Pattern Recognition 56 (2016) 1–15

Contents lists available at ScienceDirect

Pattern Recognition

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

Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy

Xiao Zhang [a,*], Changlin Mei [b], Degang Chen [c], Jinhai Li [d]

[a] Department of Applied Mathematics, School of Sciences, Xi'an University of Technology, Xi'an, PR China
[b] Department of Statistics, School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, PR China
[c] Department of Mathematics and Physics, North China Electric Power University, Beijing, PR China
[d] Faculty of Science, Kunming University of Science and Technology, Kunming, PR China

ARTICLE INFO

Article history:
Received 1 July 2015
Received in revised form
2 February 2016
Accepted 22 February 2016
Available online 3 March 2016

Keywords:
Mixed data
Feature selection
Fuzzy rough set theory
Information entropy

ABSTRACT

Feature selection in the data with different types of feature values, i.e., the heterogeneous or mixed data, is especially of practical importance because such types of data sets widely exist in real world. The key issue for feature selection in mixed data is how to properly deal with different types of the features or attributes in the data set. Motivated by the fuzzy rough set theory which allows different fuzzy relations to be defined for different types of attributes to measure the similarity between objects and in view of the effectiveness of entropy to measure information uncertainty, we propose in this paper a fuzzy rough set-based information entropy for feature selection in a mixed data set. It is proved that the newly-defined entropy meets the common requirement of monotonicity and can equivalently characterize the existing attribute reductions in the fuzzy rough set theory. Then, a feature selection algorithm is formulated based on the proposed entropy and a filter-wrapper method is suggested to select the best feature subset in terms of classification accuracy. An extensive numerical experiment is further conducted to assess the performance of the feature selection method and the results are satisfactory.

1. Introduction

Feature selection is a common technique to preprocess data in data mining, machine learning and pattern recognition [10,11,15,24,25,27,29,34,37,41,48,50,56,59,61,64]. The aim of feature selection is to remove redundant or irrelevant features (attributes) without significantly decreasing the prediction accuracy of the classifier built only by the selected features, or to yield a class distribution as close as possible to the original class distribution [10]. It is very common that a real-world data set generally includes heterogeneous attributes which means that the attributes in the data set take different types of values. For example, in the data concerning the medical diagnosis, the attributes or features may include sex, age, weight, and blood pressure of the patients, where the feature sex is nominal (categorical), the feature age is integer-valued, and the features weight and blood pressure are real-valued (numerical). Therefore, it is of great importance to study the issue of feature selection for a data set with different types of attribute values which we call mixed data or heterogeneous data as usual.

Currently, there are two categories of methods to process mixed data. One category is to transform heterogeneous data into homogeneous data [6,7,12,33,39,51]. For example, real-valued features are discretized as nominal ones by some discretization methods. However, as indicated in [17], discretization leads to the loss of neighborhood structures and order structures of the real-valued features in real spaces. As another kind of the transformation methods, nominal features are treated as real-valued features by coding a series of integer numbers. Nevertheless, it may be nonsense in some instances for a nominal feature to be measured by real values. The other category is to measure different types of features using different criteria and then integrate the different measuring results [5,17,26,42,46]. For example, two kinds of information entropy measures are respectively defined by Liang et al. [26] for real-valued features and nominal features and are combined to deal with mixed data for clustering analysis.

Although the fuzzy rough set theory, a generalization of the traditional rough set theory [36], was originally developed to deal with the real-valued data, it can also be used to process mixed data, because fuzzy rough sets are based on fuzzy relations which are allowable to be defined for different kinds of attributes to measure the similarity between objects. For example, the

* Corresponding author.
E-mail addresses: zhangxiaoo@126.com (X. Zhang), clmei@mail.xjtu.edu.cn (C. Mei), wscz973@163.com (D. Chen), jhlixjtu@163.com (J. Li).

http://dx.doi.org/10.1016/j.patcog.2016.02.013
0031-3203/

equivalence relation defined for a nominal attribute can be taken as a special fuzzy relation in which the values 1 and 0 are used to respectively measure the complete similarity and dissimilarity (indiscernibility and discernibility) between objects. As done in [4,20,46], the distance functions can also be employed to define fuzzy relations for real-valued attribute based on the fact that the smaller the distance between two objects is, the higher the similarity between them is. According to [9,44,62], many measures such as the set-operator, the T-similarity relations and the interval-inclusion degree can be used to define fuzzy relations for set-valued attributes, fuzzy attributes and interval-valued attributes, respectively. Therefore, from the viewpoint of fuzzy relations, the fuzzy rough set theory provides a new framework for dealing with mixed data. This framework is different from the aforementioned two categories of the methods in which the fist category could be regarded as a preprocessing technique of data while the second category is a strategy of "divide-and-conquer".

