1. Introduction

High-dimensional data is a phenomenon in real-world data mining applications. Text data is a typical example. In text mining, a text document is viewed as a set of pairs < ti; fi >, where ti is a term or word, and fi is a measure of ti, for example, the frequency of ti in the document. The total number of unique terms in a text data set represents the number of dimensions, which is usually in the thousands. High-dimensional data occurs in business as well. In retail companies, for example, for effective supplier relationship management (SRM), suppliers are often categorized in groups according to their business behaviors. The supplier's behavior data is high dimensional because thousands of attributes are used to describe the supplier's behaviors, including product items, ordered amounts, order frequencies, product quality, and so forth. Sparsity is an accompanying phenomenon of high dimensional data. Clearly, clustering of high-dimensional sparse data requires special treatment. This type of clustering methods is referred to as subspace clustering, aiming at finding clusters from subspaces of data instead of the entire data space. In a subspace clustering, each cluster is a set of objects identified by a subset of dimensions and different clusters are represented in different subsets of dimensions. According to the ways that the subspaces of clusters are determined, subspace clustering methods can be divided into two types. The first type is to find out the exact subspaces of different clusters called as hard subspace clustering. The second type is to cluster data objects in the entire data space but assign different weighting values to different dimensions of clusters in the clustering process, based on the importance of the dimensions in identifying the corresponding clusters .We call these methods soft subspace clustering. In this paper, we present a new k-means type algorithm for soft subspace clustering of large high-dimensional sparse data. We consider that different dimensions make different contributions to the identification of objects in a cluster.

2. Entropy Weighting K-Means

In this section, we present a new k-means type algorithm for soft subspace clustering of high-dimensional sparse data. In the new algorithm, we consider that the weight of a dimension in a cluster represents the probability of contribution of that dimension in forming the cluster. The entropy of the dimension weights represents the certainty of dimensions in the identification of a cluster. Therefore, we will modify the objective function by adding the weight entropy term to it so that we can simultaneously minimize the within cluster dispersion and maximize the negative weight entropy to stimulate more dimensions to contribute to the identification of clusters. In this way, we can avoid the problem of identifying clusters by few dimensions in sparse data. The objective function is written as follows:

$$F(W,Z,\Lambda)=\sum_{i=1}^{k}\left[\sum_{j=1}^{n}\sum_{i=1}^{m}w_{ij}\lambda_{ti}(z_{ti}-x_{ji})^2+\gamma\sum_{i=1}^{m}\lambda_{ti}\log\lambda_{ti}\right]$$

Subject to

$$\begin{cases} \sum_{l=1}^{k} w_{lj} = 1, & 1 \le j \le n, \quad 1 \le l \le k, \quad w_{lj} \in \{0,1\} \\ \sum_{i=1}^{m} \lambda_{li} = 1, & 1 \le l \le k, \quad 1 \le i \le m, \quad 0 \le \lambda_{li} \le 1. \end{cases}$$

The first term in is the sum of the within cluster dispersions, and the second term the negative weight entropy. The positive parameter controls the strength of the incentive for clustering on more dimensions. Next, we present the entropy weighting k-means algorithm (EWKM) to solve the above minimization problem.

The usual method toward optimization of F is to use the partial optimization for ʌ, Z and W. In this method, we first fix Z and ʌ and minimize the reduced F with respect to W. Then, we fix W and ʌ and minimize the reduced F with respect to Z. afterward; we fix W and Z and minimize the reduced F to solve ʌ We can extend the standard k-means clustering process to minimize F by adding an additional step in each iteration to compute weights ʌ for each cluster.

$$\begin{cases} w_{lj} = 1, & \text{if } \sum_{i=1}^{m} \lambda_{li}(z_{li} - x_{ji})^2 \le \sum_{i=1}^{m} \lambda_{ri}(z_{ri} - x_{ji})^2 \\ & \text{for } 1 \le r \le k, \\ w_{lj} = 0, & \text{otherwise.} \end{cases}$$

wij =1 means that the jth object is assigned to the lth cluster. If the distances between an object and two cluster centers are equal, the object is arbitrarily assigned to the cluster with the smaller cluster index number. Given W and are fixed, Z is updated as:

$$z_{li} = \frac{\sum_{j=1}^{n} w_{lj} x_{ji}}{\sum_{j=1}^{n} w_{lj}} \quad \text{for } 1 \le l \le k \text{ and } 1 \le i \le m.$$

The objective of this analysis was to help a food retail company to categorize its suppliers according to suppliers' business behaviors. Supplier categorization refers to the process of dividing suppliers of an organization into different groups according to the characteristics so that each group of suppliers can be managed differently within the organization. It is an important step in SRM for creating better supplier management strategies to reduce the product sourcing risk and costs and improve business performance.

3. Conclusion

In this paper, we have discussed about an enhanced k-means type algorithm for high-dimensional data where we simultaneously minimize the cluster dispersion and maximize the negative weight entropy in the clustering process. Because this clustering process awards more dimensions to make contributions to identification of each cluster, the problem of identifying clusters by few sparse dimensions can be avoided. As such, the sparsity problem of high-dimensional data is tackled.