A Graph Based Approach to Multiview Clustering

Moumita Saha

Department of Computer Science and Engineering,
Indian Institute of Technology Kharagpur
moumitasaha@cse.iitkgp.ernet.in

Abstract. Rich and complex data sets prevalent in many applications can often be explored from multiple perspectives. Examples include clustering of multimedia, multilingual and heterogeneously linked data sets. Multiview clustering attempts to discover clusters from different views of the same data set. In this article, construction of subspace representation of the views and subsequently clustering the subspaces to produce multiview clusters is been proposed. The subspaces are obtained by a separate clustering procedure on the nearest neighbour graphs of the individual features. Three graph similarity measures are used for this clustering. Empirical results on three benchmark data sets shows that the proposed method provides superior performance in terms of classification accuracy using known class labels as compared to single view clustering of the entire data sets.

Keywords: Multiview clustering, graph measures, subspace clustering.

1 Introduction

Multi-view Clustering is an emerging topic in the field of data mining. In richly structured data the entities can be observed or modelled from various perspectives leading to multiple views or representations. Multi-view learning is an useful approach to effectively explore and exploit the information from heterogeneous data for the purpose of improving the learning performance. Multi-view algorithms deals with each view of the data independently and then merge the solutions to obtain a complete, robust pattern which is superior compared to its single-view representation.

In this paper, an approach for multiview clustering of data is proposed. A graph based methodology is adopted. Multiple clusterings of the graphs are obtained and used to represent the data from multiple views.

The paper is organised as follows. Review of some related work on multiview clustering is presented in section 2. The proposed graph based approach is delivered in section 3. Analysis and illustration of the experimental results is presented in section 4 and finally the paper is concluded in section 5.

P. Maji et al. (Eds.): PReMI 2013, LNCS 8251, pp. 128–133, 2013.

2 Related Work

Multi-view clustering is becoming increasingly important in several application. Authors have proposed several approaches to address this problem. Among the earliest efforts, Bickel and Scheffer [1] propose partitioning and agglomerative, hierarchical multi-view clustering and apply them to text data. It is shown that multi-view versions of k-Means and EM outperform their respective single view as they optimize agreement between views. However, it shows negative result for agglomerative hierarchical clustering as the mixture components have smaller overlap when the views are concatenated.

In [2], both textual and visual contents of the retrieved images are considered and a multi-view clustering approach is proposed to re-rank the web retrieved result provided by text based search engine. Kim et al. [3] proposes a multiview clustering method for multilingual documents. The proposed approach is an incremental algorithm which first groups documents having the same patterns assigned by view-specific probabilistic latent semantic analysis (PLSA) models, and then the remaining un-clustered documents are assigned to the groups using a constrained PLSA model.

Greene and Cunningham [4] present an analysis of the research themes in a bibliographic literature network, based on the integration of both co-citation links and text similarity relationships between papers in the network. In [5], Bruno and Maillet investigated a late fusion approach for multi-view clustering based on the latent modelling of cluster-cluster relationships and it is shown to outperform an early-fusion approach based on multi-view feature correlation analysis.

Chaudhuri et al. [6] propose method for projection of higher dimensional data into its lower dimensional-subspace using multiple views of data by using Canonical Correlation Analysis (CCA). Results for mixtures of Gaussian and mixtures of log concave distributions are shown to be effective.

Several other approaches to multiview clustering are reported in literature. Many of them explore projected subspaces to obtain individual clustering views and finally combine them to provide different perspectives of the data. The graph based approach for finding appropriate single view subspace clusters for the data in multiview framework is described in the next section.

3 Graph Based Approach to Multi-view Clustering

Multi-view clustering is performed utilizing a graph representation of subspaces and hierarchical agglomerative clustering. A three stage approach is proposed for this. First, the data set is represented as graphs with nodes as data points and the link as feature wise similarity. Next various graph similarity measures is used to evaluate the similarity between the graphs and utilize them to cluster the graphs into different groups. Each of these clusters represents separate view or representation of the data. Finally, for each of the view, clustering on the corresponding projected subspaces is applied in order to obtain final clusters of instances for each of the views. The schema of the approach is shown in figure 1.

