashed edges indicate distances of 2/3, and dotted edges indicate distances of 1. The circles depict the clusters of clustering C that minimizes the number of disagreements.

II. APPLICATIONS OF CLUSTERING AGGREGATION

Clustering aggregation can be applied in various settings. We will now present some of the main applications and features of our framework. Clustering categorical data. An important application of clustering aggregation is that it provides a very natural method for clustering categorical data. Consider a dataset with tuples t1, ... , tn over a set of categorical attributes A1, ... , Am. The idea is to view each attribute Aj as a way of producing a simple clustering of the data, that is, if Aj contains kj distinct values, then Aj partitions the data in kj clusters, one cluster for each value. Then, clustering aggregation considers all those m clusterings produced by the m attributes and tries to find a clustering that agrees as much as possible with all of them. For example, consider a Movie database. Each tuple in the database corresponds to a movie that is defined over a set of attributes such as Director Actor, Actress, Genre, Year, etc, some of which take categorical values. Note that each of the categorical attributes naturally defines a clustering. For example, the Movie.Genre attribute groups the movies according to their genre, while the Movie.Director according to who directed the movie. The objective is to combine all these clusterings into a single clustering.

Clustering heterogeneous data: The clustering aggregation method can be particularly effective in cases where the data are defined over heterogeneous attributes that contain incomparable values. Consider for example the case that there are many numerical attributes whose units are incomparable (say, Movie.Budget and Movie.Year) and so it does not make sense to compare numerical vectors directly using an Lp-type distance measure. A similar situation arises in the case where the data contains a mix of categorical and numerical values. In such cases, the data can be partitioned vertically into sets of homogeneous attributes, obtain a clustering for each of these sets by applying the appropriate clustering algorithm, and then aggregate the individual clusterings into a single clustering.

Identifying the correct number of clusters. One of the most important features of the formulation of clustering aggregation is that there is no need to specify the number of clusters in the result. The automatic identification of the appropriate number of clusters is a deep research problem that has attracted significant attention (see, e.g., Schwarz [1978]; Hamerly and Elkan [2003]; Smyth [2000]). For most clustering approaches, the quality of the solution (likelihood, sum of distances to cluster centers, etc.) improves as the number of clusters is increased. Thus, the trivial solution of all singleton clusters is the optimal. There are two ways of handling the problem. The first is to have a hard constraint on the number of clusters or on their quality. For example, in agglomerative algorithms, one can either fix in advance the number of clusters in the final clustering or impose a bound on the distance beyond which no pair of clusters will be merged. The second approach uses model selection methods.

Detecting outliers: The ability to detect outliers is closely related to the ability to identify the correct number of clusters. If a node is not close to any other nodes, then from the point of view of the objective function, it would be bene- ficial to assign that node in a singleton cluster. In the case of categorical data clustering, the scenarios for detecting outliers are very intuitive. If a tuple contains many uncommon values, it does not participate in clusters with other tuples, and it is likely that it will be identified as an outlier. Another scenario where it pays off to consider a tuple as an outlier is when the tuple contains common values (and therefore it participates in big clusters in the individual input clusterings), but there is no consensus on a common cluster (e.g., a horror movie featuring actress Julia.Roberts and directed by the independent director Lars.vonTrier).

Improving clustering robustness: Different clustering algorithms have different qualities and different shortcomings. Some algorithms might perform well in specific datasets but not in others, or they might be very sensitive to parameter settings. For example, the single-linkage algorithm is good at identifying elongated regions, but it is sensitive to clusters connected with narrow strips of points. The k-means algorithm is a widely-used technique, but it favors spherical clusters, it is sensitive to clusters of uneven size, and it can get stuck in local optima.

Privacy-preserving clustering: Consider a situation where a database table is vertically split and different attributes are maintained in different sites. Such a situation might arise in cases where different companies or governmental administrations maintain various sets of data about a common population of individuals. For such cases, our method offers a natural model for clustering the data maintained in all sites as a whole in a privacy-preserving manner, that is, without the need for the different sites to reveal their data to each other and without the need to rely on a trusted authority. Each site clusters its own data independently, and then all resulting clusterings are aggregated. The only information revealed is which tuples are clustered together; no information is revealed about data values of any individual tuples.

The AGGLOMERATIVE Algorithm: The AGGLOMERATIVE algorithm is a standard bottom-up procedure for the correlation clustering problem. It starts by placing every node into a singleton cluster. It then proceeds by considering the pair of clusters with the smallest average distance. The average distance between two clusters is defined as the average weight of the edges between the two clusters. If the average distance of the closest pair of clusters is less than 1/2, then the two clusters are merged into a single cluster. If there are no two clusters with average distance smaller than 1/2, then no merging of current clusters can lead to a solution with improved cost d(C). Thus, the algorithm stops, and it outputs the clusters it has created so far. The AGGLOMERATIVE algorithm has the desirable feature that it creates clusters where the average distance of any pair of nodes is at most 1/2. The intuition is that the opinion of the majority is respected on average. Using this property, we are able to prove that when m = 3, the AGGLOMERATIVE algorithm produces a solution with cost at most 2 times that of the optimal solution. The proof appears in Section 6. The complexity of the algorithm is O(mn2) for creating the matrix plus O(n2 log n) for running the algorithm.

