Pattern Recognition 48 (2015) 2761–2771

Contents lists available at ScienceDirect

Pattern Recognition

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

Fast multi-label feature selection based on information-theoretic feature ranking

Jaesung Lee, Dae-Won Kim*

School of Computer Science and Engineering, Chung-Ang University, 221, Heukseok-Dong, Dongjak-Gu, Seoul 156-756, Republic of Korea

ARTICLE INFO

Article history:
Received 30 June 2014
Received in revised form
9 March 2015
Accepted 5 April 2015
Available online 15 April 2015

Keywords:
Multi-label feature selection
Mutual information
Interaction information
Entropy

ABSTRACT

Multi-label feature selection involves selecting important features from multi-label data sets. This can be achieved by ranking features based on their importance and then selecting the top-ranked features. Many multi-label feature selection methods for finding a feature subset that can improve multi-label learning accuracy have been proposed. In contrast, computationally efficient multi-label feature selection methods have not been studied extensively. In this study, we propose a fast multi-label feature selection method based on information-theoretic feature ranking. Experimental results demonstrate that the proposed method generates a feature subset significantly faster than several other multi-label feature selection methods for large multi-label data sets.

1. Introduction

Multi-label feature selection is useful for reducing the computational burden of learning, while maintaining or possibly improving accuracy. It has been used widely in application areas such as music emotion recognition, gene function classification, semantic image annotation, and text categorization [25,26,19,3,7,31]. Let $W \subset \mathbb{R}^d$ denote input data constructed from a set of features F, where $|F| = d$ and patterns drawn from W are assigned to a joint state of multiple labels $L = \{l_1, \dots, l_t\}$, where $|L| = t$. Multi-label feature selection can be achieved through a ranking process of assessing the importance of d features based on a score function and selecting the top-ranked n features from F ($n \ll d$).

Several researchers have dedicated their efforts to selecting important features for multi-label learning [4,8,10,15,21,28]. Multi-label feature selection methods can be categorized into three types, wrapper, embedded, and filter approaches, according to how they assess the importance of candidate feature subsets. Wrapper-based multi-label feature selection methods assess the importance of feature subsets based on the accuracy of multi-label learning algorithm [35]. Some multi-label learning algorithms have a feature selection process embedded in their learning process [10,13,23]. In contrast, filter-based multi-label feature selection methods find a feature subset by focusing on the characteristics of candidate feature subsets and multiple labels [15,16,18,29]. Although multi-label problems can be solved in a simpler manner by assuming that labels are independent to each other [17,29], label dependency is considered to be a key factor in determining a better feature subset [6,20,38]. Because multi-label feature selection can boost the efficacy of label dependency by discarding noisy features, it is regarded as an effective method for multi-label learning [16,18,35]. To consider label dependency, an algorithm must examine various label combinations from input labels [36]. Therefore, considering label dependency can be computationally prohibitive when the number of input labels is large. However, most multi-label feature selection methods focus on improving multi-label learning accuracy solely; and hence, research on fast multi-label feature selection is still lacking.

In this paper, we propose a multi-label feature selection method with a concern for computational efficiency. To demonstrate the efficiency issue of multi-label feature selection theoretically, we derive a score function based on information theory for assessing the importance of each feature [5,12,27] and then analyze it in terms of computational cost. A derived score function indicates that significant computational cost will be expended to calculate the entropy involved in the interaction of information terms. To circumvent this efficiency issue, we propose an efficient feature ranking method based on three components:

- Relaxing the derived score function by constraining the maximum size of label combinations to be considered.
- Discarding unnecessary entropy calculations for feature ranking and reusing pre-calculated entropy terms.
- Identifying promising labels for considering label dependency.

* Corresponding author. Tel.: +82 2 820 5304; fax: +82 2 820 5301.
E-mail address: dwkim@cau.ac.kr (D.-W. Kim).

http://dx.doi.org/10.1016/j.patcog.2015.04.009
0031-3203/© 2015 Elsevier Ltd. All rights reserved.

Thus, the contribution of this work can be summarized as follows:

- A computationally efficient score function for the multi-label feature selection problem is proposed. This function is obtained by discarding unnecessary or ineffective entropy calculations from the score function of Lee and Kim [18].
- The proposed score function considers promising label combinations based on the information-theoretic perspective, while the score function in the previous study considers all the label-pairs, irrespective of the expense.
- The actual benefit provided by each component is shown by a mathematical analysis based on the computational cost for entropy calculation. The previous study does not present such an analysis.

Experimental results indicate that the proposed method outputs a feature subset significantly faster than other multi-label feature selection methods for large multi-label data sets.

2. Proposed method

In this section, we propose our multi-label feature selection method based on efficient feature ranking. First, the score function is derived from Shannon's mutual information (MI) [27] between a feature and labels. To circumvent examining all possible combinations of labels, the derived score function is relaxed in Section 2.1. Second, to make the score function computationally less expensive, methods of avoiding unnecessary and redundant calculations for entropy terms are introduced in Sections 2.2 and 2.3, respectively. Third, to reduce computational cost for considering label combinations, a strategy for identifying promising labels is proposed in Section 2.4. An algorithmic sketch of the proposed method is presented in Section 2.5.

