International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 10, Issue 6 (June 2014), PP.25-30

Enhancing Clustering Results In Hierarchical Approach By Mvs Measures

RAMANA REDDY BANDI
Asst.Prof In CSE Dept Annamacharya Institute Of Technology And Sciences , Tirupathi, ANDHRA PRADESH, INDIA

Abstract:- Clustering is a process of grouping set of objects into classes of clusters, so that the objects in the same cluster are more similar to each other than to those in other clusters. Data Clustering has galore techniques in representing the relationship among data objects by similarity in features. All similarity measures play a vital role in achieving the accuracy of results in clustering. So to measure this similarity some form of measurement is needed. In this paper, we introduce multi-viewpoint based similarity (MVS) approach with traditional dissimilarity/ similarity measure which utilizes more than one viewpoint. The important contribution of our work is to find clustering tendency by visual assessment tendency(VAT)analysis and to discover clusters using MST based clustering.

Index terms- Data clustering, Similarity measure, VAT analysis, MST based clustering

I. INTRODUCTION

Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often proximity according to some defined distance measure. Data clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. The computational task of classifying the data set into k clusters is often referred to as k-clustering. Document clustering aims to group, in an unsupervised way, a given document set into clusters such that documents within each cluster are more similar between each other than those in different clusters. It is an enabling technique for a wide range of information retrieval tasks such as efficient organization, browsing and summarization of large volumes of text documents. Cluster analysis aims to organize a collection of patterns into clusters based on similarity. Clustering has its root in many fields, such as mathematics, computer science, statistics, biology, and economics. In different application domains, a variety of clustering techniques have been developed, depending on the methods used to represent data, the measures of similarity between data objects, and the techniques for grouping data objects into clusters.

II. MULTI-VIEW POINT BASED SIMILARITY

It is possible to use more than one point of reference to construct a new concept of similarity and we may have a more accurate assessment of how close or distant pair of points is, if we look at them from many different viewpoints. From a third point d_h, the distances and directions to d_i and d_j are indicated respectively by the difference vectors $(d_i - d_h)$ and $(d_j - d_h)$. By standing at various reference points d_h to view d_i, d_j and working on their difference vectors, the similarity between the two documents can be defined as:

$$Sim(d_i, d_j)_{d_i, d_j \in S_r} = \frac{1}{n - n_r} \sum_{d_h \in S \setminus S_r} Sim(d_i - d_h, d_j - d_h)$$

Two objects that are to be measured must be in the same cluster, while the points from where measurement should be established must be outside the cluster. We call this proposal the multi-Viewpoint based similarity, or MVS. From this point onwards, we will denote the proposed similarity measure between two document vectors d_i and d_j by $\mathrm{MVS}\,(d_i, d_j \setminus d_i, d_j \in S_r)$,or occasionally $\mathrm{MVS}(d_i, d_j)$for short.

If the relative similarity is defined by dot-product of the difference vectors, we have:

$$\mathrm{MVS}(d_i, d_j \setminus d_i, d_j \in S_r)$$

$$= \frac{1}{n - n_r} \sum_{d_h \in S \setminus S_r} (d_i - d_h)^t (d_j - d_h)$$

$$= \frac{1}{n - n_r} \sum_{d_h} \cos (d_i - d_h, d_j - d_h) \parallel d_i - d_h \parallel \parallel d_j - d_h \parallel$$

III. MULTI-VIEWPOINT BASED CLUSTERING

3.1 Two clustering criterion functions I_R and I_v

We now formulate our clustering criterion functions. The first clustering criterion function is I_R which is the cluster size-weighted sum of average pair-wise similarities of documents in the same cluster. The sum in a general form can be expressed by the function F:

$$F = \sum_{r=1}^{k} nr \left[\frac{1}{n_r^2} \sum_{d_i, d_j \in S_r} sim(d_i, d_j) \right]$$

To perform the optimization procedure in a simple, fast and efficient way we would like to transform the above objective function into some suitable form

$$\sum_{d_i, d_j \in S_r} sim(d_i, d_j)$$

$$= \sum_{d_i, d_j \in S_r} \frac{1}{n - n_r} \sum_{d_h \in S \setminus S_r} (d_i - d_h)^t (d_j - d_h)$$

$$= \frac{1}{n - n_r} \sum_{d_i, d_j} \sum_{d_h} (d_i{}^t d_j - d_i{}^t d_h - d_j{}^t d_h + d_h{}^t d_h)$$

