Applied Soft Computing 30 (2015) 113–122

Contents lists available at ScienceDirect

Applied Soft Computing

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

Non-dominated sorting genetic algorithm using fuzzy membership chromosome for categorical data clustering

Chao-Lung Yang*, R.J. Kuo, Chia-Hsuan Chien, Nguyen Thi Phuong Quyen

Department of Industrial Management, National Taiwan University of Science and Technology, Taipei, Taiwan, ROC

ARTICLE INFO

Article history:
Received 6 May 2014
Received in revised form
12 November 2014
Accepted 9 January 2015
Available online 31 January 2015

Keywords:
Categorical attributes
Multi-objective optimization
Genetic algorithm
Fuzzy clustering

ABSTRACT

In this research, a data clustering algorithm named as non-dominated sorting genetic algorithm-fuzzy membership chromosome (NSGA-FMC) based on K-modes method which combines fuzzy genetic algorithm and multi-objective optimization was proposed to improve the clustering quality on categorical data. The proposed method uses fuzzy membership value as chromosome. In addition, due to this innovative chromosome setting, a more efficient solution selection technique which selects a solution from non-dominated Pareto front based on the largest fuzzy membership is integrated in the proposed algorithm. The multiple objective functions: fuzzy compactness within a cluster (π) and separation among clusters (*sep*) are used to optimize the clustering quality. A series of experiments by using three UCI categorical datasets were conducted to compare the clustering results of the proposed NSGA-FMC with two existing methods: genetic algorithm fuzzy K-modes (GA-FKM) and multi-objective genetic algorithm-based fuzzy clustering of categorical attributes (MOGA (π, *sep*)). Adjusted Rand index (ARI), π, *sep*, and computation time were used as performance indexes for comparison. The experimental result showed that the proposed method can obtain better clustering quality in terms of ARI, π, and *sep* simultaneously with shorter computation time.

1. Introduction

A clustering procedure is a process to partition a given dataset into several subsets based on a similarity or dissimilarity measure. The standard distance measurement such as Euclidean distance is used to calculate the distance between two points of the given dataset in the clustering algorithm. However, there is not any natural order or distance among the parties that can be directly applied on the categorical dataset. Categorical attribute such as gender and blood type can be identified as ordinal or non-ordinal are very common in real world dataset. Each categorical attribute is represented with a small set of unique categorical values such as [A, B, AB and O] for the blood type attribute. Due to the discreteness and unordered of categorical data, a new clustering algorithm is needed to accommodate the dissimilarity measurement of categorical data.

Several methods were proposed to handle dissimilarity measurement on categorical data. For example, converting categorical data to numerical data and calculating the dissimilarity by the existing dissimilarity method is one way to handle the categorical data clustering. However, if the data is nominal with no ordering, the assigning numerical value might cause bias or misleading on clustering result [1]. Another approach is counting the value occurrence (frequency-based) to calculating the dissimilarity. For instance, K-modes algorithm, which is modified from K-means algorithm [2–4] uses modes instead of mean as centroid of a cluster [5]. Because the frequency-based dissimilarity can be adaptive to all kinds of categorical data without the limitation, in this research, K-mode clustering method is utilized on studying on categorical datasets.

For either continual or categorical data clustering, most of clustering algorithms rely on optimizing a single objective function such as the intra-distance within a cluster to obtain the data partition. For example, genetic algorithm (GA) based clustering method which is based on the rule of Darwinian evolution generally uses a single objective function to search for a better data partitioning in a dataset. The clustering result based on the single objective function might be only good on one perspective (lower total intra-distance in a cluster), but not be able to fulfill other clustering objective such as enlarging the separation among clusters. Please note the ideal clustering result might be the data partitioning where data points

* Corresponding author. Tel.: +886 227303621; fax: +886 227376344.
E-mail addresses: clyang@mail.ntust.edu.tw (C.-L. Yang), rjkuo@mail.ntust.edu.tw (R.J. Kuo), lucky6844@gmail.com (C.-H. Chien), quyen.ntp@gmail.com (N.T.P. Quyen).

http://dx.doi.org/10.1016/j.asoc.2015.01.031
1568-4946/

in one cluster should be closer to each other (compactness), and each cluster is far away from other clusters (more separated).

To solve the real-world problems where multiple objectives might exist, developing a clustering algorithm to solve the clustering problems which have more than one objective function concurrently is necessary. There are several validity measures were introduced in [6] such as Minkowski score and adjusted Rand index (ARI) for internal cluster measures; Xie-Beni index and $\mathfrak{I}$ index for external cluster measure. Besides, there are two indices to evaluate the clustering results proposed in [7] including the compactness of the clusters and separation among the clusters. Furthermore, a combination of both criteria can be considered at the same time for clustering problem on real world data.

Some researchers have used a linear combination technique to integrate the multiple criterions to a single objective for clustering [8]. However, the weights for each objective functions are required in advance and it might not be easily defined the relative weight of different clustering criteria. Therefore, the compactness and separation should be optimized separately instead of joining them into a single measure for optimization.

