Applied Soft Computing xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Applied Soft Computing

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

Cluster ensemble framework based on the group method of data handling

Q1 Geer Teng[a], Changzheng He[b,*], Jin Xiao[b], Yue He[b], Bing Zhu[b], Xiaoyi Jiang[c]

[a] Institute of Social Development and Western China Development Studies, Sichuan University, Chengdu, Sichuan, China
[b] Business School, Sichuan University, Chengdu, Sichuan, China
[c] Department of Mathematics and Computer Science, University of Münster, Münster, Germany

ARTICLE INFO

Q3 *Article history:*
Received 21 January 2015
Received in revised form 9 November 2015
Accepted 24 January 2016
Available online xxx

Keywords:
Cluster ensemble
GMDH
Evolutionary algorithm
Least squares
CSPA
Semidefinite programming

ABSTRACT

Cluster ensemble is a powerful method for improving both the robustness and the stability of unsupervised classification solutions. This paper introduced group method of data handling (GMDH) to cluster ensemble, and proposed a new cluster ensemble framework, which named cluster ensemble framework based on the group method of data handling (CE-GMDH). CE-GMDH consists of three components: an initial solution, a transfer function and an external criterion. Several CE-GMDH models can be built according to different types of transfer functions and external criteria. In this study, three novel models were proposed based on different transfer functions: least squares approach, cluster-based similarity partitioning algorithm and semidefinite programming. The performance of CE-GMDH was compared among different transfer functions, and with some state-of-the-art cluster ensemble algorithms and cluster ensemble frameworks on synthetic and real datasets. Experimental results demonstrate that CE-GMDH can improve the performance of cluster ensemble algorithms which used as the transfer functions through its unique modelling process. It also indicates that CE-GMDH achieves a better or comparable result than the other cluster ensemble algorithms and cluster ensemble frameworks.

1. Introduction

Clustering is a process of grouping objects into different groups, such that the common properties of data in each cluster is high, and between different clusters is low. Clustering has been studied for several decades in the areas of pattern recognition, machine learning, applied statistics, communications and information theory and so on [1,2]. It has also been widely applied in many real world application domains, such as data mining [3], image segmentation [4], marketing [5,6] and customer relationship management [7].

Cluster ensemble algorithms have shown to be effective in improving the accuracy and stability of single clustering algorithms [8]. Consensus functions in the cluster ensemble algorithms play an important role in the modelling process. The most often used consensus functions are based on hypergraph, voting, mutual information, co-association and mixture model [9]. To the best of our knowledge, most of these consensus functions generate the final result by combining multiple labels of a given dataset in a direct approach (without iteration). Therefore, the final ensemble clustering results obtained by these direct approaches are usually inaccuracy compared with a given partition of the data (also called ground truth).

In this paper, we proposed a novel cluster ensemble framework, cluster ensemble framework based on the group method of data handling (CE-GMDH), aimed at improving the accuracy of the cluster ensemble algorithms. This framework consists of three components according to the theory of the group method of data handling (GMDH), including initial solutions, a transfer function and an external criterion. The initial solutions in the consensus function are the clustering solutions generated by the basic clustering algorithms. The transfer function is used to combine the initial solutions to generate some middle candidate solutions. External criteria in the GMDH are measures which evaluate the quality of the middle clustering solutions.

The contribution of this work is that a clustering ensemble framework based on the group method of data handling was proposed. With this framework, different base cluster analysis algorithms, transfer functions and external criteria can be used as its components to generate the final optimal ensemble through an iterative approach.

The basic idea of the CE-GMDH is to start from the initial solutions, generate new candidate solutions in every layer through

Q2 * Corresponding author. Tel.: +86 02885418891; fax: +86 02885418891.
E-mail address: hechangzheng@scu.edu.cn (C. He).

http://dx.doi.org/10.1016/j.asoc.2016.01.043
1568-4946/

combining every two middle candidate solutions from the preceding layer, then utilise external criteria to evaluate and select middle candidate solutions. This process continues until a final solution has been found.

According to different initial solutions, transfer functions and external criteria chosen, several CE-GMDH models can be built. In this paper, we proposed three CE-GMDH models based on different transfer functions: one based on the least squares approach, one based on the cluster-based similarity partitioning algorithm (CSPA) proposed by Strehl and Ghosh [8] and another one based on the semidefinite programming (SDP) [10].

CE-GMDH was compared with cluster ensemble algorithms which used as the transfer functions, other well-known cluster ensemble algorithms and cluster ensemble frameworks on several synthetic and real datasets. The experimental results on these datasets demonstrated that the CE-GMDH can improve the performance of cluster ensemble through its special modelling process. Statistical test results also indicated that CE-GMDH achieved better performances than the other cluster ensemble algorithms and cluster ensemble frameworks in terms of the evaluation measures.

The remainder of this paper is organised as follows: Section 2 reviews the literature. Then, we introduce the mechanism of GMDH in Section 3. Details of the proposed cluster ensemble framework are provided in Section 4. Section 5 presents the experimental design, results and analysis, and finally, concluding remarks are given in Section 6.

