Information Sciences 267 (2014) 35–57

Contents lists available at ScienceDirect

Information Sciences

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

Incremental learning based multiobjective fuzzy clustering for categorical data

Indrajit Saha *, Ujjwal Maulik

Department of Computer Science and Engineering, Jadavpur University, Kolkata 700032, West Bengal, India

ARTICLE INFO

Article history:
Received 4 July 2013
Received in revised form 19 October 2013
Accepted 30 December 2013
Available online 14 January 2014

Keywords:
Categorical data
Fuzzy clustering
Multiobjective differential evolution
Random forest
Statistical test

ABSTRACT

The problem of clustering categorical data, where attribute values cannot be naturally ordered as numerical values, has gained more importance in recent time. Due to the special properties of categorical attributes, the clustering of categorical data seems to be more complicated than that of numerical data. Although, a few clustering algorithms that optimize single clustering objective have been proposed. It has found that such single measure may not be appropriate for all kind of datasets. Hence, in this article, an Incremental Learning based Multiobjective Fuzzy Clustering for Categorical Data is proposed. For this purpose, a multiobjective modified differential evolution based fuzzy clustering algorithm is developed. Thereafter, it integrates with the well-known supervised classifier, called random forest, using incremental learning to propose the aforementioned technique. Here, the multiobjective algorithm produces a set of optimal clustering solutions, known as Pareto-optimal solutions, by optimizing two conflicting objectives simultaneously. Subsequently, through incremental learning using random forest classifier final solution is evolved from the ensemble Pareto-optimal solutions. The results of the proposed method are demonstrated quantitatively and visually in comparison with widely used state-of-the-art methods for six synthetic and four real life datasets. Finally, statistical test is conducted to show the superiority of the results produced by the proposed method.

1. Introduction

The clustering problem in generally deals with partitioning a dataset consisting of n patterns in d-dimensional space into K distinct set of clusters, such that the data points within the same cluster are more similar to each other than to data points in other clusters [1–3]. The main objective of any clustering technique is to produce a $K \times n$ partition matrix $U(X)$ of the given dataset X, consisting of n patterns, $X = \{x_1, x_2, \ldots, x_n\}$. The partition matrix may be represented as $U = [u_{ij}]$, $i = 1, \ldots, K$ and $j = 1, \ldots, n$, where u_{ij} is the membership of pattern x_j to the ith cluster. For fuzzy clustering of the data, $0 < u_{ij} < 1$, i.e., u_{ij} denotes the degree of belongingness of pattern x_j to the ith cluster. Hence, the data clustering help us to get the insight of the data distribution.

Recent literature review shows that various types of clustering algorithms have been developed [2,4,5], which are primarily focused on numerical data. However, growing demands in the field of environmental study [6], market basket study [7], DNA or protein sequence study [8,9], text mining [10], computer security [11] and psychological study [12,13], where data are mostly categorical in nature, have been attracted researchers in recent years to analyze the categorical data. In this regards, some of the widely used and recently proposed categorical data clustering methods are available in the literature like Partitioning Around Medoids (PAM or KMdd) [14], K-Modes (KMd) [15], Fuzzy K-Modes (FKMd) [16] and Fuzzy K-Medoids

* Corresponding author. Tel.: +91 9433845689.
E-mail address: indra@icm.edu.pl (I. Saha).

0020-0255/$ - see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.12.057

(FKMdd) [17], CAUTUS [18], STIRR [19], ROCK [20], COOLCAT [11], Squeezer [21], TSFKMd [22], ALG-RAND [23], CLICK [24], LIMBO [25], TCSOM [26], ccdByEnsemble [27], QROCK [28], CCDV [29], MMR [30], *k*-ANMI [31], G-ANMI [32] and MoDEFCCD [33]. Few more techniques are also reported in [34–47]. Each of these algorithms has its own strengths and weaknesses. For a particular dataset, different algorithms, or even the same algorithm with different parameters, usually provide distinct solutions. Moreover for the purpose of clustering, they use a single objective function to optimize. A single objective function may not work uniformly well for different kinds of categorical data. Hence, optimizing multiple conflicting objectives simultaneously is expected to provide better solution.

