Knowledge-Based Systems xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Multi-granularity feature selection on cost-sensitive data with measurement errors and variable costs

Shujiao Liao[*,a,b], Qingxin Zhu[b], Yuhua Qian[c], Guoping Lin[a]

[a] School of Mathematics and Statistics, Minnan Normal University, Zhangzhou 363000, China
[b] School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China
[c] Institute of Big Data Science and Industry, Shanxi University, Taiyuan 030006, China

ARTICLE INFO

Keywords:
Feature-granularity selection
Measurement errors
Multi-granularity
Neighborhood
Rough sets
Variable costs

ABSTRACT

In real applications of data mining, machine learning and granular computing, measurement errors, test costs and misclassification costs often occur. Furthermore, the test cost of a feature is usually variable with the error range, and the variability of the misclassification cost is related to the object considered. Recently, some approaches based on rough sets have been introduced to study the error-based cost-sensitive feature selection problem. However, most of them consider only single-granularity cases, thus are not feasible for the case where the granularity diversity between different features should be taken into account. Motivated by this problem, we propose a multi-granularity feature selection approach which considers measurement errors and variable costs in terms of feature-value granularities. For a given feature, the feature-value granularity is evaluated by the error confidence level of the feature values. In this way, we build a theoretic framework called confidence-level-vector-based neighborhood rough set, and present a so-called heuristic feature-granularity selection algorithm, and a relevant competition strategy which can select both features and their respective feature-value granularities effectively and efficiently. Experiment results show that a satisfactory trade-off among feature dimension reduction, feature-value granularity selection and total cost minimization can be achieved by the proposed approach. This work would provide a new insight into the cost-sensitive feature selection problem from the multi-granularity perspective.

1. Introduction

Feature selection is one of the most frequently-used techniques in data mining, machine learning and granular computing [4,14,19,42,65]. A dataset often contains many features, thus posing great difficulty in processing. By using the feature selection technique, irrelevant or redundant features can be removed to reduce the data complexity. Consequently, the efficiency of data processing can be improved significantly [10]. Rough set theory [26,34,35,41] is a powerful mechanism to handle uncertain data. Feature selection is also called attribute reduction in rough set society [16,30,39,50].

Cost-sensitive learning has received much attention in data mining and machine learning [1,43,54,60,66,67]. Among various kinds of cost in cost-sensitive learning [48], test cost (also called feature cost) and misclassification cost are the most commonly considered. Usually, the feature values of an object could not be obtained for free. Test cost refers to the money, time, or other resources consumed in acquiring a data item of the object. In addition, an object may be misclassified into a class that it does not belong to. Misclassification cost is the penalty paid for the wrong decision. Cost-sensitive feature selection, also called cost-sensitive attribute reduction in rough set community, aims at finding a feature subset to minimize some types of cost and meanwhile to keep the properties of original decision system as many as possible [15,21,25,31,44,46,61].

In practical applications, it is hard to obtain the accurate value of a data item because the measurement errors are ubiquitous and ineffaceable. For a quantity in reality, its measurement errors usually satisfy a normal (or near-normal) distribution. The existence of measurement errors poses great difficulty in distinguishing two objects if their measured values are close to each other. In view of this problem, some researchers have addressed objects in groups instead of addressing them individually [2,12]. The groups are referred to as information granules. Objects with measured values closing to each other are drawn into the same granule. In this case, the sizes of information granules are related to the error ranges, or equivalently, the lengths of error intervals. Granularity selection, namely selecting the sizes of information granules, plays an important role in granular computing [63].

Recently, three main kinds of approaches have been presented to

* Corresponding author at: School of Mathematics and Statistics, Minnan Normal University, Zhangzhou 363000, China.
E-mail addresses: sjliao2011@163.com (S. Liao), qxzhu@uestc.edu.cn (Q. Zhu), jinchengqyh@126.com (Y. Qian), guoplin@163.com (G. Lin).

https://doi.org/10.1016/j.knosys.2018.05.020
Received 30 June 2017; Received in revised form 13 May 2018; Accepted 16 May 2018

study the error-based cost-sensitive feature selection problem by using the rough set theory. The first kind [6,32] considers only test costs but not misclassification costs; the second kind [62,63] considers both test costs and misclassification costs, and the two types of costs are assumed to be fixed values; while in the third kind of approaches [23,64], both types of costs are seen to be variable, and all features are supposed to have the same feature-value granularity, namely have the same data precision. Unfortunately, these approaches are not feasible in some real-world scenarios. Firstly, test costs and misclassification costs often occur simultaneously in real applications, thus it is more realistic to take the two types of costs into consideration. Secondly, acquiring fine-grained data items usually costs more than acquiring coarse-grained ones, so the test cost of a feature is often monotonically decreasing with the enlargement of the feature values' error range. While the variability of the misclassification cost depends on the environment involved and the object considered. Taking the risk evaluation of granting credit as an example, if a customer is misclassified, both the cost (also called benefit if the cost is negative) of the customer and that of the finance company are usually constants. Taking the medical diagnosis as another example, for the misdiagnosis of a specific disease, the misclassification cost of the patient is often fixed, but that of the doctor is usually monotonically increasing with the total test cost paid by the patient. Concretely, if the patient is misdiagnosed with a total test cost of $100, he may require just a little compensation, namely the misclassification cost of the doctor is low; whereas if the patient is misdiagnosed with a total test cost of $1000, the misdiagnosis may make him angry and result in high misclassification cost of the doctor. Finally, in many real applications, different features may have different feature-value granularities, namely have different data precision. For example, electrocardiogram and color ultrasound are two different medical check-up items. Their metrics are not the same; naturally, the precision requirements are not necessarily identical for them. Therefore, the granularity diversity between different features, also called the multi-granularity characteristic of features, should be discussed in the research. However, most of existing rough-set-based feature selection approaches are essentially single-granularity approaches.

