Information Sciences 366 (2016) 134–149

Contents lists available at ScienceDirect

Information Sciences

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

Cost-sensitive feature selection based on adaptive neighborhood granularity with multi-level confidence

Hong Zhao[a,b], Ping Wang[a,c], Qinghua Hu[d,e,*]

[a] School of Computer Software, Tianjin University, Tianjin 300072, China
[b] Lab of Granular Computing, Minnan Normal University, Zhangzhou 363000, China
[c] School of Science, Tianjin University, Tianjin 300072, China
[d] School of Computer Science and Technology, Tianjin University, Tianjin 300072, China
[e] Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin 300072, China

ARTICLE INFO

Article history:
Received 18 August 2015
Revised 10 March 2016
Accepted 22 May 2016
Available online 27 May 2016

Keywords:
Cost-sensitive learning
Feature selection
Granular computing
Neighborhood granularity
Neighborhood rough sets

ABSTRACT

Neighborhood rough set model is considered as one of the effective granular computing models in dealing with numerical data. This model is now widely discussed in feature selection and rule learning. However, there is no theoretical analysis on the issue of neighborhood granularity selection, the influence of sampling resolution, test and misclassification costs on modeling. In this paper, we design an adaptive neighborhood rough set model according to data precision and develop a fast backtracking algorithm for neighborhood rough sets based cost-sensitive feature selection by considering the trade-off between test costs and misclassification costs. In the proposed model, the neighborhood granularity, based on the 3σ rule of statistics, is adaptive to data precision that is described by the multi-level confidence of the feature subsets. Our experiments, thoroughly performed on 12 datasets, demonstrate the effectiveness of the model and the efficiency of the backtracking algorithm.

1. Introduction

Granular computing is potentially used in the machine learning and data mining domains, and aims to develop a granular view to interpret and solve problems [18,50]. In granular computing [8,20,30,41], elements in a granule may be drawn together by indistinguishability, similarity, proximity, or functionality [38,43]. There are representative models for granular computing, such as Pawlak rough sets [28,31], covering rough sets [51], neighborhood rough sets [13,22], and so on. Neighborhood rough set model [14] is one of the effective granular computing models in dealing with numerical data which widely exists in domains of scientific research [23], knowledge engineering [37], medical analysis [19], cancer recognition [12], and tumor classification [34].

Cost-sensitive learning is also a key problem in the machine learning and it is listed as one of the top 10 challenging problems in the data mining domain [36]. During the past ten years, many researchers have investigated issues related to cost-sensitive learning [4,15,24,44]. Cost-sensitive feature selection is one of the most fundamental problems in this domain [6,11,16,21]. In addition, Zhou et al. discussed cost-sensitive neural networks in [48], and introduced multi-class cost-sensitive learning in [49]. Min et al. built a hierarchical model for cost-sensitive decision systems in [25]. Among these researches,

* Corresponding author. Tel.: +86 15122108020.
E-mail address: huqinghua@tju.edu.cn, huqinghua@hit.edu.cn (Q. Hu).

http://dx.doi.org/10.1016/j.ins.2016.05.025
0020-0255/

two major costs are the test costs and the misclassification costs. These costs are the two most important aspects in practical applications.

Selecting a proper neighborhood granularity plays a crucial role in granular computing, and there are close relationships between the test costs and the misclassification costs. However, current neighborhood rough set models [26,47] suffer from a major limitation: there is no theoretical analysis on the issue of neighborhood granularity selection, the influence of sampling resolution, test and misclassification costs on modeling. The size of neighborhood has effect on consistency of neighborhood spaces and their approximation ability [50]. Therefore, the neighborhood should be constructed from data, and the size of neighborhood can be adaptively derived from the precision of the data.

