Knowledge-Based Systems 67 (2014) 162–168

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Feature selection via neighborhood multi-granulation fusion

Yaojin Lin [a,*], Jinjin Li [a,b], Peirong Lin [a], Guoping Lin [b], Jinkun Chen [b]

[a] School of Computer Science, Minnan Normal University, Zhangzhou 363000, PR China
[b] School of Mathematics and Statistics, Minnan Normal University, Zhangzhou 363000, PR China

ARTICLE INFO

Article history:
Received 7 January 2014
Received in revised form 18 May 2014
Accepted 29 May 2014
Available online 10 June 2014

Keywords:
Granular computing
Feature selection
Multi-granulation
Neighborhood rough sets
Granularity influence

ABSTRACT

Feature selection is an important data preprocessing technique, and has been widely studied in data mining, machine learning, and granular computing. However, very little research has considered a multi-granulation perspective. In this paper, we present a new feature selection method that selects distinguishing features by fusing neighborhood multi-granulation. We first use neighborhood rough sets as an effective granular computing tool, and analyze the influence of the granularity of neighborhood information. Then, we obtain all feature rank lists based on the significance of features in different granularities. Finally, we obtain a new feature selection algorithm by fusing all individual feature rank lists. Experimental results show that the proposed method can effectively select a discriminative feature subset, and performs as well as or better than other popular feature selection algorithms in terms of classification performance.

1. Introduction

Many data mining and pattern recognition systems suffer from the curse of dimensionality. This motivates the search for suitable feature selection methods [7,9,15,24,29,46]. In practice, many application fields, such as bio-informatics and text categorization, involve databases in which both the number of rows (objects) and columns (features) increase rapidly. The high-dimensional nature of the data presents a challenge to learning algorithms. This is because not all features can contribute to the discriminative power, and the correlated features may bring many disadvantages to traditional learning algorithms, such as low efficiency, over-fitting, and poor performance. To ease this problem, it is desirable to reduce the high-dimensionality of data, as this enhances the accuracy of pattern recognition and produces a more compact classification model with better generalization.

As we know, the feature selection technique plays a non-trivial role in speeding up learning and improving classification performance [7,9,11,15,46]. To date, a number of feature selection algorithms have been developed for classification learning. The process of feature selection can be divided into two steps. First, metrics such as mutual information [2,3,28], consistency [4,12], dependency [9,10,48], and the classification margin [33] are used to evaluate the quality of candidate features. Second, a search strategy is designed to solve the given optimization function. Feature selection may employ a heuristic search, genetic optimization and greedy search, or another intelligent search algorithm [34].

Although many algorithms have been developed for feature selection, very little work has considered a multi-granulation view. Granular computing, proposed by Zadeh [44], is an approximation schema that can effectively solve a complex program at a certain level or at multiple levels of granulation, and has attracted increasing interest [26,27,38,43,45]. There are many representative granular computing models, such as rough sets [25], fuzzy sets [27,44], probabilistic rough sets [41,42], covering rough sets [49], and neighborhood rough sets [8–10,18,39]. Of these, neighborhood rough sets provide an effective granular computing model for the problem of heterogeneous feature subset selection, and have been widely applied in cancer recognition, image annotation, and vibration diagnosis. For multi-granulation rough sets, neighborhood rough sets compute the neighborhoods of samples from which to extract information granularity (*in this paper, neighborhood size and granularity are equivalent terms*), and different information granularities can be induced by different neighborhood sizes [16,17].

It has been shown that diverse results can be obtained from different granular spaces for a given learning task. Indeed, given the same set of objects, different granular spaces can provide complementary predictive powers, and the prediction accuracy is significantly improved by combining their information [20–22,30]. The method of combining multiple granularities from different

[*] Corresponding author. Tel.: +86 13960044089.
E-mail addresses: yjlin@mail.hfut.edu.cn (Y. Lin), jinjinli@mnnu.edu.cn (J. Li), zzlprfj@163.com (P. Lin), gplin@163.com (G. Lin), cjk99@163.com (J. Chen).

http://dx.doi.org/10.1016/j.knosys.2014.05.019
0950-7051/© 2014 Elsevier B.V. All rights reserved.

granular spaces is a key issue. There are two key factors that improve learning performance in a granular computing model: constructing multiple granular spaces, or combining multiple granular spaces. Constructing multiple granular spaces mainly focuses on multiple binary relations or different neighborhood sizes in neighborhood rough sets. Many methods of constructing multiple granular spaces have been proposed. Qian et al. [31] extended Pawlak's rough sets model to a multi-granulation rough sets model, using set approximations defined by multiple equivalence relations on the universe. Wu et al. [36,37] presented a formal approach to granular computing with multi-scale data measured at different granulation levels, whereas Du et al. [5] adopted neighborhood cover reduction for rule learning. Hu et al. [9,10] used a neighborhood rough sets model as a uniform framework in which to understand and implement neighborhood classifiers. In terms of combining multiple granular spaces, multi-granulation fusion or multi-granulation cooperative learning are usually applied for a given task. Zhu et al. [47] explored an ensemble learning technique for evaluating and combining models derived from multi-granulation based on margin distribution optimization. Liang et al. [15] proposed an efficient rough feature selection algorithm for large-scale datasets by stimulating multi-granulation, considering a sub-table of a dataset as a small granularity.

