Pattern Recognition 48 (2015) 3688–3702

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Stratified feature sampling method for ensemble clustering of high dimensional data

Liping Jing [a,*], Kuang Tian [a], Joshua Z. Huang [b]

[a] Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing, China
[b] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China

ARTICLE INFO

Article history:
Received 18 June 2013
Received in revised form
19 August 2014
Accepted 2 May 2015
Available online 13 May 2015

Keywords:
Stratified sampling
Ensemble clustering
High dimensional data
Consensus function

ABSTRACT

High dimensional data with thousands of features present a big challenge to current clustering algorithms. Sparsity, noise and correlation of features are common characteristics of such data. Another common phenomenon is that clusters in such high dimensional data often exist in different subspaces. Ensemble clustering is emerging as a prominent technique for improving robustness, stability and accuracy of high dimensional data clustering. In this paper, we propose a stratified sampling method for generating subspace component data sets in ensemble clustering of high dimensional data. Instead of randomly sampling a subset of features for each component data set, in this method we first cluster the features of high dimensional data into a few feature groups called feature strata. Using stratified sampling, we randomly sample some features from each feature stratum and merge the sampled features from different feature strata to generate a component data set. In this way, the component data sets have better representations of the clustering structure in the original data set. Comparing with random sampling and random projection methods in synthetic data analysis, the component clustering by stratified sampling has demonstrated that the average clustering accuracy was increased without sacrificing clustering diversity. We carried out a series of experiments on eight real world data sets from microarray, text and image domains to evaluate ensemble clustering methods using three subspace component data generation methods and four consensus functions. The experimental results consistently showed that the stratified sampling method produced the best ensemble clustering results in all data sets. The ensemble clustering with stratified sampling also outperformed three other ensemble clustering methods which generate component clusters from the entire space of the original data.

1. Introduction

High dimensionality is a big challenge in cluster analysis [24]. Sparsity, noise, correlation and uninformativeness of features are common characteristics of high dimensional data in real applications. Another common phenomenon is that clusters in such data often exist in different subspaces. To effectively cluster high dimensional data, researchers have proposed various clustering methods, including subspace clustering methods [1,2,32,7,20] (see the survey reports [30,24,36]. However, most algorithms fail in obtaining good clustering performance [34,33,8]. The ensemble clustering methods are promising to solve this problem.

Ensemble clustering is an emerging clustering technique that integrates multiple clusterings generated from samples of a given data set into a single clustering with a result which is usually much better than the results of individual clusterings on the data set [34,16]. Ensemble clustering is more effective in clustering high dimensional complex data than the clustering methods that produce single clustering results because the ensemble of clusterings created from a high dimensional data set is more stable and robust and usually with an accuracy higher than any of the individual component clusterings. Due to this advantage, ensemble clustering becomes attractive in clustering high dimensional data such as text data [8], microarray data [39] and image data [15,18].

Given a data set, the process of ensemble clustering is conducted in two steps, generating a set of individual component clusterings from the data set and integrating the component clusterings into a clustering ensemble. The quality of the final clustering ensemble is determined by the methods to carry out these two steps. Different methods result in different ensemble clustering algorithms.

Generation of component clusterings is mainly focused on the diversity of component clusterings [25]. Three strategies are commonly used in this step [18]: (a) using one clustering algorithm with different parameter settings such as initial cluster centers, the number of clusters, the cluster density threshold,

* Corresponding author.
E-mail addresses: lpjing@bjtu.edu.cn (L. Jing), zx.huang@szu.edu.cn (J.Z. Huang).

http://dx.doi.org/10.1016/j.patcog.2015.05.006
0031-3203/

etc., to generate homogeneous clusterings [11,13,26,17]; (b) using different clustering algorithms to cluster the same data set to create heterogeneous clusterings [3,28]; (c) sampling the given data set to generate different component data sets and using a clustering algorithm to cluster them to produce component clusterings [34,9,4,35,39,8]. The ensemble clustering is created by means of a consensus function to integrate multiple component clusterings into one final clustering. The consensus functions used in this step include the direct method, the feature-based method [4], the graph-based approach [34,33] and the pairwise similarity approach [18]. The main objective of integration is to produce an ensemble clustering with a higher accuracy than the accuracies of component clusterings.

For high dimensional data, two randomization methods are employed to generate low dimensional component data sets. One is to randomly sample different subsets of features to generate subspace component data sets [34,4] and the other is to project the given high dimensional data into low dimensional component data sets by randomly generated projection matrices [9]. The major advantage of these methods is that they can generate diverse component clusterings. A serious shortcoming is that low dimensional component data sets are severely deviated from the original given data set, which results in a strong discrepancy in the clustering structures between the component data sets and the original data. As a consequence, the strength or the quality of component clusterings is significantly affected.

