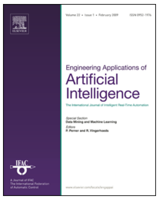
Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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



Hierarchical cluster ensemble selection

Ebrahim Akbari [a,b,*], Halina Mohamed Dahlan [a], Roliana Ibrahim [a], Hosein Alizadeh [c]

[a] *Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bahru, Malaysia*
[b] *Department of Computer Engineering, Islamic Azad University, Sari Branch, Sari, Iran*
[c] *Computer Engineering Department, Iran University of Science and Technology, 1684613114 Narmak, Tehran, Iran*

ARTICLE INFO

Article history:
Received 5 August 2014
Received in revised form
12 November 2014
Accepted 9 December 2014
Available online 31 December 2014

Keywords:
Cluster ensemble selection
Hierarchical clustering
Diversity
Consensus function

ABSTRACT

Clustering ensemble performance is affected by two main factors: diversity and quality. Selection of a subset of available ensemble members based on diversity and quality often leads to a more accurate ensemble solution. However, there is not a certain relationship between diversity and quality in selection of subset of ensemble members. This paper proposes the Hierarchical Cluster Ensemble Selection (HCES) method and diversity measure to explore how diversity and quality affect final results. The HCES uses single-link, average-link, and complete link agglomerative clustering methods for the selection of ensemble members hierarchically. A pair-wise diversity measure from the recent literature and the proposed diversity measure are applied to these agglomerative clustering algorithms. Using the proposed diversity measure in HCES leads to more diverse ensemble members than that of pairwise diversity measure. Cluster-based Similarity Partition Algorithm (CSPA) and Hypergraph-Partitioning Algorithm (HGPA) were employed in HCES method for obtaining the full ensemble and cluster ensemble selection solution. To evaluate the performance of the HCES method, several experiments were conducted on several real data sets and the obtained results were compared to those of full ensembles. The results showed that the HCES method led to a more significant performance improvement compared with full ensembles.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Clustering is one of the unsupervised rules for searching and analyzing data, which is used in different fields such as statistics, pattern recognition, machine learning, data mining, and bio-informatics (Jain, 2010; Quintana et al., 2003; De Angelis and Dias, 2014; Sun et al., 2012). Wide usage of clustering algorithms proves their usefulness in exploratory data analysis (Jain et al., 2000). The major aim of data clustering is to find groups of patterns (clusters) in such a way that patterns in one cluster can be more similar to each other than to patterns of other clusters. Because of characteristics of dataset, different clustering algorithms obtain different clustering results. It is difficult to choose a suitable algorithm for a given data set. Based on the Kleinberg theorem, there is no the best single clustering algorithm (Kleinberg, 2003).

Clustering ensemble, which is an approach in clustering problem, combines multiple clustering results (clusterings) to achieve final clusters without accessing the features or algorithms that obtain the clusterings. The combination of the clusterings is performed by a consensus algorithm. The clustering ensemble approach attempts to improve the quality and robustness of clustering results (Strehl and Ghosh, 2003; Fred and Jain, 2005; Mimaroglu and Erdil, 2013). Furthermore, clustering ensemble can achieve some properties such as novelty, stability, and scalability (Topchy et al., 2005). There are some applications of clustering ensemble in bio-informatics, image processing, and marketing (Strehl and Ghosh, 2003; Avogadri and Valentini, 2009; Ma et al., 2009; Mimaroglu and Erdil, 2010). Since clustering ensemble only needs to gain access to the base clusterings instead of the data itself, it provides a convenient approach to privacy preservation and knowledge reuse (Strehl and Ghosh, 2003). In many applications, for the objects under consideration, various clusterings may already come to exist. In this condition, these clusterings can be integrated into a single solution. For example, in market basket analysis, assume that a company already has various legacy customer segmentations based on geographical region, credit rating, demographics, and purchasing patterns in their retail stores, and so on. They want to reuse this pre-existing knowledge in order to form a single consolidated clustering. Because the legacy clusterings are provided largely by experts or by other companies by means of proprietary methods, for reusing this knowledge, there is a limited access to original features of raw data and the algorithms that obtain the clusterings (Strehl and Ghosh, 2003).

[*] Corresponding author at: Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bahru, Malaysia. Tel.: +60 177263810.
E-mail addresses: akbari@iausari.ac.ir, ebrahimakbari30@yahoo.com (E. Akbari).

http://dx.doi.org/10.1016/j.engappai.2014.12.005
0952-1976/© 2014 Elsevier Ltd. All rights reserved.

On the other hand, in contrast to the knowledge reuse, there could be a potential for greater gains when using an ensemble for the purpose of improving clustering quality (Strehl and Ghosh, 2003). Traditionally, a set of large library of clusterings is generated and then the consensus solution is obtained by a consensus function based on all base clusterings. Unlike classification problems where labels of data items are known beforehand, data items in unsupervised clustering problems are unlabeled which may some clustering results unreliability in large library of clusterings. Thus, not all obtained clusterings can truly benefit for the final solution of clustering ensembles (Azimi and Fern, 2009; Hong et al., 2009).

