Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 122 (2017) 276–283

Procedia Computer Science

www.elsevier.com/locate/procedia

Information Technology and Quantitative Management (ITQM 2017)

An Entropy based Method for Overlapping Subspace Clustering

Charu Puri[1], Naveen Kumar[1]

[a]*Dept. of Computer Science, University of Delhi, New Delhi-10065, INDIA*

Abstract

We present subspace clustering method based on entropy and modification of Gustafson-Kessel clustering. The proposed method associates with each cluster its center, variance co-variance matrix and a weight matrix. It represents clusters and their features through gradation in membership, and hence reflects a realistic representation of clusters. Evaluation of experiments on data from UCI ae well as text with the comparative methods shows better results of proposed method.

Peer-review under responsibility of the scientific committee of the 5th International Conference on Information Technology and Quantitative Management, ITQM 2017.

Keywords: Entropy; subspace; document clustering.

1. Introduction

Large document collection often contains vast amount of important scientific data which requires an automatic method for organizing documents. A collection of documents may contain multiple topics and hence, may belong to multiple categories. Also, text data has a large number of dimensions and is highly correlated. Thus, the primary issues for document categorization are hard clustering, curse of dimensionality and correlation in data. Class of fuzzy co-clustering algorithms generate descriptive clusters in high dimension and discover clusters with overlap i.e. it enables a document to be in multiple clusters. Subspace clustering finds cluster with fewer dimensions. Soft subspace clustering finds cluster in the entire space with continuously dimension grading. In high dimensional text data, only a subset of dimensions is relevant for each cluster [14] as varying key words categorize the feature space of each document cluster. Also, a document may belong to multiple categories and multiple specific set of key words may belong to a document cluster. Some of the techniques that have been proposed for variable weighting subspace clustering for document categorization are LAC [6], FWKM [15], and EWKM [14]. These algorithms discover soft subspaces for soft partitioning of document categories.

We propose an algorithm for entropy based overlapping subspace clustering and demonstrate the performance on document categorization. The proposed algorithm associates a center, variance co-variance matrix and a weight matrix with each of the clusters. It represents cluster and their features through a gradation in memberships, and hence reflects a realistic representation of document clusters.

It takes into account the correlation between the features at two stages: (i) during the clustering (ii) while finding the subspaces. The proposed algorithm has the feature of local adaptation of Mahalanobis distance metric

*Charu Puri Tel.: +91-971-755-6021;
E-mail address: cpuri.cs.du@gmail.com, nk.cs.du@gmail.com.

1877-0509
Peer-review under responsibility of the scientific committee of the 5th International Conference on Information Technology and Quantitative Management, ITQM 2017.
10.1016/j.procs.2017.11.370

used as the similarity measure which takes into account the co-variance of the data for clustering. For finding subspaces locally for each cluster the weight matrix assigns the degree of relevance of features and is inversely proportional to ratio of dispersion. In order to effectively capture the correlation in the data, we incorporated mahalanobis distance metric in our algorithm. The primary contributions of this paper are:

We present an entropy based gustafon kessel subspace clustering algorithm with locally adaptive distance metric. Entropy based subspace clustering algorithm finds meaningful subspaces and clusters with fewer dimensions from text. The contributions of EGKS clustering algorithm are:

1. A correlation-based subspace clustering objective function is proposed which minimize cluster error.
2. Finds multiple categories with multiple set of keywords.
3. Captures the correlation in data at level of clustering as well to find subspaces.
4. Extracts arbitrarily positioned overlapping subspace clusters simultaneously.
5. Has feature of local adaptation of distance metric and weight matrix.
6. Ability to identify subspace clusters to categorize text data.

The paper is structured in the following manner. Section 2 presents the existing research literature close to the proposed method on subspace clustering. In Section 3 the proposed Entropy based Gustafson Kessel Subspace algorithm is presented. Section 4 presents the results of experiments on both UCI as well as text datasets along with the experimental comparisons with other comparative methods. Finally conclusion is presented in Section 5.

2. Related Work

$Kmeans$ algorithm is a widely used partitional clustering algorithm which gives equal weights to each of the variables in the clustering process. However, in many real-world applications where the true clusters tend to appear only in a subset of dimensions it is more desirable to generalize the algorithm to incorporate feature weights. Huang et al. [12] presented $kmeans$ type method $Wkmeans$ which automatically assigns weights to the variables on the basis of their relevance to the clustering solution. $Wkmeans$ modifies the basic k-means by including the weights to the features as well an update mechanism for the weight variables on the basis of current data partitioning.

Wang et al. have presented fuzzy c-means algorithm with varying weights called Fuzzy W-k means. It creates partition of soft clusters [20]. They have modified K-means to create soft partitions of the clusters and introduced the parameters α and β as fuzzification parameters. They computes memberships of data points and dimensions simultaneously.

Jing et al. have presented clustering method FW-KMeans [15] for finding subspaces. It assigns weight to dimensions specific to clusters. Thus a dimension may be assigned a high weight for a cluster (implying high relevance) and a low weight for another cluster (implying low relevance).

Subsequently Jing et al. have modified the FW-KMeans for clustering textual data. In text clustering, large weight identifies a key. In text data many words may not be a part of documents of certain classes. This adds sparsity to the text data. However, cases where a word may not appear in any document of the corresponding cluster or a word may appear in every document with equal frequency results in zero dispersion which makes ω_{ir} infinite. For text data sets FW-Kmeans algorithm becomes unsolvable as weights for these terms become infinite.

Jing et al. have proposed an extension of k-means i.e. Entropy Weighting K-Means. EWKM is a soft subspace finding approach [14]. They have extended k-means to compute a membership of every dimension in each cluster. These weight values identify the subset of relevant dimensions which categorize different every cluster. The weight ω_{ir} of a feature measures the contribution of that dimension in forming the cluster where as the entropy of the feature measures the certainty of a feature in the cluster identification.

In this paper, we present an entropy and Gustafson Kessel based subspace finding method for document categorization. It discovers multiple partitions of document in variable weighting subspaces of words. It finds overlapping clusters of documents in the overlapping subspaces i.e. it clusters documents and words simultaneously where word membership helps in generating more useful descriptions of document clusters. The proposed algorithm takes into account the correlation in the data calculated using the inverse of variance co-variance matrix.