2.1. Deriving score function

To perform multi-label feature selection, an algorithm must be able to measure the dependency (importance score) between each feature and multiple labels. The dependency between a feature f and labels L can be measured using MI.

$$M(f;L)=H(f)-H(f,L)+H(L) \tag{1}$$

where $H(\cdot)$ of Eq. (1) represents a joint entropy that measures the extent of self-information carried by multiple variables, defined as

$$H(X)=-\sum P(X)\log\, P(X) \tag{2}$$

where $P(\cdot)$ is a probability mass function of a given set of variables X. Let S be a set of n features, S' a power set of S without $\{\phi\}$, and X a possible element of $S'_p=\{e\,|\,e\in S', |e|=p\}$. Because Eq. (1) suffers from estimating high-dimensional joint entropy, when $|L|$ is large, it can be rewritten using the work of Lee and Kim [18]:

$$M(S;L)=\sum_{k=2}^{|L|+n}\sum_{p=1}^{k-1}(-1)^k V_k(S'_p\times L'_{k-p}) \tag{3}$$

where $\times$ denotes the Cartesian products of two sets of variables, and $V_k(\cdot)$ is defined as

$$V_k(S')=\sum_{X\in S'_k} I(X) \tag{4}$$

where $I(X)$ is the interaction information (refer it to *interaction* in this paper) for a given variable set X [2,5,12], defined as

$$I(X)=-\sum_{Y\in X'}(-1)^{|Y|}H(Y) \tag{5}$$

where X' is a power set of X without $\{\phi\}$. It should be noted that MI takes into consideration the shared information between S and L, but it ignores the information that lies within S or L; however, interaction information considers the shared information of all the involved variables [18].

Eq. (3) was derived to consider dependencies among multiple features and multiple labels. In contrast, Eq. (1) considers dependency between a feature and multiple labels because of computational efficiency, hence Eq. (3) can be further simplified. Because S is f from Eq. (1), and S'_p of Eq. (3) represents a set of possible subsets of S with p cardinality, there is no element in S'_p, when $p>1$. Therefore, n and p in Eq. (3) are fixed to one. As a result, Eq. (3) is simplified to

$$M(S;L)=M(f;L)=\sum_{k=2}^{|L|+1}(-1)^k V_k(f'_1\times L'_{k-1}) \tag{6}$$

Because $X'_1=X$, where $|X|=1, f'_1$ in Eq. (6) can be simplified to f, whereby we get Eq. (7) as follows:

$$M(f;L)=\sum_{k=2}^{|L|+1}(-1)^k V_k(f\times L'_{k-1}) \tag{7}$$

Eq. (7) indicates that $M(f;L)$ can be separated into interaction terms involving a feature and all possible label combinations. For example, MI between f and $L=\{l_1,l_2,l_3\}$ can be rewritten as

$$\begin{aligned}M(f;L)=&I(f,l_1)+I(f,l_2)+I(f,l_3)-I(f,l_1,l_2)\\&-I(f,l_1,l_3)-I(f,l_2,l_3)+I(f,l_1,l_2,l_3)\end{aligned}$$

Eq. (7) also indicates that the number of interaction terms increases exponentially with the size of labels. To circumvent prohibitive computations, we relax the score function by constraining $|L|$ of (7) to b. This allows the score function to consider label combinations with a maximum b cardinality:

$$M_b(f;L)=\sum_{k=2}^{b+1}(-1)^k V_k(f\times L'_{k-1}) \tag{8}$$

For example, if we set b to one, then Eq. (8) can be written as

$$M_1(f;L)=\sum_{k=2}^{2}(-1)^k V_k(f\times L'_{k-1})=V_2(f\times L'_1)=\sum_{l_i\in L} I(f,l_i)$$

As a result, the score function will not consider dependency between a feature and label combinations. To circumvent this, b can be set to two to consider dependency between a feature and label-pairs:

$$\begin{aligned}M_2(f;L)&=\sum_{k=2}^{3}(-1)^k V_k(f\times L'_{k-1})=V_2(f\times L'_1)-V_3(f\times L'_2)\\&=\sum_{l_i\in L} I(f,l_i)-\sum_{l_i,l_j\in L} I(f,l_i,l_j)\end{aligned}$$

where $l_i\neq l_j$. These examples show that the computational cost is relaxed according to b. Because the calculation of interaction terms is performed by obtaining entropy terms involved in interaction terms, we analyzed Eq. (8) in terms of the computational cost of the entropy calculation. Let k be the number of variables involved in an entropy term. The number of patterns is a constant value, and hence, the computational cost for calculating an entropy term will increase linearly according to k; the number of values to be examined for calculating entropy is $|W|\cdot k$, where $|W|$ is the number of patterns in a given data set. For simplicity, we assume a computational cost of $H(X)$, where $|X|=k$ is k unit cost, with one unit cost being the computational cost to calculate an entropy term involving one variable. Assuming that an algorithm assesses