Information Sciences 267 (2014) 16–34

Contents lists available at ScienceDirect

Information Sciences

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

Probabilistic cluster structure ensemble

Zhiwen Yu [a,*], Le Li [a], Hau-San Wong [b], Jane You [c], Guoqiang Han [a], Yunjun Gao [d], Guoxian Yu [e]

[a] *School of Computer Science and Engineering, South China University of Technology, China*
[b] *Department of Computer Science, City University of Hong Kong, Hong Kong*
[c] *Department of Computing, Hong Kong Polytechnic University, Hong Kong*
[d] *College of Computer Science, Zhejiang University, China*
[e] *College of Computer and Information Science, Southwest University, Chongqing, China*

ARTICLE INFO

Article history:
Received 7 April 2013
Received in revised form 12 November 2013
Accepted 12 January 2014
Available online 24 January 2014

Keywords:
Cluster ensemble
Gaussian mixture model
Normalized cut
Structure ensemble

ABSTRACT

Cluster structure ensemble focuses on integrating multiple cluster structures extracted from different datasets into a unified cluster structure, instead of aligning the individual labels from the clustering solutions derived from multiple homogenous datasets in the cluster ensemble framework. In this article, we design a novel probabilistic cluster structure ensemble framework, referred to as Gaussian mixture model based cluster structure ensemble framework (GMMSE), to identify the most representative cluster structure from the dataset. Specifically, GMMSE first applies the bagging approach to produce a set of variant datasets. Then, a set of Gaussian mixture models are used to capture the underlying cluster structures of the datasets. GMMSE applies *K*-means to initialize the values of the parameters of the Gaussian mixture model, and adopts the Expectation Maximization approach (EM) to estimate the parameter values of the model. Next, the components of the Gaussian mixture models are viewed as new data samples which are used to construct the representative matrix capturing the relationships among components. The similarity between two components corresponding to their respective Gaussian distributions is measured by the Bhattycharya distance function. Afterwards, GMMSE constructs a graph based on the new data samples and the representative matrix, and searches for the most representative cluster structure. Finally, we also design four criteria to assign the data samples to their corresponding clusters based on the unified cluster structure. The experimental results show that (i) GMMSE works well on synthetic datasets and real datasets in the UCI machine learning repository. (ii) GMMSE outperforms most of the previous cluster ensemble approaches.

1. Introduction

With the development of cluster ensemble techniques, a growing number of related approaches have been successfully applied to different fields [32,33,11,64,39,55], such as medicine, bioinformatics, and multimedia data mining. For example, Iam-On et al. [32] proposed a link-based cluster ensemble approach based on the similarity between clusters, and successfully applied it to both artificial and real datasets. They also applied this approach to solve categorical data clustering problem [33]. Christou [11] explored the optimization-based cluster ensemble approach which is formulated in terms of intra-cluster criteria, and applied it to the TSPLIB benchmark data sets. Yu et al. [64] studied the knowledge based cluster ensemble approach which is applied to perform cancer discovery from gene expression profiles. Mimaroglu and Aksehirli

[*] Corresponding author. Tel.: +86 20 62893506; fax: +86 20 39380288.
E-mail address: zhwyu@scut.edu.cn (Z. Yu).

0020-0255/$ - see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2014.01.030

[39] designed a divisive clustering ensemble approach called DICLENS, which is able to identify the cluster number automatically and achieved good performance on gene expression data sets. Weber et al. [55] gave a general definition of optimal clustering related to overlapping clustering solutions, which is useful for cluster ensemble approaches.

Compared with traditional clustering algorithms, cluster ensemble approaches represent a more effective technique since they have the ability to generate a unified clustering solution from multiple clustering solutions in the ensemble, and improve the effectiveness, stability and robustness of the clustering process. Most of the previous cluster ensemble approaches focus on the alignment of the labels of data samples derived from diverse clustering solutions, and do not take into account the fusion of multiple cluster structures obtained from various data sources into a unified structure. The cluster structure which summarizes information about the distribution of the data samples is more useful in a lot of scenarios. For example, as time passes, some data sources will gradually change, which will lead to the variation of the labels of data samples in different clustering solutions. In this scenario, the cluster structure of the data is more important than the labels of data samples. This raises an interesting question of how to construct a cluster structure ensemble, and identify the most representative cluster structure among the datasets.