Since

$$\sum_{d_i \in S_r} d_i = \sum_{d_j \in S_r} d_j = D_r,$$

$$\sum_{d_h \in S \setminus S_r} d_h = D - D_r \text{ and } \parallel d_h \parallel = 1,$$

We have,

$$\sum_{d_i, d_j \in S_r} Sim(d_i, d_j)$$

$$= \sum_{d_i, d_j \in S_r} d_i{}^t d_j - \frac{2n_r}{n - n_r} \sum_{d_i \in S_r} d_i{}^t \sum_{d_h \in S \setminus S_r} d_h + n_r{}^2$$

$$= D_r{}^t D_r - \frac{2n_r}{n - n_r} D_r{}^t (D - D_r) + n_r{}^2$$

$$= \frac{n + n_r}{n - n_r} \parallel D_r \parallel^2 - \frac{2n_r}{n - n_r} D_r{}^t D + n_r{}^2$$

Substituting into F to get:

$$F = \sum_{r=1}^{k} \frac{1}{n_r} \left[\frac{n + n_r}{n - n_r} \| D_r \|^2 - \left(\frac{n + n_r}{n - n_r} - 1 \right) D_r{}^t D \right] + n$$

Where n is a constant, maximizing F is equivalent to maximizing $\overline{F}$:

$$\overline{F} = \sum_{r=1}^{k} \frac{1}{n_r} \left[\frac{n + n_r}{n - n_r} \| D_r \|^2 - \left(\frac{n + n_r}{n - n_r} - 1 \right) D_r{}^t D \right]$$

Hence the λ is integrated into the expression of $\overline{F}$ to become:

$$\overline{F}_{\lambda} = \sum_{r=1}^{k} \frac{\lambda_r}{n_r} \left[\frac{n + n_r}{n - n_r} \| D_r \|^2 - \left(\frac{n + n_r}{n - n_r} - 1 \right) D_r{}^t D \right]$$

Now let us a parameter α called the regulating factor, which has a constant value αϵ[0,1] and from the above we know that $\lambda_r = {n_r}^{\alpha}$.Now the final form of our criterion function I_R is:

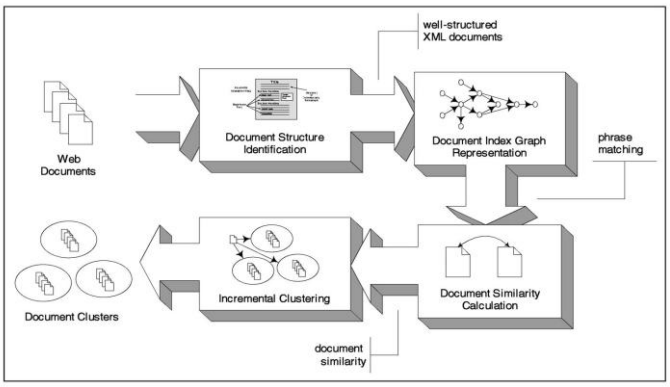
$$I_R = \sum_{r=1}^{k} \frac{1}{{n_r}^{1-\alpha}} \left[\frac{n + n_r}{n - n_r} \| D_r \|^2 - \left(\frac{n + n_r}{n - n_r} - 1 \right) D_r{}^t D \right]$$

The second criterion function is I_V which calculates the weighted difference between the two terms: $\| D_r \| \text{ and } D_r{}^t D / \| D_r \|$

$$I_V = \sum_{r=1}^{k} \left[\frac{n + \| D_r \|}{n - n_r} \| D_r \| - \left(\frac{n + \| D_r \|}{n - n_r} - 1 \right) \frac{D_r{}^t D}{\| D_r \|} \right]$$

Which represent an intra-cluster similarity measure and an inter-cluster similarity measure, respectively.

