ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol. 3 , No.1, Pages : 288 – 291 (2014)
Special Issue of ICETETS 2014 - Held on 24-25 February, 2014 in Malla Reddy Institute of Engineering and Technology, Secunderabad– 14, AP, India

A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification

[1]B.Archana, [2]Y.Prathima

[1] I M.Tech C.S.E Malla Reddy College of Engineering &Technology, Hyderabad.
[2] AssociateProfessor C.S.E Malla Reddy Collegeof Engineering&Technology, Hyd.

Abstract: All clustering methods have to assume some cluster relationship among the data objects that they are applied on. Similarity between a pair of objects can be defined either explicitly or implicitly. In this paper, we introduce a novel multiviewpoint-based similarity measure and two related clustering methods. The major difference between a traditional dissimilarity/similarity measure and ours is that the former uses only a single viewpoint, which is the origin, while the latter utilizes many different viewpoints, which are objects assumed to not be in the same cluster with the two objects being measured. Using multiple viewpoints, more informative assessment of similarity could be achieved. Theoretical analysis and empirical study are conducted to support this claim. Two criterion functions for document clustering are proposed based on this new measure. We compare them with several well-known clustering algorithms that use other popular similarity measures on various document collections to verify the advantages of our proposal.

Key word: Dimension reduction, feature clustering, text classification, self-constructing clustering, multi-label document classification.

INTRODUCTION

CLUSTERING is one of the most interesting and important topics in data mining. The aim of clustering is to find intrinsic structures in data, and organize them into meaningful subgroups for further study and analysis. There have been many clustering algorithms published every year. They can be proposed for very distinct research fields, and developed using totally different techniques and approaches.Nevertheless, according to a recent study ,more than half a century after it was introduced, the simple algorithm k-means still remains as one of the top 10 datamining algorithms nowadays. It is the most frequently used partitional clustering algorithm in practice. Another recent scientific discussion [2] states that k-means is the favorite algorithm that practitioners in the related fields choose to use. Needless to mention, k-means has more than a few basic drawbacks, such as sensitiveness to initialization and to cluster size, and its performance can be worse than other state-of-the-art algorithms in many domains. In spite of that,its simplicity, understandability, and scalability are the reasons for its tremendous popularity. An algorithm with adequate performance and usability in most of application scenarios could be preferable to one with better performance in some cases but limited usage due to high complexity.While offering reasonable results, k-means is fast and easy to combine with other methods in larger systems.A common approach to the clustering problem is to treat it as an optimization process. An optimal partition is found by optimizing a particular function of similarity (or distance) among data. Basically, there is an implicit assumption that the true intrinsic structure of data could be correctly described by the similarity formula defined and embedded in the clustering criterion function. Hence,effectiveness of clustering algorithms under this approach depends on the appropriateness of the similarity measure to the data at hand. For instance, the original k-means has sum-of-squared-error objective function that uses euclidean distance. In a very sparse and high-dimensional domain like text documents, spherical k-means, which uses cosine similarity (CS) instead of euclidean distance as the measure,is deemed to be more suitable .In, Banerjee et al. showed that euclidean distance was indeed one particular form of a class of distance measures called Bregman divergences. They proposed Bregman hardclustering algorithm, in which any kind of the Bregman divergences could be applied. Kullback-Leibler divergence was a special case of Bregman divergences that was said to give good clustering results on document data sets. Kullback-Leibler divergence is a good example of nonsymmetric measure. Also on the topic of capturing dissimilarity in data,Pakalska et al. found that the discriminative power of some distance measures could increase when their non-Euclidean and nonmetric attributes were increased. They concluded that noneuclidean and