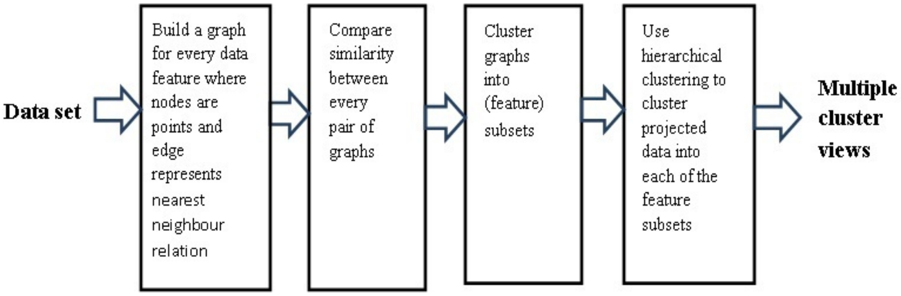

Data set
Build a graph for every data feature where nodes are points and edge represents nearest neighbour relation
Compare similarity between every pair of graphs
Cluster graphs into (feature) subsets
Use hierarchical clustering to cluster projected data into each of the feature subsets
Multiple cluster views

Fig. 1. Block diagram of graph theoretic multiview clustering

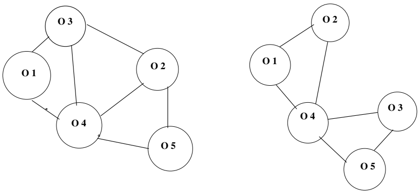
3.1 Feature Graph Construction

The data set is represented as graph where nodes are the data points and edges exist between nearest neighbour nodes. d number of graphs is build for a d-dimensional dataset representing each of the features of the data set. As a pre-processing step all the duplicate values of the data points are removed and data points are arranged in sorted order in order to obtain the nearest neighbour fast.

Illustration of representation of a toy data set is shown in table 1.

Table 1. Illustration of an example graph representation

Point	F1	F2
O1	2	200
O2	7	190
O3	5	150
O4	6	180
O5	12	160
O6	2	200



3.2 View Extraction

The subsequent step is view extraction of the data. Different views are extracted by assembling the similar features together. Agglomerative clustering is used to cluster the features into different groups which will represent the views. Similarity measures considered are the following.

Cumulative Degree Centrality (CDC) : The degree centrality of a node v, for a given graph $G := (V, E)$ with $\mid V \mid$ vertices and $\mid E \mid$ edges, is defined as $C_D(\mathrm{v}) = \mathrm{deg(v)}$.

If n is the number of nodes of the graph G then cumulative degree centrality is evaluated as in equation 1.

$$\mathit{Cumulative\ degree\ centrality\ of\ G} \ = \ \Sigma_{i=1}^{n}\ \mathit{deg}(v_i) \qquad (1)$$

Edges Count of the Graph (EC) : Edges are the connections between the vertices of the graph. This set is denoted by $E(G)$. If $G := (V, E)$ represents a graph, individual edges are pairs $\{u, v\}$ where u and v are vertices in V. The total number of edges of each feature graph is considered as a measure of the corresponding graphs.

Similar Edge Count (SEC) : The third approach taken is evaluating the number of similar edges between two graphs.

If $G := (V, E)$ and $F := (V', E')$ are two graphs, then two edges $(u, v) \in E$ and $(u', v') \in E'$ are said to be similar if their end vertices are same, that is either

$$u = u' \quad \text{and} \quad v = v' \qquad \text{or} \qquad u = v' \quad \text{and} \quad v = u'.$$

Thus, each pair of graph is compared on the basis of number of similar edges between the graphs.

Three more measure are evaluated which are combination of the above measures. They are combination of degree centrality and number of similar edges (CDC + SEC), combination of degree centrality and number of edges (CDC + EC), combination of degree centrality, number of edges and number of similar edges (CDC + EC + SEC) between the graphs.

The approach considered is as follows:

1. The above mentioned measures are evaluated for each of the feature graph.
2. This measures are taken as the similarity measure for executing agglomerative clustering to obtain the cluster of features.
3. Dendograms obtained from clustering results are analysed and the features are grouped into different clusters. Each cluster of features represents a view of the data. Three clusters, i.e. three views of the data based on best accuracy performance are considered.

3.3 Clustering the Views

The final step of the algorithm is clustering the instances according to the different views obtained to get a better analysis of the group of instances with respect to the views. The approach followed is depicted below.

1. The data set is projected according to the features obtained in each view.
2. The reduced projected view of data is clustered utilizing complete linkage agglomerative clustering. The number of cluster is set by user. If class label is available it is set as the number of classes.

The accuracy of clustering is evaluated in terms of classification accuracy as follows.

(a) Evaluate the count of each class label in each group of instances.

(b) Calculate the maximum count and assign the particular group with that maximum count class label.

(c) Repeat the process until all the group is assigned a class label, keeping in mind no two group can be assigned the same class label.

The above process is repeated for every view of the data and this multi-view approach is compared to its single view counterpart.