IV. CLUSTER ANALYSIS ARCHITECTURE



The figure represents the process of forming the clusters. The web documents are taken as the inputs and the structure of the documents is verified. Now the documents are represented in the form of graph and then the document similarity will be calculated between two documents. Finally the clusters are formed based on the similarity measure.
The similarity between two documents is done by the cosine-similarity measure.

$$TF = C / T$$

$$IDF = D / DF$$

D −>quotient of the total number of documents

DF−> number of times each word is found in

the entire corpus

C−> quotient of no of times a word appears in

each document

T −>total number of words in the document

$$TFIDF = TF * IDF$$

V. HIERARCHICAL DOCUMENT CLUSTERING USING FREQUENT ITEM SETS

Document clustering has been studied intensively because of its wide applicability in areas such as web mining, search engines, information retrieval, and topological analysis. Unlike in document classification, in document clustering no labeled documents are provided. Although standard clustering techniques such as k-means can be applied to document clustering, they usually do not satisfy the special requirements for clustering documents: high dimensionality, high volume of data, ease for browsing, and meaningful cluster labels. In addition, many existing document clustering algorithms require the user to specify the number of clusters as an input parameter and are not robust enough to handle different types of document sets in a real-world environment. For example, in some document sets the cluster size varies from few to thousands of documents. This variation tremendously reduces the clustering accuracy for some of the state-of-the art algorithms. Frequent Itemset-based Hierarchical Clustering (FIHC), for document clustering based on the idea of frequent itemsets proposed by Agrawalet. al. The intuition of our clustering criterion is that there are some frequent itemsets for each cluster (topic) in the document set, and different clusters share few frequent itemsets. A frequent itemset is a set of words that occur together in some minimum fraction of documents in a cluster. Therefore, a frequent itemset describes something common to many documents in a cluster. In this technique we use frequent itemsets to construct clusters and to organize clusters into a topic hierarchy. Here are the features of this approach.

- *Reduced dimensionality*. This approach uses only the frequent items that occur in some minimum fraction of documents in document vectors, which drastically reduces the dimensionality of the document set. Experiments show that clustering with reduced dimensionality is significantly more efficient and scalable. This decision is consistent with the study from linguistics (Longman Lancaster Corpus) that only 3000 words are required to cover 80% of the written text in English and the result is coherent with the Zipf's law and the findings in Mladenic et al. and Yang et al.

- *High clustering accuracy*. Experimental results show that the proposed approach FIHC outperforms best documents clustering algorithms in terms of accuracy. It is robust even when applied to large and complicated document sets.

- *Number of clusters as an optional input parameter*. Many existing clustering algorithms require the user to specify the desired number of clusters as an input parameter. FIHC treats it only as an optional input parameter. Close to optimal clustering quality can be achieved even when this value is unknown.

VI. MINIMUM SPANNING TREE CLUSTERING

The MST clustering algorithm is known to be capable of detecting clusters with irregular boundaries. Unlike traditional clustering algorithms, the MST clustering algorithm does not assume a spherical shaped clustering structure of the underlying data. The EMST clustering algorithm uses the Euclidean minimum spanning tree of a graph to produce the structure of point clusters in the n-dimensional Euclidean space. Clusters are detected to achieve some measure of optimality, such as minimum intra cluster distance or maximum inter cluster distance. The EMST clustering algorithm has been widely used in practice. An example application of the algorithm is image color clustering in web image analysis. Web images are usually supplied with shaded or multicolored complex backgrounds, often found in photographs, maps, engineering drawings and commercial advertisements. Analyzing web images is a challenging task due to their low spatial resolution and the large number of colors in the images. The purpose of color clustering in web image analysis is to reduce thousands of colors to a representative few that clearly differentiate objects of interest in an image. Once the MST is built for a given input, there are two different ways to produce a group of clusters.

6.1 PRIMS ALGORITHM

Like Kruskal's algorithm, Jarnik's algorithm, as described in CLRS, is based on a generic minimum spanning tree algorithm. The main idea of Jarnik's algorithm is similar to that of Dijkstra's algorithm for finding shortest path in a given graph. The Jarnik's algorithm has the property that the edges in the set A always form a

single tree. We begin with some vertex v in a given graph G =(V, E), defining the initial set of vertices A. Then, in each iteration, we choose a minimum-weight edge (u, v), connecting a vertex v in the set A to the vertex u outside of set A. Then vertex u is brought in to A. This process is repeated until a spanning tree is formed. Like Kruskal's algorithm, here too, the important fact about MST is that we always choose the smallest-weight edge joining a vertex inside set A to the one outside the set A. The implication of this fact is that it adds only edges that are safe for A; therefore when the Jarnik's algorithm terminates, the edges in set A form a minimum spanning tree, MST.

