Knowledge-Based Systems 77 (2015) 114–127

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Ensemble based rough fuzzy clustering for categorical data

Indrajit Saha [a,*,1], Jnanendra Prasad Sarkar [b,1], Ujjwal Maulik [b,*]

[a] Institute of Computer Science, University of Wroclaw, Wroclaw 50383, Poland
[b] Department of Computer Science and Engineering, Jadavpur University, Kolkata 700032, India

ARTICLE INFO

Article history:
Received 2 June 2014
Received in revised form 22 November 2014
Accepted 19 January 2015
Available online 28 January 2015

Keywords:
Categorical data
Ensemble
Machine learning
Rough fuzzy clustering
Similarity/dissimilarity measure
Statistical test

ABSTRACT

Categorical data is different from continuous data, where the values of attribute do not follow any natural ordering. Moreover, inherent complexities like uncertainty, vagueness and overlapping among clusters make the analysis of real life categorical data set more difficult. Recent literature review shows that the well-known categorical data clustering techniques are using different similarity/dissimilarity measures to tackle the inherent complexities of the categorical attribute values. Generally, it is hard to find single method and cluster validity measure that can be used as perfect or standard for all kinds of categorical data sets. Hence, in this paper first, a clustering method for categorical data is proposed by fusing rough set and fuzzy set theories. Subsequently, an ensemble based framework is designed with the recently proposed similarity/dissimilarity measures in order to have better clustering results for different types of categorical data sets. For this purpose, the proposed rough fuzzy clustering method is used sequentially with the integration of different measures to evolve the clustering solutions. Using consensus of these solutions, pure classified, semi rough and pure rough points are identified. Thereafter, machine learning method, called Random Forest, is used in incremental way to classify the semi and pure rough points using pure classified points to yield better clustering results. The performance of the proposed method has been demonstrated in comparison with several other recently developed clustering methods. Additionally, the selection of Random Forest in the proposed framework is justified by comparing its performance with other well-known machine learning methods like K-Nearest Neighbor and Support Vector Machine. Ten categorical data sets are used for the experimental purpose. Finally, statistical significance test has been conducted to judge the superiority of the results.

1. Introduction

Data clustering is an important area of research over the past few decades. The aim of clustering is to identify interesting patterns in the data set and group them according to some similarities among themselves. Clustering methods are extensively used in different fields including pattern recognition [1,2], medical system [3], trend analysis [4], customer segmentation [5] etc. Most of the literatures about data clustering are focused mainly on continuous data. The geometric structure of such data can be easily exploited using some defined distance functions. The problem becomes difficult while processing categorical data, because of some special inherent properties within the attributes. For example, the categorical attribute, *color*, can have different values like *red*, *green*, *blue* etc, where natural ordering is missing.

In recent times, the clustering of categorical data has drawn much attention. Some challenges of clustering categorical data are discussed in [6,7]. A number of methods for clustering categorical data have been proposed in [8–38]. Among them *K*-Modes (KMd) [12] and Fuzzy *K*-Modes (FKMd) [13] are widely used. In [30], Kim et al. improved the FKMd by proposing a fuzzy centroid based clustering approach, while in [39], another fuzzy approach was introduced by Umayahara et al. for document classification. However, as fuzzy set theory is not always able to handle uncertainty and vagueness very well, the advantages of rough set theory is exploited in [40,41]. Generally in rough clustering, a point either belongs to a cluster with membership degree 1, or it belongs to the boundary regions of multiple clusters. Hence, the points belonging to the boundary regions can be thought to be situated in the overlapping portions of two or more clusters. Min-Min-Roughness (MMR) [27] based clustering is the first step toward clustering categorical data using rough set theory. However, MMR is sensitive to out layers, which directly effect on cluster size. Therefore, it is

[*] Corresponding authors.
E-mail addresses: indra@icm.edu.pl (I. Saha), umaulik@cse.jdvu.ac.in (U. Maulik).
[1] Joint first author and contributed equally.

http://dx.doi.org/10.1016/j.knosys.2015.01.008
0950-7051/© 2015 Elsevier B.V. All rights reserved.

natural to use rough and fuzzy sets together to cluster simple as well as uncertain, vague and overlapping categorical data sets.

The similarity/dissimilarity measure between two objects also plays an important role for classifying categorical data set. In this regard, a number of research papers can be found in [14,20,24,28,42,43]. However, mostly they are using one particular measure. For example, KMd [12] uses simple matching [8] between objects. Similarly, some other techniques, such as FKMd and Tabu Search based Fuzzy *K*-Modes (TSFKMd) [19] etc are also using the simple matching measure. However, the simple matching measure often fails to discover some hidden properties within the categorical values [42,43]. Hence in [42], a dissimilarity measure was proposed by using concept of simple matching by considering the frequency of mode in current cluster. Other than the simple matching approach for formulating distance metric, Gibson et al. [15] proposed co-occurrence approach. This approach assumes that the similarity between two categorical values lies on their co-occurrence with common values or a set of values. Later a distance metric based on co-occurrence probability of two categorical values was presented in [44]. However, this metric is also unable to properly identify the significance of an attribute. Most recent time, Cao et al. proposed important and effective dissimilarity measures in [43].