The BESTCLUSTERING Algorithm: This is the simple algorithm that was mentioned in the previous section. Given m clusterings C1,..., Cm, BESTCLUSTERING finds the input clustering Ci that minimizes the total number of disagreements D(Ci). Using the data structures described in Barthelemy and Leclerc [1995] or techniques similar to those described in Mielikainen et al. [2006] the best ¨ clustering can be found in time O(m2n). As discussed, this algorithm yields a solution with an approximation ratio at most 2(1−1/m). In Section 6, we show that this bound is tight, that is, there exists an instance of the clustering aggregation problem where the algorithm BESTCLUSTERING produces a solution of cost exactly 2(1 − 1/m) times the cost of the optimal solution.

The algorithm is specific to clustering aggregation— it cannot be used for correlation clustering. In fact, it is not always possible to construct a clustering aggregation instance that gives rise to the given correlation clustering instance. Any metric X uv that arises out of clustering aggregation is a convex combination of cut metrics and is, therefore, an L1 metric (see Deza and Laurent [1997]). Thus, a metric X uv that is not an L1 metric cannot be represented by a clustering aggregation instance.

III. CONCLUSION

In this paper we considered the problem of clustering aggregation. Simply stated, the idea is to cluster a set of objects by trying to find a clustering that agrees as much as possible with a number of preexisting clustering. We motivated the problem by describing in detail various applications of clustering aggregation including clustering categorical data, dealing with heterogeneous data, improving clustering robustness, and detecting outliers. We formally de- fined the problem, and we showed its connection with the problem of correlation clustering. We proposed various algorithms for both the clustering aggregation and the correlation clustering problem including a sampling algorithm that allows us to handle large datasets with no significant loss in the quality of the solutions.

REFERENCES

[1] AILON, N., CHARIKAR, M., AND NEWMAN, A. 2005. Aggregating inconsistent nformation: Ranking and clustering. In Proceedings of the ACM Symposium on Theory of Computing (STOC). 684– 693.

[2] ANDRITSOS, P., TSAPARAS, P., MILLER, R. J., AND SEVCIK, K. C. 2004. LIMBO: Scalable clustering of categorical data. In Proceedings of the International Conference on Extending Database Technology (EDBT). 123–146.

[3] BANSAL, N., BLUM, A., AND CHAWLA, S. 2004. Correlation clustering. Machine Learn. 56, 1–3, 89– 113. BARTHELEMY, J.-P. AND LECLERC, B. 1995. The median procedure for partitions. DIMACS Series in Discrete Mathematics, 3–34.

[4] BLAKE, C. L. AND MERZ, C. J. 1998. UCI repository of machine learning databases. BOULIS, C. AND OSTENDORF, M. 2004. Combining multiple clustering systems. In Proceedings of the European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD). 63–74.

[5] CHARIKAR, M., GURUSWAMI, V., AND WIRTH, A. 2003. Clustering with qualitative information. In Proceedings of the IEEE Symposium on Foundations of Computer Science (FOCS). 524–533. CRISTOFOR, D. AND SIMOVICI, D. A. 2001. An information-theoretical approach to genetic algorithms for clustering. Tech. rep. TR-01-02, UMass, Boston,

[6] MA. DEMAINE, E. D., EMANUEL, D., FIAT, A., AND IMMORLICA, N. 2006. Correlation clustering in general weighted graphs. Theoret. Computer. Science 361, 2–3, 172–187.

[7] DEZA, M. AND LAURENT, M. 1997. Geometry of Cuts and Metrics. Springer-Verlag. DWORK, C., KUMAR, R., NAOR, M., AND SIVAKUMAR, D. 2001. Rank aggregation methods for the Web. In Proceedings of the International World Wide Web Conference. 613– 622.

[8] FAGIN, R., KUMAR, R., AND SIVAKUMAR, D. 2003. Comparing top k lists. In Proceedings of the ACMSIAM Symposium on Discrete Algorithms (SODA). 28–36.

[9] FERN, X. Z. AND BRODLEY, C. E. 2003. Random projection for high dimensional data clustering: A cluster ensemble approach. In Proceedings of the International Conference on Machine Learning (ICML). 186–193.

[10] FILKOV, V. AND SKIENA, S. 2004. Integrating microarray data by consensus clustering. Int. J. AI Tools 13, 4, 863–880. FRED, A. AND JAIN, A. K. 2002. Data clustering using Evidence accumulation. In Proceedings of the International Conference on Pattern Recognition (ICPR). 276–280.

[11] Snehal Mahajan, Dharamaraj Patil, ”Image Retrieval Using Contribution-based Clustering”, Algorithm with Different Feature Extraction Techniques”978-1-4799-3064-7/14 IEEE 2016

[12] Y. Poornima, P.S. Hiremath,”Efficient Modelling of Visual Art Color Image Clustering”, International Journal of Computer Applications (0975 – 8887) Volume 91 – No.11, April 2014

[13] Yogita Mistry, D. T. Ingole, and M. D. Ingole,” Efficient Content Based Image Retrieval Using Transform and Spatial Feature Level Fusion”, The 2nd International Conference on Control, Automation and Robotics. 978-1-4673-9859-61/16 IEEE.

[14] Naveena A K and N K Narayanan,"Image Retrieval using Combination of Color, Texture and Shape Descriptor,"©(2016) IEEE, pp-958-962