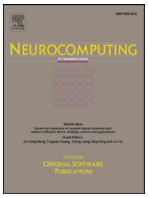
Neurocomputing 318 (2018) 271–286

Contents lists available at ScienceDirect

Neurocomputing

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



Feature selection for multi-label learning based on kernelized fuzzy rough sets

Yuwen Li[a], Yaojin Lin[b,c,*], Jinghua Liu[a], Wei Weng[a], Zhenkun Shi[d,e], Shunxiang Wu[a,*]

[a] *Department of Automation, Xiamen University, Xiamen 361000, PR China*
[b] *School of Computer Science, Minnan Normal University, Zhangzhou 363000, PR China*
[c] *Key Laboratory of Data Science and intelligence Application, Fujian Province Unversity, PR China*
[d] *College of Computer Science and Technology, Jilin University, Changchun 130012, PR China*
[e] *Key Laboratory of Symbolic Computation and Knowledge Engineering, Ministry of Education, Jilin University, Changchun, 130012 PR China*

ARTICLE INFO

Article history:
Received 13 February 2018
Revised 14 June 2018
Accepted 20 August 2018
Available online 31 August 2018

Communicated by Dr. Ivor Tsang

Keywords:
Feature selection
Fuzzy rough set
Multi-label learning
Multiple kernel learning

ABSTRACT

Feature selection is an essential pre-processing part in multi-label learning. Multi-label learning is usually used to deal with many complicated tasks, in which each sample is associated with multiple labels simultaneously. Fuzzy rough set model is one of the most effective ways for multi-label learning. However, it treats feature space and label space separately, and only uses features to describe sample structure information. In this paper, we fully consider the internal correlation between feature space and label space while fusing kernelized information from respective spaces. Moreover, we integrate fuzzy rough set with multiple kernel learning to finally realize feature selection. To be specific, firstly, we leverage one kind of kernel function to reveal the similarity between samples in feature space, and another one to assess the degree of label overlap between samples in label space. Secondly, we combine the kernelized information from the two spaces through linear combination to achieve precisely the lower approximation and construct a robust multi-label kernelized fuzzy rough set model, called RMFRS in this paper. Meanwhile, we discuss its properties and give theoretical analysis. Finally, we define a measurement criterion for selecting optimal features to evaluate the performance of the proposed algorithm. As many as 10 publicly available data sets are used to validate the effectiveness of our methods, and the result shows a distinct advantage over the state-of-the-art.

1. Introduction

Multi-label learning [45,47] provides an important technique for a number of applications, such as image recognition [16], text categorization [27], diagnosis code assignment [32], and computer vision [34]. In multi-label learning, samples can be linked to a set of labels, which characterizing the multiple semantic meanings. As we know, a huge number of features are stored in multi-label data, and some features may be irrelevant and/or redundant, which usually degrade the performance of learning. Therefore, multi-label learning is also affected by the curse of dimensionality [33]. Feature selection [3,15,22] is an essential pre-processing part in multi-label classification which can mitigate the curse of dimensionality. Therefore, some feature selection algorithms are designed to improve the classification performance by selecting an optimal feature subset from the original feature set. Moreover, feature selection can keep most useful information of the data set, maintain the physical meaning of original feature set, and provide better model readability and interpretability.

Feature selection with multi-label data [20,46] can be divided into three classes: filter, wrapper, and embedded. Filter [2] utilizes some measure criteria to sort features by using a greedy search process. The key factor of the filter is to construct an optimal criterion to evaluate the quality of the candidate features, such as mutual information [18,19,40] in information theory, dependency [6] and fuzzy dependency [11,30] in different rough set models. Rough set theory [25] has become an increasing need for expanding the application of feature selection [36] and rule learning [5]. However, the classical rough set model can't operate the hybrid features effectively. The fuzzy rough set (FRS) theory [31] is proposed to deal with this problem. In the framework of FRS, the fuzzy upper and lower approximation operators are defined according to fuzzy similarity relations. Then, we can obtain fuzzy dependency directly. Finally, as an evaluation metric, the fuzzy dependency is used to select feature subset. It was reported, however, that FRS model is sensitive to noisy samples. To alleviate this shortcoming, robust FRS models were developed.

[*] Corresponding author.

https://doi.org/10.1016/j.neucom.2018.08.065
0925-2312/© 2018 Elsevier B.V. All rights reserved.

Existing framework of robust FRS models can be roughly classified into two groups: in the first group, samples locate around classification boundary that are considered as noisy samples, such as β-precision FRS (β-PFRS) [8], probabilistic variable precision FRS (P-VP-FRS) [1], soft FRS (SFRS) [12], k-trimmed FRS [13] and data-distribution-aware FRS (PFRS) [2]. The other group uses robust approximation operators, such as k-median FRS [13], k-means FRS [13], fuzzy variable precision rough set (FVPRS) [44], and vaguely quantified rough set (VQRS) [4].

