Applied Soft Computing 38 (2016) 244–256

Contents lists available at ScienceDirect

Applied Soft Computing

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

Multi-label feature selection based on neighborhood mutual information

Yaojin Lin[a,b,*], Qinghua Hu[a,*], Jinghua Liu[b], Jinkun Chen[c], Jie Duan[a]

[a] School of Computer Science and Technology, Tianjin University, Tianjin 300072, PR China
[b] School of Computer Science, Minnan Normal University, Zhangzhou 363000, PR China
[c] School of Mathematics and Statistics, Minnan Normal University, Zhangzhou 363000, PR China

ARTICLE INFO

Article history:
Received 11 April 2015
Received in revised form 8 September 2015
Accepted 1 October 2015
Available online 20 October 2015

Keywords:
Feature selection
Multi-label learning
Neighborhood
Neighborhood mutual information

ABSTRACT

Multi-label learning deals with data associated with a set of labels simultaneously. Like traditional single-label learning, the high-dimensionality of data is a stumbling block for multi-label learning. In this paper, we first introduce the margin of instance to granulate all instances under different labels, and three different concepts of neighborhood are defined based on different cognitive viewpoints. Based on this, we generalize neighborhood information entropy to fit multi-label learning and propose three new measures of neighborhood mutual information. It is shown that these new measures are a natural extension from single-label learning to multi-label learning. Then, we present an optimization objective function to evaluate the quality of the candidate features, which can be solved by approximating the multi-label neighborhood mutual information. Finally, extensive experiments conducted on publicly available data sets verify the effectiveness of the proposed algorithm by comparing it with state-of-the-art methods.

1. Introduction

In classical supervised learning, each instance only belongs to one label relative to a number of candidate labels. In many real-world applications, however, one instance is usually associated with multiple concepts simultaneously [8,7,2,27,45,44]. For example, a newspaper article concerning the reactions of the scientific circle to the release of the Da Vinci Code film can be classified into any of the three classes: arts, science, and movies; an image showing a tiger in woods is associated with several keywords such as trees and tiger. As we know, one label per object is unable to fully describe such scenario, and therefore the research on multi-label classification task has attracted increasing interest [11,20,21,31,48,50,52,62]. In which, Zhang et al. [61] presented a profound review on multi-label learning algorithms, which includes the fundamentals on multi-label learning, eight representative algorithms, and several related learning settings.

In many real-world applications, the multi-label data usually have thousands or even tens of thousands of features [15,22,61]. This is a more common characteristic in image annotation and text categorization. For example, millions of informative words are extracted from a collection of documents or web pages to represent their topics. Also, from an image thousands of features are extracted to reflect its all kinds of semantics. Generally speaking, many features are redundant and/or irrelevant for a given learning task, and high dimensional data may brings many disadvantages to learning algorithms, such as computational burden, over-fitting, and poor performance [4,14,17,18,29,33,46,47]. To solve this problem, a number of dimensionality reduction based multi-label learning methods have been presented. Those methods can be grouped into two categories: multi-label feature extraction and multi-label feature selection. Multi-label feature extraction is a method that converts original high-dimensional feature space into a new low-dimensional feature space through transforming or mapping, and the new constructed features are usually combinations of original features. However, it is difficult to link the features from original feature space to new features. At present, some popular feature extraction methods have been proposed, such as Partial Least Squares (PLS) [49], Linear Discriminant Analysis (LDA) [16], Canonical Correlation Analysis (CCA) [10], and multi-label informed latent semantic indexing (MLSI) [56]. To sum up, the characteristics of feature extraction include (1) the results of feature extraction are

[*] Corresponding authors. Tel.: +86 13960044089.
E-mail addresses: yjlin@mnnu.edu.cn (Y. Lin), huqinghua@tju.edu.cn (Q. Hu), zzliujinghua@163.com (J. Liu), cjk99@163.com (J. Chen), duanjie@tju.edu.cn (J. Duan).

http://dx.doi.org/10.1016/j.asoc.2015.10.009
1568-4946/© 2015 Elsevier B.V. All rights reserved.

lack of interpretation; (2) feature extraction blurs the information of original features and loses physical interpretation.

Different from multi-label feature extraction, multi-label feature selection selects the feature subset from the original feature space directly, and keeps the physical meaning for the selected features. Multi-label feature selection methods are usually classified into three main groups: filter, wrapper, and embedded [37,42,43]. The filter approach separates feature selection from classifier learning [39,58]. The wrapper approach uses the predictive accuracy of a predetermined learning algorithm to determine the quality of selected features [12,60]. The embedded approach achieves model fitting and feature selection simultaneously [15]. As we know, the key step of the filter approach is to design effective metrics to evaluate the quality of the candidate features, such as mutual information [6,23–25], dependency [58], and the classification margin [40,41]. As for mutual information, Lee et al. [23] proposed a multivariate mutual information based feature selection method for multi-label classification, which selects an effective feature subset via maximizing the multivariate mutual information between the selected features and labels. In [27], information gain between a feature and label set is exploited to measure the importance of the feature and label correlation. Yu et al. [57] proposed a multi-label feature selection algorithm based on mutual information and genetic algorithm. In addition, the mutual information measure is applied in [6] according to a modified LP approach [35], which considers label dependence.