Recently, a subset of diversity is selected rather than all for combining the diversity to obtain the final result (Hadjitodorov et al., 2006). Cluster ensemble selection is mainly aimed to select a subset from a large library of clustering solutions to form a smaller cluster ensemble that performs as properly as or better than the set of all available clustering solutions (Kuncheva and Hadjitodorov, 2004; Fern and Lin, 2008; Azimi and Fern, 2009). Selective ensembles method is also on the basis of the supervised classification area in which it has been recognized that selective classifier ensembles always outperform the conventional ensemble methods in terms of achieving better solutions (Banfield et al., 2005; Zhang et al., 2006). In a straightforward classifiers selection method, the classifiers are ranked based on their individual performance on a held-out test set and the best ones are picked (Caruana et al., 2014). Whereas, in unsupervised clustering area, data items are unlabeled beforehand. As a result, this is not possible to estimate the quality of a single clustering result by computing its quality on the test set.

In ensemble selection, diversity and quality are two important factors that affect ensemble performance (Fern and Brodley, 2003). A few recent studies have investigated heuristically the question how a subset of ensemble members should be selected based on diversity and quality (Minaei-Bidgoli et al., 2014; Alizadeh et al., 2014; Naldi et al., 2013). The most successful method proposed by Fern and Lin (2008) is called the Cluster And Select (CAS) that combines quality and diversity. This, first, partitions the ensemble members into k (the number of clusters) clusters based on their similarities. Then, CAS selects the clusterings with the highest quality from each obtained cluster for the ensemble. They concluded that the use of both quality and diversity in cluster ensemble selection (CES) can make a higher improvement in the results compared to full ensembles. However, the drawback of CAS is that the number of k that can obtain the most appropriate ensemble size is uncertain and the concept of quality and diversity is loosely defined. To address the above problems, a hierarchical diversity selection strategy based on both diversity and quality is proposed to improve the traditional clustering ensemble performance. This strategy also solves the drawback of ensemble selection strategy in the CAS method.

This paper proposes a new combinational method called the Hierarchical Cluster Ensemble Selection (HCES). In the first step of the HCES method, a pairwise matrix of all available clustering members is constructed using two different diversity measures. It then employs three hierarchical methods including single-link, average-link, and complete-link to build a nested tree. An appropriate cut on the obtained tree creates diverse groups of primary partitions that guide us in targeted selection of smaller yet better performing ensemble. Finally, the HCES uses HGPA and CSPA consensus clustering algorithms for obtaining consensus clustering solutions. The HCES method obtains the final solution through the selection of an appropriate layer of the hierarchy. In the HCES method, there is no need to specify the value of k. The HCES empirically is compared to the full ensemble. The evaluation results obtained from different real data sets demonstrate statistically more significant performance improvement compared to the full ensembles. In addition, because of interpretability of the proposed method, the results are improved even with removing only one clustering; the removed clustering is considered as a noise. As a brief, the contributions of the present paper are as follow:

1. Proposing an automatic hierarchical cluster ensemble selection.
2. Proposing a diversity measure and applying to the proposed HCES method.
3. Applying three agglomerative clustering algorithms to the method and showing their effect on the performance of the method.

The rest of the paper is organized as follows. Section 2 gives an overview of related work. Section 3 introduces different diversity and quality measures. Section 4 presents the hierarchical ensemble selection method. Section 5 presents the experiments carried out on several real data sets and the obtained results. Finally, Section 6 concludes the paper and recommends future work.

2. Related work

Clustering ensemble is an approach that is widely adopted in clustering research to improve the quality and robustness of clustering results. Clustering ensemble includes two main parts: diversity (creating multiple clusterings) and consensus function (combining multiple clusterings). Recently, researchers have suggested the selection of diversity to improve the ensemble performance (Hadjitodorov et al., 2006; Jia et al., 2011; Hong et al., 2009). Fig. 1 shows the steps of clustering ensemble selection approach. In this section, some clustering ensemble methods and recent studies conducted on cluster ensemble design are reviewed.

2.1. Diversity generation

In ensemble classifier/clustering techniques, generating diversity is commonly used in supervised and unsupervised combining approaches. Various methods have been proposed in the literature for creating diversity or ensemble members, including

1. *Different parameter initializations*: Primary clusterings are created using repeated runs of a single clustering algorithm with several sets of parameter initializations such as cluster centers of the k-means clustering technique, which are known as homogeneous ensembles (Fred and Jain, 2005).
2. *Different clustering algorithms*: A number of different clustering algorithms are used together to generate primary clusterings,

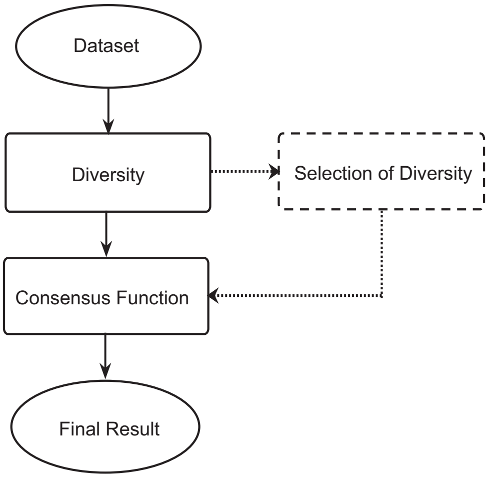


Fig. 1. Steps of the clustering ensemble selection approach.

Download English Version:

https://daneshyari.com/en/article/380413

Download Persian Version:

https://daneshyari.com/article/380413

Daneshyari.com