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol. 3 , No.1, Pages : 288 – 291 (2014)
Special Issue of ICETETS 2014 - Held on 24-25 February, 2014 in Malla Reddy Institute of Engineering and Technology, Secunderabad– 14, AP, India

nonmetric measures could be informative for statistical learning of data. In , Pelillo even argued that the symmetry and nonnegativity assumption of similarity measures was actually a limitation of current state-of-the-art clustering approaches. Simultaneously, clustering still requires more robust dissimilarity or similarity measures; recent works such as [8] illustrate this need.The work in this paper is motivated by investigations from the above and similar research findings. It appears to us that the nature of similarity measure plays a very important role in the success or failure of a clustering method. Our first objective is to derive a novel method for measuring similarity between data objects in sparse and high-dimensional domain, particularly text documents. From the proposed similarity measure, we then formulate new clustering criterion functions and introduce their respective clustering algorithms, which are fast and scalable like k-means, but are also capable of providing high-quality and consistent performance.

Background and Related work:

Feature Reduction : In general, there are two ways of doing feature reduction, feature selection, and feature extraction. By feature selection approaches, a new feature set W'={w'1',w2',…,wk'} is obtained, which is a subset of the original feature set W. Then W0 is used as inputs for classification tasks. Information Gain (IG) is frequently employed in the feature selection approach [10]. It measures the reduced uncertainty by an information-theoretic measure and gives each word a weight.

Feature Clustering : Feature clustering is an efficient approach for feature reduction [25], [29], which groups all features into some clusters, where features in a cluster are similar to each other. The feature clustering methods proposed in [24], [25], [27], [29] are "hard" clustering methods, where each word of the original features belongs to exactly one word cluster. Therefore each word contributes to the synthesis of only one new feature. Each new feature is obtained by summing up the words belonging to one cluster. Let D be the matrix consisting of all the original documents with m features and D' be the matrix consisting of the converted documents with new k features. The new feature set W'={w1',w2',…,wk'} corresponds to a partition {w1',w2',…,wk'} of the original feature set W, i.e., Wt∩Wq =0, where 1 ≤q, t ≤ k and t != q. Note that a cluster corresponds to an element in the partition. Then, the tth feature value of the converted document d'iis calculated as follows: d'it = ∑ wj Є wt dij which is a linear sum of the feature values in w

Self-Constructing Clustering : Our clustering algorithm is an incremental, self-constructing learning approach. Word patterns are considered one by one. The user does not need to have any idea about the number of clusters in advance. No clusters exist at the beginning, and clusters can be created if necessary. For each word pattern, the similarity of this word pattern to each existing cluster is calculated to decide whether it is combined into an existing cluster or a new cluster is created. Once a new cluster is created, the corresponding membership function should be initialized. On the contrary, when the word pattern is combined into an existing cluster, the membership function of that cluster should be updated accordingly.

Let k be the number of currently existing clusters. The clusters are G1, G2, . . .,Gk, respectively. Each cluster Gj has mean mj=mj1,mj2,…..mjn and deviation σj= < σj1 σj2 ……..j σmj >. Let Sj

be the size of cluster Gj. Initially, we have k =0. So, no clusters exist at the beginning. For each word pattern

Xi=< Xi1, Xi2 ,……. Xin >

we calculate the similarity of xi to each existing clusters, i.e., for 1 ≤ j ≤ k. We say that xi passes the similarity test on cluster Gj if μGJ(Xi) >ρ where p, 0≤p≤1, is a predefined threshold. If the user intends to have larger clusters, then he/she can give a smaller threshold. Otherwise, a bigger threshold can be given. As the threshold increases, the number of clusters also increases. Note that, as usual, the power in above function is 2 [34], [35]. Its value has an effect on the number of clusters obtained. A larger value will make the boundaries of the Gaussian function sharper, and more clusters will be obtained for a given threshold.

Feature Extraction : Word patterns have been grouped into clusters, and words in the feature vector W are also clustered accordingly. For one cluster, we have one extracted feature.