The major purpose of this research is to develop a clustering algorithm which is able to handle multi-objective optimization on a categorical data. Based on the literature review, several multi-objective clustering algorithms such as GA based fuzzy K-means algorithm (GA-FKM) [9] were studied. Particularly, multi-objective genetic algorithm MOGA (π, *sep*) [2] which aims to optimize the compactness and separation simultaneously by fuzzy GA algorithm is our learning benchmark. After reviewing the performance of MOGA (π, *sep*), we proposed a new method called non-dominated sorting genetic algorithm-fuzzy membership chromosome (NSGA-FMC) which uses different chromosome settings against MOGA (π, *sep*) and has a relatively simple solution selection algorithm by taking advantage of using fuzzy membership chromosome. In this research, the performance of the proposed NSGA-FMC is compared based on the different performance measures: ARI, compactness π, fuzzy separation *sep*, and computational time.

The rest of paper is organized as follows. Section 2 describes the literature review in categorical clustering and fuzzy K-modes clustering algorithm. In Section 3, the proposed NSGA-FMC algorithm is introduced and compared to MOGA (π, *sep*). In Section 4, the experimental results are presented to demonstrate the clustering quality and efficiency of the proposed algorithm. Finally, the conclusion and future research direction are shown in Section 5.

2. Literature review

2.1. Categorical clustering

Essentially, a data clustering is used to reveal patterns where data points with more similar characteristics are assigned in the same cluster. The data points belong to the same cluster are given the same label. The distance-based dissimilarity which measures distance between data elements is used to group data elements based on the intra-distance evaluation [10].

Most of the clustering algorithms are developed for data patterns whose the dissimilarity between two points of dataset can be measured by the standard distance measurement [2]. However, categorical datasets might not have any inherent distance or natural order among the elements. For example, gender, profession, position, and hobby of customers are usually defined as categorical attributes in the customer table. Each categorical attribute is represented with a small set of unique categorical values such as [Female, Male] for the gender attribute. Unlike numerical data, categorical values are discrete and usually unordered. Therefore, the clustering algorithms for numeric data cannot be directly used to cluster categorical data that exists in many real world applications.

In the past, several methods for clustering categorical data were proposed. Gibson et al. proposed a non-linear dynamic system approach to handle categorical clustering [11]. Zhang et al. integrated Renyi entropy with an ant-based clustering algorithm to solve the clustering problem based on the ants' behavior [12]. Another approach also used the optimization algorithm which named artificial bee colony algorithm (ABC) by Karaboga and Ozturk [13]. Both ant-based clustering and ABC algorithm can handle the clustering problem not only for categorical dataset but also for multivariate dataset. Besides, there are some previous works for categorical clustering such as CACTUS – a fast summarization-based algorithm [14], ROCK – a robust hierarchical algorithm which utilizes links and not distances for merging cluster [15], or COOLCAT – a entropy-based algorithm which aims to minimize the expected entropy of the clusters [16]. However, those methods only consider single objective function for categorical clustering.

In data mining research, some research efforts have been done on the development of new techniques for clustering categorical data. The K-modes clustering algorithm [17] is one of the first algorithms for clustering large categorical data. The K-modes approach modifies the standard K-means process for clustering categorical data by replacing the Euclidean distance function with the simple matching dissimilarity measure. It uses modes to represent cluster centers and updates modes with the most frequent categorical values on each iteration of the clustering process. These modifications guarantee that the clustering process converges to a local minimal result.

2.2. Fuzzy K-modes clustering

In order to deal with categorical data clustering problems, the K-modes algorithm which is the modification of K-means algorithm was proposed by Huang [18]. The main differences of K-modes against K-means method are listed as follows:

i K-modes uses a simple matching dissimilarity measure for the categorical objects;
ii K-modes uses a frequency-based method to find the modes; and
iii K-modes replaces the means of clusters with the modes.

The fuzzy version of K-modes has improved the K-modes algorithm by assigning confidence to objects in different clusters [5,9,19]. These confidence values can be used to decide the core and boundary objects of clusters, thereby providing more useful information for dealing with boundary objects. The objective of the fuzzy K-modes clustering is to find W and Z that minimize:

$$F_c(W,Z) = \sum_{l=1}^{k}\sum_{i=1}^{n} w_{li}^{\alpha} d_c(x_i, z_l), \tag{1}$$

Subject to

$$0 \le W_{li} \le 1, \quad 1 \le l \le k, \quad 1 \le i \le n, \tag{2}$$

$$\sum_{l=1}^{k} w_{li} = 1, \quad 1 \le i \le n, \tag{3}$$

$$0 < \sum_{i=1}^{n} w_{li} < n, \quad 1 \le l \le k, \tag{4}$$

where $\alpha > 1$ is the weighting component, $W = (w_{li})$ is the $k \times n$ fuzzy membership matrix, and $Z = \{z_1, z_2, \ldots, z_k\}$ is the set of cluster centers. Note that $\alpha = 1$ gives the hard K-modes clustering, i.e., the K-modes algorithm.