Neurocomputing 205 (2016) 116–129

Contents lists available at ScienceDirect

Neurocomputing

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

An application of a metaheuristic algorithm-based clustering ensemble method to APP customer segmentation

R.J. Kuo [a,*], C.H. Mei [a], F.E. Zulvia [a], C.Y. Tsai [b]

[a] Department of Industrial Management, National Taiwan University of Science and Technology, No. 43, Section 4, Kee-Lung Road, Taipei, Taiwan 106, ROC
[b] Department of Industrial Engineering and Management, Yuan Ze University, No. 135, Yuan-Tung Road, Chungli, Taoyuan, Taiwan, ROC

ARTICLE INFO

Article history:
Received 26 July 2015
Received in revised form
3 January 2016
Accepted 12 April 2016
Communicated by: Dr. Swagatam Das
Available online 7 May 2016

Keywords:
Cluster analysis
Clustering ensembles
Genetic algorithm
Particle swarm optimization
Artificial bee colony
Customer segmentation

ABSTRACT

This study proposes a metaheuristic-based clustering ensemble method. It integrates the clustering ensembles algorithm with the metaheuristic-based clustering algorithm. In the clustering ensembles, this study performs an improved generation mechanism and a co-association matrix in the co-occurrence approach. In order to improve the efficiency, a principle component analysis is employed. Furthermore, three metaheuristic-based clustering algorithms are proposed. This paper uses a real-coded genetic algorithm, a particle swarm optimization and an artificial bee colony optimization to combine with clustering ensembles algorithm. The experimental results indicate that the proposed metaheuristic-based clustering ensembles algorithms have better performance than metaheuristic-based clustering without clustering ensembles method. Furthermore, the proposed algorithms are applied to solve a customer segmentation problem. The real problem is come from a mobile application. Among all of the proposed algorithms, the artificial bee colony optimization-based clustering ensembles algorithm outperforms other algorithms. Therefore, the marketing strategy for the real application is made based on the best result.

1. Introduction

Cluster analysis has been applied in many different fields including data mining, image segmentation and document exploration. It aims to identify the data structure by organizing the data into several groups. A large variety of clustering algorithms have been proposed. They are K-means, fuzzy c-means, single linkage agglomerative hierarchical clustering (SLHC) and etc. However, there is no one algorithm which is suitable for all types of datasets [1–3]. Therefore, this paper proposes a combination of some clustering algorithms using a clustering ensemble method in order to improve the clustering methods and obtain better clustering results. Clustering ensemble gives a better solution than a single clustering algorithm in term of robustness, novelty and stability [1,4,5]. The proposed clustering ensemble method comprises of two steps. They are generation mechanism and consensus function. The generation mechanism generates the data partition. Then the process is continued by consensus function to integrate the results of the generation mechanism. Furthermore, three metaheuristic-based clustering algorithms are combined with the clustering ensembles. This paper employs three single clustering algorithms namely, the real-coded GA-based K-means clustering ensembles (GKCE), PSO-based K-means clustering ensembles (PSOKCE) and ABC-based K-means clustering ensembles (ABCKCE).

The validation is conducted using iris, wine, tae, flame, bank authentication and D31 datasets. Furthermore, the proposed method is applied to solve a customer segmentation problem. This real application is taken from a weight control APP.

The remainder of this paper is organized as follows. Section 2 presents the relevant background of this paper. Section 3 presents a novel clustering ensembles method. Furthermore, the validation is given in Section 4. In Section 5, application of the proposed method to the weight loss APP is discussed. Finally, concluding remarks are made in Section 6.

2. Literature review

This section briefly recalls some theories applied in this paper, including clustering methods, meta-heuristic algorithm-based clustering methods and clustering ensembles.

2.1. Cluster analysis

Clustering is an unsupervised data processing which classifies instances into some groups. Herein, a cluster should comprise of

* Corresponding author. Tel.: +886-2-27376328, fax: +886-2-27376344 .
E-mail address: rjkuo@mail.ntust.edu.tw (R.J. Kuo).

http://dx.doi.org/10.1016/j.neucom.2016.04.017
0925-2312/

similar data and should be different with other clusters [6]. Cluster analysis has been applied in many applications. Therefore, many clustering methods have been proposed. In general, clustering methods can be divided in two categories, hierarchical and partition clustering methods. Hierarchical clustering constructs the clusters step-by-step based on the similarity of two data points. There are some similarity measurements such as single linkage, complete linkage, average linkage, etc [7]. Partition clustering methods generate a number of clusters simultaneously. Basically, they use a clusters center and assign each data into the nearest cluster. K-means is the most popular partition clustering method. Many papers have employed K-means algorithm since it can generate a quite good clustering result with a relatively simple algorithm [8–10].