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol. 3 , No.1, Pages : 288 – 291 (2014)
Special Issue of ICETETS 2014 - Held on 24-25 February, 2014 in Malla Reddy Institute of Engineering and Technology, Secunderabad– 14, AP, India

Since we have k clusters, we have k extracted features. The elements of T are derived based on the obtained clusters, and feature extraction will be done. We propose three weighting approaches: hard, soft, and mixed. In the hard-weighting approach, each word is only allowed to belong to a cluster, and so it only contributes to a new extracted feature. In the soft-weighting approach, each word is allowed to contribute to all new extracted features, with the degrees depending on the values of the membership functions. The mixed-weighting approach is a combination of the hard-weighting approach and the soft-weighting approach.

Text Classification :

Given a set D of training documents, text classification can be done as follows: Get the training document set and specify the similarity threshold ρ. Assume that k clusters are obtained for the words in the feature vector W. Then find the weighting matrix T and convert D to D`. Using D` as training data, a classifier based on support vector machines (SVM) is built. Note that any classifying technique other than SVM can be applied. Joachims [36] showed that SVM is better than other methods for text categorization. SVM is a kernel method, which finds the maximum margin hyperplane in feature space separating the images of the training patterns into two groups [37], [38], [39]. To make the method more flexible and robust, some patterns need not be correctly classified by the hyperplane, but the misclassified patterns should be penalized. Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. Weka is a collection of machine learning algorithms for data mining tasks. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Using weak we classify the text

Our Method:

There are some issues pertinent to most of the existing feature clustering methods. First, the parameter k indicating the desired number of extracted features, has to be specified in advance.This give burden to user and-error has to be done until the appropriate number of extracted features is found. Second, when calculating similarities, the variance of the underlying cluster is not considered. Intuitively, the distribution of the data in a cluster is an important factor in the calculation of similarity. Third, all words in a cluster have the same degree of contribution to the resulting feature clusteringalgorithm is proposed to deal with these issues. Suppose, we are given a document set D of n documents d1, d2, . . . , dn, together with the feature vector W of m words w1, w2, . . . , wm and p classes c1, c2, . . . , cp, as specified. We construct one word pattern for each word in W. For word wi, its word pattern xi is defined, similarly as in [27], byfor $i \leq j \leq p$.

Note that dqi indicates the number of occurrences of wi in document dq

Also, ∂qj is defined as either 1 or 0

Therefore, we have m word patterns in total.

Proposed System: The work in this paper is motivated by investigations from the above and similar research findings. It appears to us that the nature of similarity measure plays a very important role in the success or failure of a clustering method. Our first objective is to derive a novel method for measuring similarity between data objects in sparse and high-dimensional domain, particularly text documents. From the proposed similarity measure, we then formulate new clustering criterion functions and introduce their respective clustering algorithms, which are fast and scalable like k-means, but are also capable of providing high-quality and consistent performance.

Conclusion: In this paper we propose a Multiviewpoint-based Similarity measuring method, named MVS. Theoretical analysis and empirical examples show that MVS is potentially more suitable for text documents than the popular cosine similarity. Based on MVS, two criterion functions, IR and IV , and their respective clustering algorithms, MVSC-IR and MVSC-IV , have been introduced. Compared with other state-of-the-art clustering methods that use different types of similarity measure, on a large number of document data sets and under different evaluation metrics, the proposed algorithms show that they could provide significantly improved clustering performance. The key contribution of this paper is the fundamental concept of similarity measure from multiple viewpoints. Future methods could make use of the same principle, but define alternative forms for the relative similarity in (10), or do not use average but have other methods to combine the relative similarities according to the different viewpoints. Besides, this paper focuses on partitioned clustering of documents. In the

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol. 3 , No.1, Pages : 288 – 291 (2014)
Special Issue of ICETETS 2014 - Held on 24-25 February, 2014 in Malla Reddy Institute of Engineering and Technology, Secunderabad– 14, AP, India

future, it would also be possible to apply the proposed criterion functions for hierarchical clustering algorithms. Finally, we have shown the application of MVS and its clustering algorithms for text data. It would be interesting to explore how they work on other types of sparse and high-dimensional data.