In order to compute mutual information for hybrid data, we should know the probability distributions of variables and their joint distributions. However, these distributions are not known in advance. In addition, the process of discretization easily loss useful information. Therefore, Hu et al. [17] presented an assumption that samples with the similar feature values should be classified into the same class or neighborhood class. Based on this assumption, the equivalent relation is extended into neighborhood relation [19,51,54], where neighborhood, computed with distance, is looked as the subset of instances which have the similar feature values with the centroid. Moreover, Hu et al. [17] integrated the concept of neighborhood into Shannon's information theory [38], and proposed a new information measure, called neighborhood information entropy. Then, joint neighborhood entropy, conditional neighborhood entropy, and neighborhood mutual information can be defined directly. However, these concepts cannot be used to multi-label learning directly. Different from single-label learning, each instance belongs to a set of labels in multi-label learning. Therefore, we need redefine the concept of neighborhood information entropy and its relative concepts. In this work, we generalize neighborhood entropy in single-label learning to fit multi-label learning, and propose three new measures of neighborhood mutual information, which can be used to evaluate the quality of the candidate features.

Our work is focused on three problems. First, we introduce the margin of instance to granulate all instances under different labels. Meanwhile, we present three different cognitive viewpoints, i.e., optimistical viewpoint, neutral viewpoint, and pessimistic viewpoint. Based on these viewpoints, three kinds of neighborhood for multi-label learning are introduced, and the new definitions on neighborhood information entropy and neighborhood mutual information are proposed. Second, we discuss the problem how to use the proposed measures in multi-label feature selection. In which, we present an optimization objective function to evaluate the quality of the candidate features, which can be solved by approximating multi-label neighborhood mutual information. This solution has the potential of being a general strategy to multi-label feature selection. Finally, a comprehensive set of experiments is conducted to show the effectiveness of our proposed method. The main contributions of this paper can be summarized as follows:

- Different from the traditional multi-label feature selection, the proposed algorithm derives from different cognitive viewpoints.
- A simple and intuitive metric to evaluate the candidate features is proposed.
- The proposed algorithm is applicable to both categorical and numerical features.
- Our proposed method outperforms some other state-of-the-art multi-label feature selection methods in our experiments.

The rest of this paper is organized as follows. Section 2 introduces multi-label learning and neighborhood mutual information. Then, we present the multi-label feature selection based on multi-label neighborhood mutual information method and report on experimental evaluations in Sections 3 and 4, respectively. Finally, our conclusions are given in Section 5.

2. Preliminaries

2.1. Multi-label learning

In multi-label learning with m labels, $X \subset \Re^d$ denotes a multi-label data set, and $x \in X$ is represented as a d-dimensional vector $\mathbf{x}=[x_1, x_2, \ldots, x_d]$. Let $L=\{l_1, l_2, \ldots, l_m\}$ be a set of labels. Each data point is associated with a subset of L, and this subset can be described as a m-dimensional vector $\mathbf{y}=[y^1, y^2, \ldots, y^m]$ where $y^j=1$ only if x has label l_j and 0 otherwise.

In multi-label classification learning, the evaluation functions are different from the traditional single-label classification learning ones. In experimental evaluation, we select some measures proposed in [36]. Let $T=\{(x_i, y_i)|1 \le i \le N\}$ be a given testing set where $y_i \subseteq L$ is a correct label subset, and $Y_i' \subseteq L$ be the binary label vector predicted by a multi-label classifier for instance x_i.

Average Precision (AP): this measure evaluates the average fraction of labels ranked above a particular label $\gamma \in y_i$, which is actually in y_i. The formula for AP is

$$AP = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|y_i|}\sum_{\gamma \in y_i}\frac{|\{\gamma' \in y_i : r_i(\gamma') \le r_i(\gamma)\}|}{r_i(\gamma)}$$

where $r_i(l)$ stands for the rank of label $l \in L$ predicted by the algorithm for a given instance x_i. The bigger the value, the better the performance.

Coverage (CV): this measure evaluates how far, on average, we need to go down the label ranking list to cover all the ground-truth labels of the instance. CV is defined as follows

$$CV = \frac{1}{N}\sum_{i=1}^{N}\max_{\lambda \in y_i} rank(\lambda) - 1$$

where $rank(\lambda)$ denotes the rank list of λ according to its likelihood, for example, if $\lambda_1 > \lambda_2$, then $rank(\lambda_1) < rank(\lambda_2)$. In the case of coverage smaller value shows better performance.

Hamming Loss (HL): this measure evaluates how many times an instance-label pair is misclassified. HL is

$$HL = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i' \oplus y_i|}{M}$$

where $\oplus$ denotes the XOR operation. Here the smaller value denotes better performance.