Both classical rough sets model and FRS model cannot exactly handle with heterogeneous data because two of them share the previously mentioned argument of transforming feature types. In recent years, for the sake of feature selection with heterogeneous data, much progress has been made [3,10]. Conspicuously, multi-kernel learning has been considered to integrate with FRS model. Broadly speaking, it becomes essential to map heterogeneous features into a unified representation framework [10,11]. Furthermore, FRS and kernel method are two general techniques. FRS takes advantage of fuzzy relations to granulate the universe. Evidently, kernel method maps data into a higher dimensional feature space. Although it seems there is no relation between FRS and kernel method, two methods have the key connection that they represent sample information with the same format. Theoretically, kernel matrices are used in kernel functions and relation matrices are utilized in FRS model [14]. Different kernel functions are put forward to compute the similarity between samples described by different types of features. The Gaussian kernel is employed to quantify the similarity between samples on numerical data [28]; a match kernel is utilized to induce an equivalence relation on symbolic data [24]. In recent years, many studies about multi-kernel learning have been reported on feature selection [10,14,21], and classification [26,41]. In 2011, Hu et al. defined kernelized FRS by using kernel functions to extract fuzzy relations. In these models, a key link is built up between kernel method and FRS [11].

Recently, in multi-label feature selection studies, one of the major strategies transforms a multi-label feature selection task to several binary single-label feature selection tasks [35,42,43,48], called problem transformation. But it cuts up the relationship between labels and easily generates unbalanced data. Then, algorithm adaption method [18–20,40,46] is the other strategy for multi-label feature selection. It solves the label overlapping and improves prediction results by adapting or extending existing single-label algorithms, rather than transforming the data. In this approach, a feature subset is obtained by the optimization of a certain criterion, such as a joint learning criterion that involves simultaneous feature selection and multi-label learning [46,49–52]. From the viewpoint of empirical risk minimization, Sun [49] selected the most discriminative features for all labels by using the $l_{2,\,1}$-norm as a certain criterion for both loss function and regularization. In addition, considering missing labels, Zhu [46] imposed the effective $l_{2,\,p}$-norm $(0 < p \leq 1)$ regularization item on the feature selection matrix to remove the irrelevant and noisy features from original feature set. Huang [50] considered that each class label might be determined by some specific characteristics of its own. Then, label-specific features for each class label are selected and composed a label-specific feature subset. Lee [51] constructed a scalable relevance evaluation criterion to evaluate conditional relevance more accurately and obtain an effective feature subset. Li [52] used a maximal correlation minimal redundancy as a criterion to propose a granular feature selection method for multi-label learning based on mutual information. This criterion can make sure that the selected feature subset contains the most class-discriminative information. Granular computing can be proved that it is an effective computing paradigm of information processing. Fortunately, FRS is also a kind of granular computing method. Therefore, some multi-label feature selection methods based on FRS have been presented and discussed [35,42,43], and the common characteristic among them is that these algorithms deal with feature space and label space separately and only reflect sample structure information by features. These existing multi-label feature selection approaches ignore the connection between feature space and label space. Especially, Zhang [42] transformed a multi-label learning task to several binary single-label tasks and then computed the average score of the features across all single-label tasks. In fact, this method belongs to problem transformation mentioned above. This method omits the relationship between labels and produces too many new labels leading to feature selection with significant difficulties. On the contrary, the proposed method belongs to algorithm adaption methods. We take the relationship between labels into consideration and construct multi-label FRS model by extending existing single-label FRS model, rather than transforming the data like [42]. Therefore, in this paper, we aim to perform multi-label feature selection by treating labels as a whole space as well as feature space by utilizing kernel functions, as shown in Fig. 1. To this end, we integrate FRS with multiple kernel learning to finally realize feature selection. Based on FRS model, it is an important issue to find the nearest different classes' sample for a given sample in multi-label learning with FRS. As a matter of course, we divide this issue into two key points, “nearest” and “different classes”. Firstly, feature space exploits Gaussian kernel to reveal “nearest” through the similarity between samples in feature space. Similarly, label space uses match kernel to reveal “different classes” through the label overlap ratio between samples in label space. Secondly, we combine the kernelized information from the two spaces through linear combination to achieve precisely the lower approximation and construct a robust multi-label kernelized FRS model, called RMFRS in this paper. Thirdly, we define a fuzzy dependency functions in the learning tasks and present a forward greedy feature selection algorithm. Finally, extensive experiments are carried on to make clear the effectiveness of the proposed algorithm. The major contributions of the proposed model can be summarized as follows:

- Integrate fuzzy rough set with multiple kernel learning to finally realize feature selection.
- Construct a fused kernel space for multi-label learning, which combines the kernelized information from label space and feature space. In this fused space we can calculate the lower approximation and assessment the importance of features.
- Propose a novel multi-label kernelized FRS model, which has a good robustness.
- Evaluate RMFRS extensively using 10 different open data sets to understand the working of RMFRS.

The rest of the paper is organized as follows. Section 2 briefly recalls several preliminary concepts. Section 3 describes a kernel method in multi-label learning. Section 4 introduces how we designed the feature selection algorithms based on a novel multi-label kernelized FRS models. Section 5 presents experimental settings and results. Section 6 concludes this study with future work.

2. Preliminaries

Given a nonempty universe U, R is a fuzzy equivalence relation if it satisfies reflexivity, symmetry, and sup-min transitivity. The fuzzy equivalence class $[x]_R$ is generated by a fuzzy equivalence relation R with respect to sample $x \in U$. $[x]_R$ is a fuzzy set on U, which is also referred as the fuzzy neighborhood of x, i.e., $[x]_R(y) = R(x, y)$ for all $y \in U$.

Definition 1 [6]. Given a nonempty universe U, $A(U)$ is the fuzzy power set of U and R is a fuzzy binary relation on U. Let $A \in A(U)$ be a fuzzy set, and the lower and upper approximations of x with