In real applications, the precision of the data is different because of the universality of measurement error. Normally, measurement errors obey a normal distribution, which is an important form of uncertain data [1,5]. Confidence level of normal distribution is the frequency that the interval contains the measurement error, and represents the precision of the data [7]. A smaller confidence level will result in a narrower confidence interval [3]. The confidence level is 100% which will result in the confidence interval $(-\infty, \infty)$. The neighborhood size can be set by using the confidence interval according to the 3σ rule of normal distribution. The 3σ rule states that for a normal distribution nearly all (99.73%) of the values lie within three standard deviations, and the values larger than 3σ are considered as abnormal values and rejected. The neighborhood margin based on the measurement errors can be derived from the 3σ rule.

In this paper, we study a cost-sensitive feature selection problem based on adaptive neighborhood granularity with the support of confidence level. First, we study how to adaptively set the size of neighborhood based on the confidence level and different feature subsets. For instance, the confidence level for the measurement errors is 95% when one feature is selected, and the confidence level will be 95% $\times$ 95% when two features are selected if they are independent. Obviously, it is difficult to achieve the margin when multiple features are selected in parallel. Therefore, to ensure that the total confidence level is 95% when two features are selected, the confidence level for each selected feature should be $\sqrt{95\%}$. We adopt the different confidence levels for the different numbers of the selected features, called a multi-level confidence.

Secondly, we study a new neighborhood rough set model based on an adaptive neighborhood granularity with multi-level confidence, and discuss the properties of the new model. It is notable that the monotonicity of the neighborhood rough set model does not hold in this case. The addition of a feature can result in a decrease or an increase of a total cost, which is computed by considering trade-off between test costs and misclassification costs.

Finally, we study the cost-sensitive feature selection problem and develop a fast backtracking algorithm with the 3σ rule. The algorithm requires iterating the whole objects to obtain a neighborhood set of one object. Therefore, the computational cost to obtain the neighborhood sets of each object is approximated to the square of the number of the objects [23]. According to the 3σ rule, the measurement value larger than three standard deviations is considered as an abnormal value and rejected. Therefore, we can eliminate effectively the outliers and reduce the computing space without computing the values which are larger than three standard deviations.

Twelve open datasets from the UC Irvine Machine Learning Repository (UCI) [2] are employed to study the effectiveness and efficiency of the adaptive neighborhood by multi-level confidence. The experiments are undertaken with the open source software Cost-sensitive rough sets (Coser) [27]. The experimental results indicate that the proposed model effectively handles numerical data with different precisions under different numbers of selected features. Compared with the fixed neighborhood, the total cost of the adaptive neighborhood is significantly small. In addition, we compare the efficiency of two backtracking algorithms with respect to running time to verify the efficiency of the fast backtracking algorithm. The results demonstrate that the proposed fast backtracking algorithm is competent.

The remainder of this paper is structured as follows: Section 2 reviews the neighborhood rough set model. Section 3 describes an adaptive neighborhood model with multi-level confidence. A feature selection algorithm is designed for the minimal cost problem of dealing with adaptive neighborhood with multi-level confidence in Section 4. In Section 5, we discuss experimental results, and analyze the effectiveness of the adaptive neighborhood. Finally, conclusions and further work are given in Section 6.

2. Background and problem description

In this section, the neighborhood rough set model is introduced and the monotonicity of the neighborhood rough set is discussed.

2.1. Neighborhood rough sets

As the rough set theory proposed by Pawlak cannot deal with numerical data, Hu et al. discussed a rough set model based on neighborhood granulation [14].

Given a decision system $\langle U, C, D\rangle$, where U is a finite nonempty set of the objects, $C = \{a_1, ..., a_m\}$ is a set of condition features describing the objects, and D is a decision feature that indicates the classes of the objects.

Definition 1 [14]**.** Given $x_i \in U$ and $B \subseteq C$, the neighborhood $\delta_B(x_i)$ of x_i with respect to B is defined as $\delta_B(x_i) = \{x_j | x_j \in U, \triangle_B(x_i, x_j) \leq \delta\}$, where $\triangle$ is a metric function. $\forall x_1, x_2, x_3 \in U$, it satisfies