Stratified sampling is a well-known sampling technique that can adequately capture the population characteristics [12]. Stratified sampling is conducted in two steps, dividing the whole population into sub-populations called strata, and then applying ransom sampling to each stratum to select the representative samples. Stratified sampling is widely used in many areas, such as estimating software reliability [31], approximating query processing [6], traffic data analysis [10], web mining [29], and building random forest for classification [38].

In this paper, we propose to use stratified sampling to sample subspace features of component data sets for ensemble clustering of high dimensional data. Due to the existence of correlations among features in high dimensional data, the clustering structure of data is categorized by the groups of correlated features. Because of this, we can cluster the original data on features into feature clusters and use stratified sampling to randomly sample features separately from feature clusters to form the feature subsets of component data sets. The number of features sampled from each feature cluster is proportional to the size of the feature cluster. Since the characteristics of all feature clusters are represented in the subspace features of component data sets, the discrepancy of the clustering structures between the component data sets and the original data set is reduced and the quality of component clusterings is improved without sacrificing the diversity of component clusterings too much. As a result, the robustness, stability and accuracy of the clustering ensemble are increased. To our knowledge, this is the first work to integrate stratified sampling into ensemble clustering.

We have conducted a series of experiments on both synthetic data and real life data to evaluate the stratified sampling method in ensemble clustering from different angles. We compared the results of ensemble clusterings produced with random sampling, random projection and stratified sampling. We used four consensus functions to generate clustering ensembles, i.e., the similarity-based consensus function, the hyper graph-based consensus function, the meta cluster-based consensus function and the link-based consensus function. The experimental results showed consistent improvement in accuracy of clustering ensembles produced from the stratified sampling method with all four consensus functions in comparison with the random sampling and random projection methods. We also analyzed the size of feature subsets and the number of feature clusters on the influence of the clustering ensemble accuracy. The results showed that a good clustering accuracy could be obtained with a small subspace of features in each component data set and a small number of feature clusters. This analysis indicates that the stratified sampling method in ensemble clustering is easy to use in practice.

The rest of this paper is organized as follows. In Section 2, we present the motivation of this work through analysis of synthetic data sets. In Section 3, we present the stratified sampling method for generating subspace component data sets in ensemble clustering and show its statistical advantage in sampling subspace data with better representation of the whole data set. In Section 4, we present four consensus functions that were used in this work to generate clustering ensembles. In Section 5, the experiments on eight real world high dimensional data sets are presented. The results of ensemble clusterings from 12 ensemble clustering algorithms are compared. The parameters on the number of feature strata and the sampling rate are investigated. Finally, we conclude this work and give a future research direction in Section 6.

2. Motivation

In this section, we use synthetic data analysis results to illustrate the problems of the random sampling and random projection methods in generation of low dimensional component data sets for ensemble clustering. We created six synthetic data sets, each with 100 features and three clusters. Each cluster consisted of 50 points generated from a multivariate normal distribution. The means of three clusters in each key dimension were set as $\mu_1 = 1$, $\mu_2 = 5$, $\mu_3 = 3$ respectively and the same variance of $\sigma = 1$ was used for all clusters. The three clusters were generated independently and merged into one data set. Afterwards, a number of noise features with uniform distribution between 0 and 1 were added to the data set to replace the same number of features with cluster distributions. By adding different percentages of noise features to the data set, we created six data sets which are shown in Fig. 1 where figures from (a) to (f) have noise features with percentages of 0.05, 0.1, 0.2, 0.5, 0.6, 0.7 respectively. The left three sections in these figures show the three clusters in dark ($\mu_1 = 1$), light ($\mu_2 = 5$) and gray ($\mu_3 = 3$) colors and the right section is noise. Introduction of noise features to the data sets increases the difficulty of clustering. The more the noise features were included in a data set, the more difficult to cluster the data.

For each data set X in Fig. 1, we used both random sampling and random projection methods to generate 200 low dimensional component data sets, 100 data sets with each method. The random sampling method randomly selected p features from X to generate a low dimensional component data set. Because any feature has the same chance to be selected in every random sampling, the component data sets may have partial common features. The random projection method projects X into p dimensions by multiplying a random matrix $P_{d\times p}$ to X where $d = 100$ is the number of dimensions in X. In practice, $p = q \times d$ where q is a sampling rate. The values of the random projection matrix P were randomly generated from a normal distribution.

Note that there may be common or overlapping features among the component data sets because any feature has the same chance of selection in every random sampling.

For the 100 low dimensional component data sets from each sampling method, we used the k-means algorithm to cluster each data set into 3 clusters and computed the clustering accuracy of the result. We divided the 100 clustering results into 6 accuracy groups of (0,0.5], (0.5,0.6], (0.6,0.7], (0.7,0.8], (0.8,0.9], (0.9,1]).