Although many algorithms have been proposed for feature selection, they are mainly described from the viewpoint of single granulation. In this paper, we propose a new filtering feature selection using multi-granulation. This makes full use of the power of each granularity, as well as the complementary knowledge from all reasonable granularities. We first obtain different feature rank lists with respect to different neighborhood granular spaces based on neighborhood rough sets. These lists are fused based on a cross-entropy Monte Carlo algorithm to give a final feature rank list, which should be as close as possible to representing all individual rank lists simultaneously. Finally, we get the number of selected feature subsets determined by single granularity, which can obtain a selected feature subset directly. Experimental results show that the proposed multi-granulation-based feature selection algorithm is effective.

The rest of this paper is organized as follows. Section 2 introduces the related concept of neighborhood rough sets. In Section 3, we analyze the influence of granularity, introduce the basic idea of feature rank fusion, and propose a feature selection algorithm that incorporates multi-granulation. Numerical experiments and their results are described in Section 4. Finally, conclusions are given in Section 5.

2. Preliminary knowledge on neighborhood rough sets

Rough sets theory, proposed by Pawlak [25], has been proven to be a relatively new soft computing tool for feature selection, rule extraction, decision analysis, and knowledge discovery from nominal data in recent years [3,13,14,19,23,30,35,40]. However, it cannot deal with data set with numerical feature or heterogeneous feature. Therefore, Hu et al. [10] proposed a neighborhood rough sets model based on neighborhood relation.

Formally, an information system for classification problem can be written as a quadruple $IS = (U, A, V, f)$, where: U is a non-empty finite set of samples, called a universe, A is a non-empty finite set of features, V is the union of feature domains such that $V = \bigcup_{a \in A} V_a$. For each $a \in A$ and $x \in U$, a mapping $f : U \times A \rightarrow V$ is an information function such that $f(x, a) \in V_a$. For a decision table, we can split set A into conditional features C and class (or decision) feature D, respectively. The conditional features represent measured features of the instances, while the class feature is a posteriori outcome of classification.

Definition 1 10. Given arbitrary $x_i \in U$ and conditional feature space C, the neighborhood $\delta_C(x_i)$ of x_i in the space C is defined as

$$\delta_C(x_i) = \{x_j | x_j \in U, \Delta_C(x_i, x_j) \leqslant \delta\},$$

where Δ is a metric function, $\forall x_1, x_2, x_3 \in U$, which satisfies

(1) $\Delta_C(x_1, x_2) \geqslant 0$, $\Delta_C(x_1, x_2) = 0$: if and only if $x_1 = x_2$;
(2) $\Delta_C(x_1, x_2) = \Delta_C(x_2, x_1)$;
(3) $\Delta_C(x_1, x_2) + \Delta_C(x_2, x_3) \geqslant \Delta_C(x_1, x_3)$.

Theorem 1 10. *Given a metric space $< \Omega, \Delta >$, and the non-empty set of samples $U = \{x_1, x_2, \ldots, x_m\}$. We have*

(1) $\forall x_i \in U : \delta(x_i) \neq \emptyset$;
(2) $\bigcup_{i=1}^{m} \delta(x_i) = U$.

Given a metric space $< \Omega, \Delta >$, the family of neighborhood granules $\{\delta(x_i) | x_i \in U\}$ forms an elemental granule systems, which covers the universe, rather than partition it. A neighborhood relation N on the universe U can be represented as a relation matrix $(r_{ij})_{n \times n}$, where

$$r_{ij} = \begin{cases} 1, & \Delta(x_i, x_j) \leqslant \delta, \delta \in [0, 1], \\ 0, & otherwise. \end{cases}$$

Neighborhood relations satisfy the properties of symmetry and reflexivity, which are a kind of similarity relations. Neighborhood relations draw the similarity between objects according to distances, and the objects in the same neighborhood granule are close to each other.

Definition 2 10. Given the non-empty set of samples U and a neighborhood relation N over U, and $< U, N >$ is called a neighborhood approximation space. For any $X \subseteq U$, two subsets of objects, called upper and lower approximations of X in $< U, N >$, are defined as

$$\overline{N}X = \{x_i | \delta(x_i) \cap X \neq \emptyset, x_i \in U\};$$
$$\underline{N}X = \{x_i | \delta(x_i) \subseteq X, x_i \in U\}.$$

Theorem 2 10. *Given a metric space $< \Omega, \Delta >$, and two neighborhood sizes δ_1 and δ_2, if $\delta_1 \geqslant \delta_2$, we have*

(1) $\forall x_i \in U : N_1 \supseteq N_2, \delta_1(x_i) \geqslant \delta_2(x_i)$;
(2) $\forall x_i \in U : \underline{N_1}X \subseteq \underline{N_2}X; \overline{N_2X} \supseteq \overline{N_1X}$.

Definition 3 10. Given a neighborhood decision system $NDT =< U, C \cup D, V, f >$, and D divides U into N equivalence classes: $X_1, X_2, \ldots, X_N, B \subseteq C$ generates the neighborhood relation N_B on $U, \delta_B(x_i)$ is a neighborhood granule generated on feature space B. Then, the upper and lower approximations of D with respect to attributes B are defined as

$$\overline{N_B}D = \cup_{i=1}^{N} \overline{N_B}X_i;$$
$$\underline{N_B}D = \cup_{i=1}^{N} \underline{N_B}X_i,$$

where

$$\overline{N_B}X = \{x_i | \delta_B(x_i) \cap X \neq \emptyset, x_i \in U\};$$
$$\underline{N_B}X = \{x_i | \delta_B(x_i) \subseteq X, x_i \in U\}.$$

Definition 4 10. Given a neighborhood decision system $NDT =< U, C \cup D, V, f >$, the neighborhood dependency degree of $B \subseteq C$ with respect to D is defined as