There are a lot of useful applications for a cluster structure ensemble approach. For example, multiple sensors will generate a lot of datasets which have their own cluster structures in the area of mobile Internet. At the same time, the cluster structures of these datasets share a large number of similar characteristics. How to construct a unified cluster structure which captures the similarity of the cluster structures in different datasets generated from multiple sensors is an interesting problem deserving intensive exploration. For another example, the objective of clustering analysis on lung cancer datasets is to assign samples to their corresponding classes. The Lung adenocarcinomas dataset in [7] contains 203 samples assigned to 5 classes: adenocarcinoma, small-cell lung cancer, pulmonary carcinoids, squamous cell lung carcinomas, and normal lung tissues. Since there are a large number of datasets obtained by different research groups in the area of lung cancer research [16], it raises the question of how to find the most representative cluster structure from the cluster structures obtained from the different datasets.

In this paper, we design a new probabilistic cluster structure ensemble framework, referred to as the Gaussian mixture model based cluster structure ensemble framework (GMMSE), to identify the most representative cluster structure from the dataset. Specifically, GMMSE first integrates the bagging technique, the *K*-means algorithm and the Expectation–Maximization approach to generate diversity, and estimate the various cluster structures from different data sources. Then, it adopts the normalized cut algorithm [47] and the representative matrix constructed based on the set of cluster structures from different data sources to find the most representative cluster structure. Finally, GMMSE applies four assignment criteria, which are the nearest Gaussian model criterion (NGM), the average Gaussian model criterion (AGM), the nearest group center criterion (NGC) and the Gaussian model based majority voting criterion (GMV), to assign the data samples to their corresponding clusters based on this most representative cluster structure. The results in the experiment show that GMMSE achieves good performance on both synthetic datasets and real datasets in the UCI machine learning repository.

The contribution of the paper is fourfold. First, we proposed a Gaussian mixture model based cluster structure ensemble framework (GMMSE) to identify the most representative cluster structure. Second, four criteria are designed to assign data samples to their corresponding clusters based on this representative cluster structure. Third, the time and space complexity of GMMSE are analyzed. Fourth, the representative matrix is designed to capture the relationships among the components of the Gaussian mixture models. The Bhattacharya distance function is adopted to measure the similarity between two components with respect to their respective Gaussian distributions.

The remainder of the paper is organized as follows. Section 2 introduces the related works to cluster ensemble approaches. Section 3 describes the Gaussian mixture model based cluster structure ensemble framework, and analyzes the time and space complexity of GMMSE. Section 4 evaluates the performance of GMMSE through experiments on synthetic datasets, as well as several real datasets in the UCI machine learning repository. Section 5 draws conclusions and describes possible future works.

2. Related works

Recently, ensemble learning is gaining more and more attention since these approaches have the ability to provide more accurate, stable and robust final clustering results when compared with traditional approaches. Most of the ensemble learning approaches [45,24,43] can be categorized into three types, which are supervised learning ensemble, semi-supervised learning ensemble and unsupervised learning ensemble. Supervised learning ensemble, also called classifier ensemble, includes a number of popular approaches such as bagging [9], boosting [20], random forest [10], random subspace [27,23], rotation forest [44], ensemble based on random linear Oracle [36], and ensemble based on neural networks [71,70,56]. Adaboost is an example of forming the ensemble at the learning process, which adjusts the weights of training data samples in the learning process, and integrates multiple weak classifiers into a strong one. In other words, the main focus of Adaboost is to adjust the weights of a sequence of classifiers in the learning process, and the classifier in the former will affect the performance of the classifier in the latter. On the other hand, bagging is an example of forming an ensemble at the output stage, which integrates multiple learning results into a more representative result by a suitable voting scheme. Supervised learning ensemble includes the multi-view learning approach [21] and the consensus maximization approach [22].