In actual applications, for a given dataset, if more necessary features are selected, or feature-value granularities get smaller (in this case, the similarity among the objects in each granule is enhanced), the total test cost will increase, while the misclassification rate will usually decrease. In this case, people cannot intuitively know how the total cost will change. Accordingly, it is complicated but important to choose suitable features and their corresponding feature-value granularities to achieve a trade-off between test costs and misclassification costs so that the total cost is as small as possible. Moreover, except the above-mentioned error-based cost-sensitive feature selection approaches, some existing papers of cost-insensitive feature selection [12,13,51] also addressed the granularity of feature values, but most of them have not taken the diversity between different features into consideration. Multi-granulation rough sets, which deal with multiple binary relations on the universe, have been studied extensively in recent years [18,27,38,53,56,59], but they have not touched the multi-granularity characteristic of features in the feature selection. Based on the above considerations, we introduce multi-granularity ideas into the cost-sensitive feature selection in this study.

In this paper, based on measurement errors and variable costs, we propose a multi-granularity feature selection approach to deal with the relationship among feature dimension, feature-value granularities and total cost. The approach aims at finding a suitable pair of feature subset and feature-value granularity vector to minimize the average total cost (the average value of total cost for the objects in the universe), and at the same time, to preserve the information of original decision system as much as possible. Differing from the previous methods, in the proposed approach the feature-value granularities between different features are not necessarily the same, thus we call the approach multi-granularity feature selection. Owing to the variability consideration of test costs and misclassification costs as well as the diversity of feature-value granularities between different features, the proposed approach is more versatile and practical than the existing error-based cost-sensitive feature selection approaches. Moreover, since most previous feature selection approaches, no matter whether cost-sensitive or cost-insensitive, are single-granularity in essence, this study would provide a new insight into the feature selection problem from the multi-granularity perspective.

In the proposed approach, for a given feature, the feature-value granularity is evaluated by the confidence level of the feature values' measurement errors. The measurement errors are assumed to satisfy a normal distribution, and the confidence level refers to the frequency that an observed interval contains a specific error value. So the confidence level is closely related to the data precision. In this context, we construct a confidence-level-vector-based neighborhood rough set model, in which features and their respective feature-value granularities are associated effectively. Under the new model, some fundamental concepts in neighborhood rough sets are redefined and discussed, such as the neighborhood granule, the lower and upper approximations, and the positive region. These concepts are closely relevant to the given feature subset and its corresponding confidence level vector. Moreover, some important properties in this model are also presented, such as three types of monotonicity in respect of the above-mentioned concepts. Then, some types of variable cost settings are introduced according to reality, in which the relationship among feature-value granularities, test costs and misclassification costs is considered. We also discuss how to compute the average total cost for any given feature-granularity pair (the pair of features and their respective feature-value granularities). Finally, we formally define the multi-granularity feature selection problem which takes measurement errors and variable costs into consideration.

A heuristic feature-granularity selection (the selection of features and their respective feature-value granularities) algorithm and a relevant competition strategy are proposed to deal with the multi-granularity feature selection problem. An addition-deletion strategy is adopted in the heuristic algorithm. Concretely, in the addition phase of the algorithm, for a given feature and its corresponding error confidence level, a feature-granularity significance (the significance of a feature and its feature-value granularity) function is designed by combining the size of incremental positive region with a δ-weighted test-cost-related value. The weight δ is set by the user to adjust the influence of the test cost to the feature-granularity significance. According to the significance values, best features and their corresponding best confidence levels are selected step by step. It is worthwhile to note that the above-mentioned monotonicities of the fundamental concepts in the confidence-level-vector-based neighborhood rough set model are fully used to make the process more efficient. Then in the deletion phase, the redundant feature-granularity elements (a feature-granularity element refers to a feature and its associated confidence level in the selected feature-granularity pair) are deleted to guarantee that the remaining feature-granularity pair has the minimal total cost. As for the competition strategy, it is presented to run the heuristic algorithm with different δ values and choose the best result. By using it, the users need not know the best setting for the weight δ in advance. Finally, some evaluation metrics are developed to study the performance of the proposed approach.

To evaluate the performance of the multi-granularity feature selection approach, a series of detailed experiments are undertaken on nine datasets from the UCI (University of California – Irvine) library [3]. Experimental results demonstrate that a satisfactory trade-off among feature dimension reduction, feature-value granularity selection and total cost minimization can be achieved by the approach. Both features and their respective feature-value granularities, which are often not the same between different features, can be obtained simultaneously through using the approach. This cannot be achieved by using the previous methods. The proposed multi-granularity approach