Multiple Clustering Views for Data Analysis

Ms. Harsha Sondawale[1], Prof. N. M. Shahane[2]

[1]Student, ME Computer, KKWIEER, Savitribai Phule Pune University, India

[2]Associate Professor, Department of computer Engineering, KKWIEER, Savitribai Phule Pune University, India

ABSTRACT

Many clustering algorithms provide single clustering solution which is not sufficient for the analysis of the data. Complex data can be represented in many different ways. In general, the clustering of high dimension data is a hard problem. The different views of the data and the use of relationship between these views are useful to solve the problem of clustering. For this purpose a survey is done that finds the multiple alternative clustering views which will be helpful for the purpose of exploratory data analysis and provides the multiple meaningful, different or non-redundant alternative clustering solutions.

Keywords:- alternative clustering, multiple clustering, spectral clustering, multi-view clustering

1. INTRODUCTION

Clustering is the most widely used techniques for exploratory data analysis. Clustering is nothing but the task of grouping a set of objects in such a way that objects in the same group or cluster are more similar to each other than to those in other groups or clusters. Clustering is similar to classification. Clustering is an unsupervised way of learning. The main objective of clustering is to determine the essential grouping in a set of unlabeled data. Cluster analysis divides data into groups or clusters that are meaningful and useful. Some clustering techniques characterize each cluster in terms of a cluster prototype. Data clustering is the main task in artificial intelligence. Human can easily identify the cluster in low dimension with fewer amounts of data, but in case of computer it is very difficult to instruct the computer to find such a relationship. As the dimension of the data increases human also has difficulties in finding the interesting structure for the data. The goal of exploratory data analysis is to find structure and interesting patterns in data or to extract information from data. But many clustering algorithms provide single clustering solution which is not sufficient for analysis of data. Complex data can be represented in many different ways for different purposes. In high dimensional data, different structure of the data may be shown by different feature subspaces then in such a case why to depend or provide a single clustering view while different other alternative clustering views might be useful for different purposes. It may be found that single clustering is not useful and actionable so there is a need to find the alternative to it. Most of the clustering algorithm provides only single partitioning of the data which is not sufficient for the analysis of data. Different or multiple clustering solutions or multiple views of data are helpful for various purposes. That means multiple set of clusters can provide more insights than only one solution. Data items can be grouped together in many different ways for different purposes. For example, face image can be grouped based on their pose or an identity. In the similar way in case of medical data, what is interesting to physicians might be different from what is interesting to insurance companies [1]. The goals and objectives of multiple clustering are to group each object in multiple clusters, representing the different perspective on the data. The result of multiple clustering must contain many alternative solutions and the user can choose one or use multiple of these solutions according to his/her purpose. Solutions should have to differ to a high extent, and thus, each of these solutions provides additional knowledge that is enhanced extraction of knowledge. There are various application of clustering. In gene expression analysis, clustering is performed on gene database to derive multiple functional roles of genes. Objects are the genes described by their expression or behavior under different condition. Objective of this is to form the groups of genes with similar function. As one gene may have multiple functions, combining them into single group is not sufficient. Hence the challenges are to form the multiple groups of genes according to their functional roles. Another example is customer segmentation. Here the need of clustering customer profiles is to derive their interest. In this the objects are customers described by profiles and the aim is to form group of customers with similar behavior. Text analysis is also one of the example of clustering. In text analysis the objects are text documents described by their content and the aim is to form groups of documents on similar topic. This can be performed by forming multiple alternative clustering solutions.

2. LITERATURE REVIEW

Data Clustering is the most popular aspect to derive the classes or cluster, desired groups of pattern and concept. It is very difficult to find the clustering or grouping of the data and finding interesting patterns about which nothing is known already. This also leads to the general problem of how to find out the knowledge of the data for mining and exploration. Although the literature on clustering is huge, there has been relatively little attention paid to the problem of finding multiple non-redundant clustering. There are two ways to find multiple alternative clustering solutions. One is to find multiple solutions simultaneously and other is to find alternative solutions iteratively. Gondek and Hofmann [2] suggest information-theoretic framework that makes use of the concept of conditional mutual information. In this the important problem of non-redundant data clustering is also investigated based on the idea of maximizing conditional mutual information relative to given background knowledge. This approach is dependent on distributed assumption. Bae and Bailey [3] utilize “cannot-link constraint” and agglomerative clustering to find alternative clustering. Cui et al.[4],[5], finds different clustering views by clustering the subspace orthogonal to the clustering solution found in previous iterations without making use of specific clustering algorithm. CAMI [7] simultaneously discovers two disparate clustering by optimizing cluster quality, quantifying these criteria by maximizing the mutual likelihood of Gaussian mixture models and minimizing mutual information between them. The method described in [6] is based on k-means and CAMI [7] both are limited to convex clusters. Another related work is based on subspace clustering. In the above methods they did not use spectral clustering. The aim of subspace clustering is to find the clusters which are hidden in high dimensional space. There are two major branches of subspace clustering based on search strategy that are top-down algorithm and bottom-up approach. Top-down algorithm finds an initial clustering in the full set of dimension and evaluate the subspaces of each cluster. While in case of bottom-up approach, it finds dense region in low dimension spaces and combine them to form clusters [8].