Usually, multiobjective clustering [48,49] produced a set of nondominated solutions, known as Pareto-optimal solutions [50,51]. The selection of best solution from these Pareto-optimal solutions is also challenging. In most of the cases [48,49], selection is done by the use of some other external cluster validity measure. As a result, other solutions in Pareto-optimal set are neglected, which may contain good information. Recently, cluster ensembles have emerged as an effective solution that is able to overcome these limitations, and improve the robustness as well as the quality of clustering results [52–56]. Therefore, it is quite significant in multiobjective case, if ensemble of Pareto-optimal solutions is used to create a final clustering result.

To address the above facts, in this article, an Incremental Learning based Multiobjective Fuzzy Clustering for Categorical Data is proposed. The contribution is two folded. First, a new multiobjective clustering for categorical data is proposed. Second, using *Incremental Learning* that uses well-known supervised classifier, called random forest [57], final solution is evolved from the ensemble Pareto-optimal solutions. For the multiobjective clustering, a newly proposed variant of Differential Evolution (DE) [58] is used. The developed modified differential evolution based fuzzy clustering technique has already been used for clustering numerical and categorical data [58,33]. The basic framework of it is used to develop the modified differential evolution based multiobjective clustering technique for categorical data, named as Multiobjective Modified Differential Evolution based Fuzzy Clustering for Categorical Data (MOMoDEFCCD). In MOMoDEFCCD algorithm, the index encoding scheme is used to encode the medoids in vectors as well as scaling is introduced after mutation to map the encoded index values within the range of categorical objects. Thereafter, ensemble Pareto-optimal solutions of MOMoDEFCCD are taken to yield better final clustering result. The integration of multiobjective clustering with incremental learning scheme is termed as Incremental Learning based Multiobjective Fuzzy Clustering for Categorical Data (ILMOFCCD). Effectiveness of the proposed ILMOFCCD technique is established quantitatively and visually in comparison with Multiobjective Modified Differential Evolution based Fuzzy Clustering for Categorical Data (MOMoDEFCCD) by selecting the best Pareto-optimal solution using external validity measure, single objective version of Modified Differential Evolution based Fuzzy Clustering for Categorical Data (MoDEFCCD) [33], Genetic Algorithm based Average Normalized Mutual Information Clustering (G-ANMI) [32], Min-Min-Roughness (MMR) [30], Tabu Search based Fuzzy K-Modes (TSFKMd) [22] and widely used state-of-the-art methods like K-Medoids (KMdd) [14], Fuzzy K-Medoids (FKMdd) [17], Fuzzy K-Modes (FKMd) [16] and Average Linkage (AL) [2] hierarchical clustering techniques for six artificial and four real life categorical datasets. Finally, a nonparametric statistical test, called Wilcoxon rank sum test [59], is performed to establish the superiority of the results.

The rest of the article is organized as follows: Sections 2 briefly describe the categorical data clustering algorithms, random forest classifier and the basics of multiobjective optimization. Proposed techniques are described in Section 3. Experimental results conducted on several synthetic and real life datasets along with statistical test are presented in Section 4. Finally, Section 5 concludes the article.

2. Background

In this section, categorical data clustering algorithms, supervised classifier random forest and the basics of multiobjective optimization are described briefly.

2.1. Categorical data clustering algorithms

This section describes some well-known categorical data clustering algorithms. Mathematical description and the dissimilarity measure used for clustering categorical data are mentioned before to go into the details of these algorithms. Therefore, let $X = \{x_1, x_2, \ldots, x_n\}$ be a set of n objects having categorical attribute domains. Each object $x_j,\ j = 1, 2, \ldots, n$, is described by a set of p attributes $A = \{A_1, A_2, \ldots, A_p\}$. Let $DOM(A_l),\ 1 \leqslant l \leqslant p$, denotes the domain of the lth attribute and it consists of different q_l categories such as $DOM(A_l) = a_l^1, a_l^2, \ldots, a_l^{q_l}$. Hence, the jth categorical object is defined as $x_j = [x_{j1}, x_{j2}, \ldots, x_{jp}]$, where $x_{jl} \in DOM(A_l),\ 1 \leqslant l \leqslant p$. Subsequently, to compute the dissimilarity measure between two objects $x_i, x_j \in X$, recently proposed *Cho et al.* dissimilarity measure [46] is used in this paper. It is defined as follow.

$$D(x_i, x_j) = \sum_{l=1}^{p} d(x_{il}, x_{jl}) \tag{1}$$

where

$$d(x_{il}, x_{jl}) = 1 - Sim(x_{il}, x_{jl}) \tag{2}$$