Feature selection based on the fuzzy rough set theory, usually called attribute reduction, was first investigated in [21] and further improved in [1,4,8,18–20,22,23,38,43,57,60,63]. Specifically, Jensen and Shen [21] presented a dependency function based-reduct and designed a heuristic algorithm to search for one of the reducts. However, this algorithm is not convergent for some data sets [1]. The reason is, as pointed out in [43], that the defined fuzzy lower approximation of the dependency function may result in the consequence that the membership of the fuzzy lower approximation of an attribute subset is greater than that of the original attribute set, which is a contradiction with the idea of the rough set theory. To fill this gap, [18,19] took the fuzzy lower approximation defined in [13] as that of the dependency function and studied the corresponding attribute reduction problem. Tsang et al. [43] investigated the granular structure of the fuzzy lower approximation defined in [13] and generalized the traditional positive region preserved reduct to the fuzzy positive region preserved reduct. In fact, a dependency function-based reduct is also a fuzzy positive region preserved reduct. The other attribute reductions of fuzzy rough sets mainly focused on the improvement of the fuzzy lower approximations in the dependency function or the fuzzy positive region. One can refer to [4,20,57,63] for the details.

In view of the effectiveness of information entropy to measure the uncertainty in information and the flexibility of the fuzzy rough set theory in dealing with heterogeneous attributes, using information entropy to study the fuzzy rough set-based attribute reduction provides a new way to select features in mixed data. It has been known that Hu et al. [19] defined a conditional entropy based on fuzzy rough sets to characterize the dependency function-based reduct and then used the entropy to develop a feature selection algorithm. As indicated in [16], however, the equivalent characterization of the conditional entropy for the dependency function-based reduct does not hold for general fuzzy decision systems. Furthermore, this conditional entropy cannot satisfy monotonicity for general fuzzy decision systems, which likely results in an improper evaluation function for feature selection. Therefore, the motivation of this paper is to provide a novel monotonic conditional entropy for selecting features of mixed data based on the fuzzy rough set theory and to present the corresponding feature selection algorithm. The new conditional entropy is not only monotonous but also able to equivalently characterize the dependency function-based reduct for general fuzzy decision systems.

The remainder of this paper is organized as follows. Before embarking on the main issues, we briefly review in Section 2 some basic knowledge about fuzzy rough sets. In Section 3, the new conditional entropy is defined and its properties are investigated. A conditional entropy-based heuristic algorithm is then formulated in Section 4 for feature selection. In Section 5, some numerical experiments are conducted to assess the performance of the proposed algorithm. The paper is ended with a summary.

2. Preliminaries

In order to facilitate the subsequent discussions, we first introduce the definition of fuzzy relations and some basic knowledge about fuzzy rough sets and attribute reduction in fuzzy decision systems.

2.1. Fuzzy relations

Let U be a nonempty universe of discourse and $F(U \times U)$ be the fuzzy power set on $U \times U$. R is called a fuzzy relation on $U \times U$ if $R \in F(U \times U)$, where $R(x,y)$ measures the strength of relationship between $x \in U$ and $y \in U$.

Let R be a fuzzy relation on $U \times U$. R is reflexive if $R(x,x)=1$ for any $x \in U$; R is symmetric if $R(x,y)=R(y,x)$ for any $x,y \in U$; and R is T-transitive if $R(x,y) \geq T(R(x,z),R(z,y))$ for a triangular norm T and any $x,y,z \in U$. Furthermore, R is called a T-similarity relation if R is reflexive, symmetric and T-transitive. Specially, if $T=\min$, R is called a fuzzy equivalence relation.

2.2. Fuzzy rough sets

In the pioneering work [13], a pair of lower and upper approximation operators of a fuzzy set X based on a T-similarity relation R is defined, for each $x \in U$, as

$$\underline{R}X(x) = \inf_{y \in U} \max\{1-R(x,y), X(y)\} \tag{1}$$

and

$$\overline{R}X(x) = \sup_{y \in U} \min\{R(x,y), X(y)\} \tag{2}$$

to measure the degree of x certainly belonging to X and the degree of x possibly belonging to X, respectively, on which the fuzzy rough set of X is defined by $(\underline{R}X, \overline{R}X)$.

The existing research on fuzzy rough sets mainly focuses on constructing approximations of fuzzy sets along the line of $\underline{R}$ and $\overline{R}$. In [2,13,14,28,30–32,35,40,52–55,58], constructive and axiomatic approaches were employed to study approximations of fuzzy rough sets. Since the existing research on attribute reduction is mainly based on the fuzzy rough sets in [13], we skip the reviews of other kinds of fuzzy rough sets and one can refer to [32,35,40,58] for the details of the contents. The work of this paper is also based on the fuzzy rough sets in [13].

2.3. Fuzzy information systems and fuzzy decision systems

Definition 1 (*Chen [3]*). A fuzzy information system is a pair (U,A) with a mapping $a_t : U \to V_{a_t}$ for each $a_t \in A$, where $U=\{x_1,x_2,\dots,x_n\}$ is the universe of discourse, $A=\{a_1,a_2,\dots,a_m\}$ is the attribute set on which a fuzzy relation $R_{\{a_t\}}$ is defined for each attribute $a_t \in A$, and V_{a_t} is the domain of a_t. The fuzzy relation of a subset $B \subseteq A$ is defined by $R_B = \bigcap_{a_t \in B} R_{\{a_t\}}$.

As pointed out in Introduction, it is possible to define the corresponding fuzzy relations for the attributes with different types of values. Then, a data set with heterogeneous attributes can be treated as a fuzzy information system for further analysis.

Definition 2 (*Chen [3]*). A fuzzy decision system is a pair $(U, A \cup D)$ with $A \cap D = \varnothing$, where (U,A) is a fuzzy information system, A is called the conditional attribute set and $D=\{d\}$ is called the