3. MTHODOLOGY

3.1 Iterative optimization clustering algorithm

A large number of clustering algorithm are based on iterative optimisation, such as k-means and Gaussian Expectation Maximization algorithms are the two popular algorithm. These algorithms always start with a solution and then repeatedly improve the solution until no further improvement can be made. The initialization of these algorithm are important. But it is one of the limitation of this algorithm. In general these algorithm do not give global optimality of their solutions. The time complexity of iterative optimization clustering algorithm is between O(nkd) and O(nkd2) per iteration, and the number of iteration can vary depending on initialization and the structure of data. Here n is the number of objects in dataset, k is number of clusters and d is dimension.

3.2 Hierarchical clustering

Hierarchical clustering is another approach of data clustering. Algorithm in this class use heuristic splitting and merging functions to either build a cluster tree from the top down or the bottom up. Algorithm build from top-down are called divisive while the bottom-up algorithms are called merge-based. Agglomerative algorithms begin at the top of the tree. The agglomerative hierarchical clustering method builds the hierarchy from the individual elements by progressively merging clusters. Each agglomeration occurs at a greater distance between cluster than the previous agglomeration, and one can decide to stop clustering either when the clusters are too far apart to be merged or when there is a sufficient small number of clusters. Hierarchical clustering is used for the application where it is important to see the structure of the data at various levels of granularity, for example in clustering gene expression data.

Hierarchical clustering algorithms operate on a pair wise distance matrix of size n × n, they need time O(n2d) per split or merge and after constructing the full cluster tree, it requires n merges or split for total time of O(n3d) where d is dimension.

3.3 Information bottleneck method

Information bottleneck method is introduced by Naftali Tishby et al.[9] for finding the best trade-off between accuracy and complexity when clustering a random variable X, given a joint probability distribution between X and an observed relevant variable Y. Other application involve distributed clustering and dimension reduction. It has generalized the classical notion of minimal sufficient statistics from parametric statistics to arbitrary distributions, not necessarily of exponential forms. It does so by relaxing the sufficiency condition to capture some fraction of the mutual information with the relevant variable Y. The compressed variable is T and algorithm minimises the following quantity

minp(t|x) I (X; T) – βI (T; Y) where I (X; T) and I (T; Y) are the mutual information between X;T and T;Y respectively and β is a Lagrange multiplier.

3.4 Orthogonal Partitioning Clustering

This clustering method combines a novel partitioning active sampling technique with an axis parallel strategy to identify continuous areas of high density in input space. It has two major contributions first one is that it uses statistical test to validate the quality of cutting plane and second is that it can operate on a small buffer containing a random sample from the original data set.

3.5 Spectral clustering

Spectral clustering has become one of the most popular clustering algorithm. It is simple to implement and can be solved efficiently. Spectral clustering technique has been developed from spectral graph theory. New spectral clustering algorithms search for globally optimal cuts in a graph representing the data to be clustered. The graph G can be thought of as analogous to a pair wise distance matrix, though the edges in the graph need not be metric distance.

Spectral clustering algorithm have been developed by computer vision researchers and graph theorists. An advantage of spectral clustering is the ability to form clusters of arbitrary shapes because the graph are fully connected and do not assume any underlying model for any cluster. Hence any data point may be considered connected with any other point.
Spectral clustering depends on the Eigen structure of a similarity matrix. In spectral clustering, clusters are formed by dividing data points using similarity matrix. It has three main stages[10], which are pre-processing, spectral mapping and post mapping. Construction of similarity matrix is performed through pre-processing, spectral mapping deals with the construction of Eigen vectors for the similarity matrix and grouping of the data points are performed by post processing. There are various advantages of spectral clustering such as there is no need to make any assumption on the cluster shape and it is simple to implement. The major drawback of spectral clustering is that it requires high computational complexity for large dataset.