4 Result and Analysis

The proposed graph-based approach to multiview clustering is applied on three datasets - SPECTF Heart dataset, Ozone Layer dataset and SECOM dataset. The description of dataset is shown in table 2.

Table 2. Datasets description

Dataset	No. instances	No. features	No. classes
SPECTF Heart	267	44	2
Ozone	2537	72	2
SECOM	1561	590	2

The result is depicted in two major steps. First the outcome of views extraction is evaluated and then the final accuracy of clustering of instances according to multiple views are established.

Table 3 shows as an example the resulting views of the SPECTF Heart data obtained from clustering of the feature graph taking into consideration different similarity measures. The accuracy of each views of three data sets are depicted in tables 4, 5, 6. Single view clustering is performed by taking all the features of the data to cluster the data points and the result is compared to multiview clustering outcome. It is observed that the accuracy improves in case of multiview clustering as compared to its single-view counterpart. This fact is depicted in table 7

Table 3. Multiple views of SPECTF Heart

Similarity Measures	view 1	view 2	view 3
CDC	1, 8	2-6, 9-44	7
EC	31	1-24, 27-40	25-26, 41-44
SEC	5	2-4,6-44	1
CDC + SEC	1, 8	2-6, 9-44	7
CDC + EC	31	1-24, 27-40	25-26, 41-44
CDC + EC + SEC	31	1-24, 27-40	25-26, 41-44

Table 4. Accuracy (%) of different views for SPECTF Heart

Similarity Measures	view 1	view 2	view 3
CDC	74.44	72.22	72.96
EC	77.03	72.22	73.33
SEC	75.18	64.81	62.92
CDC + SEC	74.44	72.22	72.96
CDC + EC	77.03	72.22	73.33
CDC + EC + SEC	77.03	75.92	73.70

Table 5. Accuracy (%) of different views for Ozone

Similarity Measures	view 1	view 2	view 3
CDC	97.32	94.17	94.07
EC	85.03	84.28	85.03
SEC	78.08	94.10	62.17
CDC + SEC	94.17	97.32	94.07
CDC + EC	86.14	60.07	94.10
CDC + EC + SEC	85.03	84.28	94.07

Table 6. Accuracy (%) of different views for SECOM

Similarity Measures	view 1	view 2	view 3
CDC	86.29	87.82	82.70
EC	81.74	93.01	93.08
SEC	93.27	91.15	93.27
CDC + SEC	74.95	87.82	82.70
CDC + EC	93.01	75.78	86.80
CDC + EC + SEC	93.08	71.49	85.07

Table 7. Accuracies (%) of single vs multiview clustering

Data set	Single view clustering	Multi view clustering	Best view of multi view clustering	Corresponding similarity measure
SPECTF Heart	73.70	77.03	view 1	EC & CDC + EC & CDC + EC + SEC
Ozone	94.10	97.32	view 2	CDC + SEC
SECOM	91.15	93.27	view 1 & view 3	SEC

5 Conclusion

A graph based approach is presented for multiview clustering. Nearest neighbour graphs of data point projected to individual features are first constructed using three different similarity measures between these graphs, they are clustered to feature subspaces. The subspaces correspond to individual views. Hierarchical agglomerative clustering is performed on each of these views to obtain multiview clusters. The multiview clusters are evaluated in terms of their classification accuracy and is found to provide superior performance than single view clustering. In future, use of other graph similarity measures and clustering algorithms may be explored to improve performance.

References

1. Bickel, S., Scheffer, T.: Multi-View Clustering. In: Proceeding of 4th IEEE International Conference on Data Mining, pp. 19–26 (2004)
2. Chi, M., Zhang, P., Zhao, Y., Feng, R., Xue, X.: Web Image Re-Ranking with Multi-view Clustering. In: Proceeding of 18th Int. World Wide Web Conference, pp. 1189–1190. ACM (2009)
3. Kim, Y.-M., Reza Amini, M., Goutte, C., Gallinari, P.: Multi-View Clustering of Multilingual Documents. In: Proceeding of 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 821–822 (2010)
4. Greene, D., Cunningham, P.: Multi-View Clustering for Mining Heterogeneous Social Network Data. In: Proceeding of Workshop on Information Retrieval over Social Networks, 31st European Conference on Information Retrieval, ECIR 2009 (2009)
5. Bruno, E., Maillet, S.M.: Multiview Clustering: A Late Fusion Approach using Latent Models. Proceedings of 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, 736–737 (2009)
6. Chaudhuri, K., Kakade, S.M., Livescu, K., Sridharan, K.: Multi-View Clustering via Canonical Correlation Analysis. In: Proceedings of 26th Annual International Conference on Machine Learning, ICML 2009, pp. 129–136 (2009)