VII. VAT (VISUAL ASSESSMENT TENDENCY) ANALYSIS

Our VAT algorithm is meant to replace the straight- forward visual displaying part of the VAT algorithms Or, for that matter, it can start from an ordered dissimilarity matrix from any algorithm of that kind. Instead of displaying the matrix as a 2-dimensional gray-level image for human interpretation, VAT analyzes the matrix by taking averages of various kinds along its diagonal and produces the tendency curves, with the most useful of them being the d-curve. This changes 2D data (a matrix) into a 1D array, which is certainly easier to both human eyes and the computer since the concentration is now only on one variable—the height. Possible cluster structure is reflected as high-low patterns on the d-curve with a relatively uniform range that enables the computer to catch them with thresholds. The values of thresholds may be arguable, but no more so than the "right" number of clusters that exist in a given data set. For example, some see only one single cluster in Figure 1 while we see three. Our experiments show that the computer is more sensitive to high-low patterns on the d-curve than human eyes to patterns in 2D gray- level images.

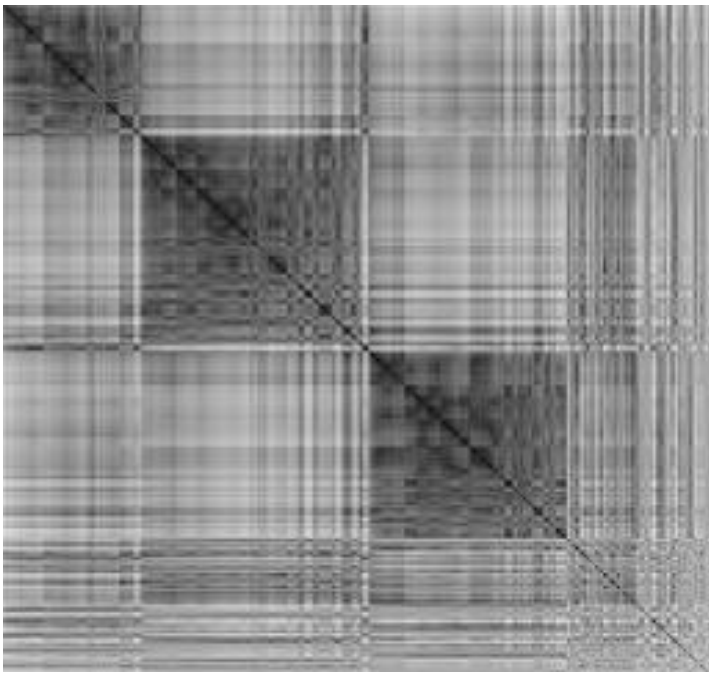


Figure 1: VAT Analysis image

VIII. CONCLUSION

Given a data set, the ideal scenario would be to have a given set of criteria to choose a proper clustering algorithm to apply. Choosing a clustering algorithm, however, can be a difficult task. Most of the algorithms generally assume some implicit structure in the data set. One of the most important elements is the nature of the data and the nature of the desired cluster. Another issue to keep in mind is the kind of input and tools that the algorithm requires. This report has a proposal of a new hierarchical clustering algorithm based on the overlap rate for cluster merging. The experience in general data sets and a document set indicates that the new method can decrease the time cost, reduce the space complexity and improve the accuracy of clustering. Specially, in the document clustering, the newly proposed algorithm measuring result show great advantages. The hierarchical document clustering algorithm provides a natural way of distinguishing clusters and implementing the basic requirement of between-cluster dissimilarity.

REFERENCES

[1]. DucThang Nguyen, Lihui Chen and CheeKeong Chan, "Clustering with Multi-Viewpoint based Similarity Measure", IEEE Transactions on Knowledge and Data Engineering, 2011.

[2]. SX. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," Knowl.Inf. Syst., vol. 14, no. 1, pp. 1–37, 2007.

[3]. "An Efficient Formulation of the Improved Visual Assessment of Cluster Tendency (iVAT) Algorithm" Timothy C. Havens, Senior Member, IEEE, and James C. Bezdek, Fellow, IEEE.

[4]. Haojun sun, zhihuiliu, lingjunkong, A Document Clustering Method Based on Hierarchical Algorithm with Model Clustering, 22nd international conference on advanced information networking and applications.

[5]. Shengrui Wang and Haojun Sun. Measuring overlap-Rate for Cluster Merging in a Hierarchical Approach to Color Image Segmentation. International Journal of Fuzzy Systems, Vol.6, No.3, September 2004.