2. Literature review

Cluster ensemble has emerged as a powerful method for improving both the robustness as well as the stability of unsupervised classification solutions [11]. It has evolved rapidly in the last decade [12,13] and regarded as an important research branch in machine learning field [14].

The consensus function is one of the vital parts in cluster ensemble algorithms. There are five types consensus functions commonly used, including hypergraph partitioning, voting approach, mutual information algorithm, co-association based functions and finite mixture model [12,15]. Strehl and Ghosh [8] proposed three effective and efficient techniques based on hypergraph for obtaining high-quality consensus functions. These techniques include the CSPA, the hypergraph partitioning algorithm (HGPA) and the meta-clustering algorithm (MCLA). Voting-based consensus clustering refers to a distinct class of consensus methods in which the cluster label mismatch problem is explicitly addressed [16]. Tumer and Agogino [17] proposed voting active clusters as a method for obtaining ensemble clusterings, and the results showed that this method performs comparably to the best existing centralised cluster ensemble methods under ideal conditions. One popular cluster ensemble method based on the information-theoretic methods is quadratic mutual information (QMI) [18]. In this method, a high agreement between two clusterings induced high values of the category utility function [19]. Fred and Jain [20] explored the idea of evidence accumulation (EAC) for combining the results of multiple clustering algorithms. The final clustering solution was formed from the co-association matrix by linking the objects whose co-association value exceeded a certain threshold. Tsaipei [21] proposed CA-tree, which was a dendrogram-like hierarchical data structure, to facilitate efficient and scalable cluster ensembles for coassociation-matrix-based algorithms. Topchy et al. [22] designed a consensus function based on a finite mixture model and the final partition was found as a solution to the maximum likelihood problem for a given cluster ensemble. Recently, Singh et al. [10] proposed a nonlinear optimisation model to maximise the new agreement measure based on a 2D string encoding and transformed it into a strict 0–1 semidefinite program (SDP). Bayesian inference based function [23] and weighted least square based function [24] are two novel consensus functions which produce the final result in a direct approach.

Some cluster ensemble frameworks were also proposed by researchers. Yu et al. [25] designed a cluster ensemble framework based on random combination of transformation operators (CE-RCTO) to provide more accurate, stable and robust final results. Iam-On et al. [26] proposed an effective link-based ensemble framework (LinkCluE) to categorical data clustering. Yu et al. [27] designed a double self-organizing map based cluster ensemble framework (SOM^2CE). It integrates the self-organizing map twice in the ensemble framework for discovering the underlying structure. Yu and Zhou [28] proposed a cluster ensemble framework based on three-way decisions (CE-TWD). Experimental results showed that CE-TWD is effective in the quality of the consensus partitions. Parvin and Minaei-Bidgoli [29] proposed an effective clustering ensemble framework based on selection of fuzzy, named fuzzy weighted locally adaptive clustering (FWLAC). Zhuang et al. [13] proposed a principled cluster ensemble framework to combine individual clustering solutions that are based on the consensus partition and applied it for Internet security.

Although the above-mentioned cluster ensemble algorithms obtained better results, the experimental results showed that these algorithms were inaccurate when handling different clustering solutions generated by different basic clustering algorithms. This is mainly due to the consensus functions used in these algorithms generated the final optimal result by combining all clustering solutions in a direct approach.

In our study, GMDH was introduced to find the final cluster ensemble result through an evolutionary approach. GMDH was first developed by Ivakhnenko [30] as an analysis method for complex systems modelling and identification. The basic idea of GMDH is to build a multilayer feed-forward neural network structure. In every layer, it constructs candidate solutions through combining two of the previously selected models, then utilise external criterion [31] to evaluate and select some best models to enter the next layer. Finally, it gets the optimal complexity model by termination principle.

GMDH has been successfully applied in a wide range of areas used for cluster analysis [32], classification [33,34], regression [35] and ensemble learning [36]. In 2008, Sarycheva [32] developed the algorithm for objective cluster analysis based on GMDH principles. But objective cluster analysis is a single cluster analysis algorithm. In our study, the basic idea of GMDH is helpful to combine the basic clustering results. Thus, a final ensemble clustering result effectively can be found.

3. Mechanism of the GMDH

In our study, the proposed cluster ensemble algorithm, CE-GMDH, is built based on the theory of GMDH. Therefore, the mechanism of the GMDH is introduced in this section.

To build a GMDH model, the following three components must exist to be fulfilled: initial solutions, a transfer function and external criteria.

GMDH is an evolutionary approach which including operations of mutation and selection. It starts from the initial model set, carries on parameter estimation by a transfer function and obtains middle candidate solutions (inherit, mutation), evaluates the middle candidate solutions by external criterion [37], and finally chooses some best ones (and thus ignores the others) for the next layer. Therefore, the modelling process of the GMDH can be described as the following four steps:

Please cite this article in press as: G. Teng, et al., Cluster ensemble framework based on the group method of data handling, Appl. Soft Comput. J. (2015), http://dx.doi.org/10.1016/j.asoc.2016.01.043