References:

1. X. Wu, V. Kumar, J.R. 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," Knowledge Information Systems, vol. 14, no. 1, pp. 1-37, 2007.
2. I. Guyon, U.V. Luxburg, and R.C. Williamson, "Clustering:Science or Art?," Proc. NIPS Workshop Clustering Theory, 2009.
3. Dhillon and D. Modha, "Concept Decompositions for Large Sparse Text Data Using Clustering," Machine Learning, vol. 42,nos. 1/2, pp. 143-175, Jan. 2001.
4. S. Zhong, "Efficient Online Spherical K-means Clustering," Proc.IEEE Int'l Joint Conf. Neural Networks (IJCNN), pp. 3180-3185, 2005.
5. A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh, "Clustering with Bregman Divergences," J. Machine Learning Research, vol. 6,pp. 1705-1749, Oct. 2005.
6. E. Pekalska, A. Harol, R.P.W. Duin, B. Spillmann, and H. Bunke,"Non-Euclidean or Non-Metric Measures Can Be Informative,"Structural, Syntactic, and Statistical Pattern Recognition, vol. 4109,pp. 871-880, 2006.
7. M. Pelillo, "What Is a Cluster? Perspectives from Game Theory,"Proc. NIPS Workshop Clustering Theory, 2009.
8. D. Lee and J. Lee, "Dynamic Dissimilarity Measure for Support Based Clustering," IEEE Trans. Knowledge and Data Eng., vol. 22,no. 6, pp. 900-905, June 2010.
9. A. Banerjee, I. Dhillon, J. Ghosh, and S. Sra, "Clustering on the Unit Hypersphere Using Von Mises-Fisher Distributions,"J. Machine Learning Research, vol. 6, pp. 1345-1382, Sept. 2005.
10. W. Xu, X. Liu, and Y. Gong, "Document Clustering Based on Non-Negative Matrix Factorization," Proc. 26th Ann. Int'l ACM SIGIR Conf. Research and Development in Informaion Retrieval, pp. 267-273,2003.
11. I.S. Dhillon, S. Mallela, and D.S. Modha, "Information-Theoretic Co-Clustering," Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 89-98, 2003.
12. C.D. Manning, P. Raghavan, and H. Schu¨ tze, An Introduction to Information Retrieval. Cambridge Univ. Press, 2009.
13. C. Ding, X. He, H. Zha, M. Gu, and H. Simon, "A Min-Max Cut Algorithm for Graph Partitioning and Data Clustering," Proc.IEEE Int'l Conf. Data Mining (ICDM), pp. 107-114, 2001.
14. H. Zha, X. He, C. Ding, H. Simon, and M. Gu, "Spectral Relaxation for K-Means Clustering," Proc. Neural Info. Processing Systems (NIPS), pp. 1057-1064, 2001.
15. J. Shi and J. Malik, "Normalized Cuts and Image Segmentation,"IEEE Trans. Pattern Analysis Machine Intelligence, vol. 22, no. 8,pp. 888-905, Aug. 2000.
16. I.S. Dhillon, "Co-Clustering Documents and Words Using Bipartite Spectral Graph Partitioning," Proc. Seventh ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD),pp. 269-274, 2001.
17. Y. Gong and W. Xu, Machine Learning for Multimedia Content Analysis. Springer-Verlag, 2007.
18. Y. Zhao and G. Karypis, "Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering,"Machine Learning, vol. 55, no. 3, pp. 311-331, June 2004.
19. G. Karypis, "CLUTO a Clustering Toolkit," technical report, Dept.of Computer Science, Univ. of Minnesota, http://glaros.dtc.umn.edu/~gkhome/views/cluto, 2003.
20. A. Strehl, J. Ghosh, and R. Mooney, "Impact of Similarity Measures on Web-Page Clustering," Proc. 17th Nat'l Conf. Artificial Intelligence: Workshop of Artificial Intelligence for Web Search (AAAI),pp. 58-64, July 2000.