2.2. Meta-heuristic algorithm-based clustering methods

Recently, many researches have applied a metaheuristic algorithm to improve the clustering methods. Genetic algorithm is one of metaheuristic algorithm which has widely combined with clustering methods [11–13]. Furthermore, application of particle swarm optimization (PSO) in clustering problem can be founded in [14–17]. In addition to the GA and PSO, a combination of artificial bee colony algorithm (ABC) with clustering methods also has been introduced in the previous paper [18]. Basically, the metaheuristic algorithm represents the clustering result in their solution representation. They start from any random initial clustering results and then iteratively improve their clustering results. GA updates its solutions using selection, crossover and mutation operators. PSO explores the search space based on a certain direction considering the particle and social best solutions. On the other hand, the updating rules in ABC are inspired from bee behavior in finding the nectar source.

2.3. Clustering ensembles

Clustering ensembles is a clustering method which combines several clustering methods to improve the stability of the result [19]. It comprises of two steps, generation mechanism and the consensus function (see Fig. 1). The generation mechanism generates the data partitions. It then projects all data points onto one dimension. Furthermore, any clustering methods with different parameter setting can be applied to generate the initial clusters. The second step combines the clustering results from the previous step using a consensus function. There are two main consensus methods, object co-occurrence and a median partition. This study applies an object co-occurrence approach using a co-association matrix since it is easier to be understood [20]. The co-association matrix determines the similarity between objects.

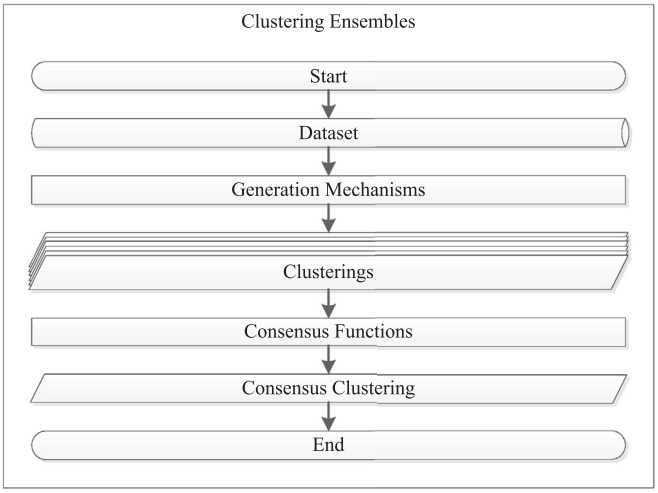

Fig. 1. The flow chart of clustering ensembles.

The most challenging problem in clustering ensembles is finding the best combination of the employed clustering methods. Azimi et al. [4] proposed an improvement of clustering ensembles algorithm with GA. Their proposed method obtains the best combination using a co-association function as the consensus functions. Yang et al. [21] proposed an innovative weighted combination model with multiple parts. It also applies a PSO algorithm to optimize the parameter used in the combining process. Furthermore, Li et al. [15] introduced an clustering ensembles algorithm using K-means and agglomerative hierarchical clustering algorithms. Herein, they applied single linkage, complete linkage and average linkage in the agglomerative hierarchical. The experimental results reveal that average linkage gives better results than single linkage or complete linkage. Clustering ensembles using a combination of partition and hierarchical clustering algorithms is also proposed by Zheng et al. [22].

3. Methodology

This paper proposes three metaheuristics-based clustering ensembles algorithms, which are a real-coded genetic algorithm-based K-means clustering ensembles (GKCE), a particle swarm optimization-based K-means clustering ensembles (PSOKCE) and an artificial bee colony-based K-means clustering ensembles (ABCKCE). This section discusses these proposed algorithms as well as the research methodology. The methodology proposed in this paper comprises of several steps. The first step applies a clustering ensembles algorithm. In order to improve the efficiency, a principle component analysis (PCA) is performed to reduce the problem complexity. Furthermore, the three metaheuristics-based K-means clustering algorithms are applied independently. Fig. 2 illustrates the framework of the proposed method.

3.1. The proposed clustering ensemble methods

The proposed clustering ensemble method comprises of two steps: generation mechanism and consensus function using a co-association matrix (see Fig. 3).

The generation mechanism divides the dataset into several subsets. In the proposed algorithm, this mechanism employs different algorithm to avoid similar subsets. There are four basic clustering algorithms applied, namely, K-means, single linkage agglomerative hierarchical clustering (SLHC), complete linkage agglomerative hierarchical clustering (CLHC) and average linkage agglomerative hierarchical clustering (ALHC). From these four algorithms, three different combinations are proposed. They are: (1) K-means algorithm, (2) SLHC, CLHC and ALHC, and (3) K-means, SLHC, CLHC and ALHC. The subsets are also generated from different attributes, $A_a, 1 \le a \le m$, and instances $I_i, 1 \le i \le n$. The combinations of attributes and instances are generated randomly. Fig. 4 illustrates the subsets generation.

The co-association matrix is a polymerization step. It combines different clustering results obtained by the generation mechanism (see Fig. 5). The elements of the co-association matrix are either 0 or 1. It is determined using Eq. (1).

$$c_{ij} = \frac{\sum_{p=1}^{q} \delta(x_{pi}, x_{pj})}{q} \tag{1}$$