4. DISCUSSION

In most of the research papers of multiple views clustering, the clustering is applied on both synthetic data and the real-world data to check whether the different algorithm gives reasonable alternative clustering solutions with good quality or not. WebKB dataset is the most popular dataset used in clustering. This dataset include number of html documents from four universities. These four universities are Cornell, University of Washington, University of Wisconsin, and University of Texas. These web pages can be grouped into different clusters. Some famous data sets coming from UCI repository are suitable for the purpose of finding multiple clustering views. For example, the face data set from UCI KDD repository [11] contains different face images of multiple people taken at varying poses and the purpose is to find the multiple alternative clustering solutions depending on their poses where each person’s identity is used as existing clustering solution. There are also a number of other multimedia datasets which are usually used in experiments to extract different multiple views of the data.

5. PROPOSED APPROACH

Spectral clustering is a well accepted method for clustering. It is not sensitive to outliers or shape of clusters. In proposed approach the main focus is on spectral clustering due to its flexibility and its increasing popularity in applications. However, because of the machine limitations, there is a serious empirical barrier in applying this method for large data sets. So there is a need to modify spectral clustering that will be applicable on large size datasets. The proposed approach is based on faithful (Information Preserving) sampling. This sampling method can be used in combination with other graph-based clustering algorithms with different objective functions to reduce size of the data.

6. CONCLUSION

In many scenarios, more than one view can be provided to describe the data. Multiple views of data are helpful for various purposes. Multiple set of clusters can provide more insights than only one clustering solution. For this purpose the survey of different clustering techniques are discussed, which are used for finding multiple clustering solutions. It is observed that among all clustering techniques, spectral clustering is widely used but it has major drawback that it has high computational complexity for large dataset. To overcome this drawback, modified spectral clustering technique is used in proposed approach which supports large dataset.

REFERENCES

[1] Donglin Niu, Jennifer G. Dy, Michael I. Jordan, “Iterative Discovery of Multiple Alternative Clustering Views”. IEEE transaction on pattern analysis and machine intelligence, vol. 36, no.7, July 2014.
[2] D. Gondek and T. Hofmann, “Non-Redundant Data Clustering,” Proc. IEEE Int’l Conf. Data Mining, pp. 75-82, 2004.
[3] E. Bae and J. Bailey, “COALA: A Novel Approach for the Extrac-tion of an Alternate Clustering of High Quality and High Dissim-ilarity,” Proc. IEEE Int’l Conf. Data Mining, pp. 53-62, 2006.
[4] Y. Cui, X.Z. Fern, and J. Dy, “Non-Redundant Multi-View Clus-tering via Orthogonalization,” Proc. Seventh IEEE Conf. Data Min-ing (ICDM ’07), pp. 133-142, 2007.
[5] Y. Cui, X.Z. Fern, and J.G. Dy, “Learning Multiple Nonredun-dant Clusterings,” ACM Trans. on Knowledge Discovery from Data, vol. 4, no. 3, Article 15, 2010.
[6] P. Jain, R. Meka, and I.S. Dhillon, “Simultaneous Unsupervised Learing of Disparate Clustering,” Proc. SIAM Int’l Conf. Data Min-ing, pp. 858-869, 2008.
[7] X.H. Dang and J. Bailey, “Generation of Alternative Clusterings Using the CAMI Approach,” Proc. SIAM Int’l Conf. Data Mining, pp. 118-129, 2010.
[8] L. Parsons, E. Haque, and H. Liu, “Subspace Clustering for High Dimensional Data: A Review,” ACM SIGKDD Explorations News-letter, vol. 6, no. 1, pp. 90-105, 2004.

[9] N. Tishby, F.C. Pereira, and W. Bialek: “The Information Bottleneck method”. The 37th annual Allerton Conference on Communication, Control, and Computing, Sep 1999: pp. 368–377

[10] M Meila, D Verma, 2001. Comparison of spectral clustering algorithms. University of Washington, technical report

[11] S.D. Bay, “The UCI KDD Archive,” 1999. http://kdd.ics.uci.edu.

AUTHOR

Harsha Ashok Sondawale received the B.E. degree in Information Technology from Smt. Rajshree Mulak College of Engineering in 2012, respectively. She is currently pursuing her Masters degree in Computer Engineering from K. K. Wagh Institute of Engineering Education and Research, Savitribai Phule Pune University Former UoP. This paper is published as a part of the research work done for the degree of Masters.

Prof. N. M. Shahane is a Associate Professor in the Department of Computer Engineering, K. K. Wagh Institute of Engineering Education and Research, Savitribai Phule Pune University. His current research interests include pattern recognition, digital signal processing, machine learning, data mining and mathematical modeling.