The aim of this article is twofold. First, rough and fuzzy sets based method, namely Rough Fuzzy *K*-Modes (RFKMd), is proposed, that can deal with uncertainty and vagueness with the concept of lower and upper approximation in rough set theory. Moreover, it can also efficiently handle the overlapping partition by exploiting the advantages of fuzzy membership function. Second, a framework, named as Ensemble based Rough Fuzzy Clustering (ERFC), is developed considering the recently proposed dissimilarity measures to tackle different types of categorical data sets. For this purpose, Rough Fuzzy *K*-Modes algorithm is used with different dissimilarity measures. It is observed that the number of rough objects varies with the use of different dissimilarity measures for the same data set. Hence, a consensus of these ensemble solutions are taken to identify *pure classified*, *semi rough* and *pure rough* points. Thereafter, pure classified points are considered as training set of Random Forest (RF) [45] classifier to test or classify semi rough and pure rough points in incremental way [46,47]. In this case, if one dissimilarity measure fails to handle any particular inherent complexity within data set, some other dissimilarity measure can take care of that particular complexity. Thus, it produces reliable training set for Random Forest to classify rest of the unclassified data objects. Two other machine learning methods, Support Vector Machine (SVM) [48] and *K*-Nearest Neighbor (*K*-NN) [49] are also used for comparison. The effectiveness of the proposed ensemble based rough fuzzy clustering is shown by comparing the results with Clustering Categorical Data By Cluster Ensemble (ccdByEnsemble) [24], Min-Min-Roughness (MMR) [27], Genetic Algorithm based Average Normalized Mutual Information Clustering (G-ANMI) [29], Tabu Search based Fuzzy *K*-Modes (TSFKMd) [19] and widely used state-of-the-art methods like *K*-Modes (KMd) [12] and Fuzzy *K*-Modes (FKMd) [13]. Experimental results are provided for six artificial and four real life categorical data sets by evaluating cluster validity indices and visual plots. Finally, a statistical significance test has been performed to establish the superiority of the proposed methods.

The rest of the paper is organized as follows: Section 2 briefly describes background of categorical data clustering techniques, rough set and machine learning methods. The different dissimilarity measures are discussed in Section 3. Section 4 presents the proposed methods. Experimental results are explained in Section 5 and finally, Section 6 concludes the paper.

2. Background

This section describes briefly the background of categorical data clustering algorithms, rough set theory and machine learning methods.

2.1. Brief description of categorical data clustering algorithms

Since the inception of *K*-Modes (KMd) [12] and Fuzzy *K*-Modes (FKMd) [13] algorithms, clustering of categorical data have drawn much attention of researchers. Before that, only few algorithms were developed [8–11]. KMd algorithm uses the *K*-Means [50] paradigm, whereas Fuzzy *K*-Modes [13] algorithm is the extension of the well-known Fuzzy C-Means [51] algorithm for clustering categorical data. Later in subsequent years, CACTUS [14], STIRR [15], ROCK [16], Squeezer [18], TSFKMd [19], COOLCAT [17], CLOPE [52], LIMBO [22], CORE [53], TCSOM [23], ccdByEnsemble [24], ANMI [24] were developed.

Each of these methods 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, this is observed that these methods are not working equally well for uncertain, vague and overlapping data sets. While in real world application, these are important issues because sharp boundary between the clusters is often lagging. For this purpose, Min-Min-Roughness (MMR) [27] based clustering technique was used. However, MMR does not remove any outliers which may affect the size of the clusters. Recently mutual information based categorical data clustering (CDC) method, called k-ANMI, was developed in [28]. It is quite similar to the *K*-Means algorithm. This algorithm may also get trapped into local optimal solution. Hence, Genetic Algorithm based Average Normalized Mutual Information (G-ANMI) was presented in [29]. The ANMI is computed on a set of given partitions by assuming that good combined partition may share as much information as possible. Note that G-ANMI is already found to perform better than k-ANMI [28], TCSOM [23], and Squeezer [18] algorithms. Therefore, in this article, Ensemble based Rough Fuzzy Clustering (ERFC) is compared with KMd [12], FKMd [13], TSFKMd [19], MMR [27], G-ANMI [29] and ccdByEnsemble [24].

2.2. Brief description of rough set theory

Rough set theory [54] which is subsequently studied in various literatures [46,55–59], is to approximate the concept of uncertainty. The theory is modeled with the concept of *Lower Approximation* and *Upper Approximation* space of a set of objects. If U and A are defined as two finite non-empty sets, called *universe* and set of attributes, then the pair (U, A) is called an *information system*. The *equivalence relation* determined by $B \subseteq A$ can be denoted as $R(B) \subseteq U \times U$. The *equivalence relation* which is also called *indiscernibility relation* in the context of rough set theory, partitions the universe, U into some subsets which are called equivalence classes. Objects in same equivalence class are indistinguishable. An equivalence class of $R(B)$, i. e., a block of partition in quotient set U/B, containing x is denoted by $B(x)$. The pair, $\langle U, R(B) \rangle$ or simply $\langle U, B \rangle$ is called a Pawlak approximation space. If $P(U)$ is the power set of U and $X \in P(U)$ is any arbitrary set, then X may not be well described in the approximation space $\langle U, B \rangle$ in crisp manner. The main objective of the rough set theory is to handle such scenario. Therefore, $X \subseteq U$ can be approximated by a pair of subsets of U. These two subsets are known as *Lower Approximation* denoted by $\underline{B}(X)$ and *Upper Approximation* denoted by $\overline{B}(X)$ respectively. The difference between *lower* and *upper approximation*