Pattern Recognition 66 (2017) 342–352

Contents lists available at ScienceDirect

Pattern Recognition

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

SCLS: Multi-label feature selection based on scalable criterion for large label set

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

Keywords:
Machine learning
Multi-label learning
Multi-label feature selection
Relevance evaluation
Conditional relevance

ABSTRACT

Multi-label feature selection involves the selection of relevant features from multi-labeled datasets, resulting in a potential improvement of multi-label learning accuracy. In conventional multi-label feature selection methods, the final feature subset is obtained by identifying the features of high relevance with low redundancy. Thus, accurate score evaluation is a key factor for obtaining an effective feature subset. However, conventional methods suffer from inaccurate conditional relevance evaluation when a large number of labels are involved. As a result, irrelevant features can be a member of the final feature subset, leading to low multi-label learning accuracy. In this paper, we propose a new multi-label feature selection method. Using a scalable relevance evaluation process that evaluates conditional relevance more accurately, the proposed method significantly improves multi-label learning accuracy compared with conventional multi-label feature selection methods.

1. Introduction

Multi-label classification is part of a base technique for recent applications, such as sentiment analysis of user texts [25,28] or tag classification of music clips [24,35,39,41], because texts and music clips can be associated with multiple concurrent labels [20,43]. In practice, applications can incur a series of labels for encoding the target concepts to be learned, especially when the target consists of multiple sub-concepts, such as humor or admiration [1,8,12]. Let $W \subset \mathbb{R}^d$ denote a set of patterns constructed from a set of features F. Then each pattern $w_i \in W$ where $1 \le i \le |W|$ is assigned to a certain label subset $\lambda_i \subseteq L$, where $L = \{l_1, \ldots, l_{|L|}\}$ and is a finite set of labels. Because there can be hidden relationships among these tags or labels that would improve multi-label learning accuracy, better performance can be achieved by exploiting useful relationships [36,44]. For this reason, multi-label feature selection can contribute to the improvement of learning accuracy by highlighting such relationships based on important features [17,33,42], and hence it is considered an important preprocessing step [16,18,19].

Given input data with an original feature set F and label set L, the goal of multi-label feature selection is to identify a feature subset $S \subset F$ with $n \ll |F|$ features that have the largest relevance on multiple labels [16,19]. Because S should support multiple labels simultaneously using only n features, the selection of a compact feature subset becomes an important task when L involves many labels [18,21,22,23]. To ensure the largest relevance on L with n features, each feature in S should carry individual information on labels [19]. If two features carry the same information, it becomes unnecessary to select one of them to compose S because this feature will not carry any additional discriminating power under the selection of the other feature. Thus, relevance evaluation that considers dependency among the selected features is important for identifying an effective feature subset.

Under the incremental selection for efficiently finding a near-optimal solution, the selection of the i-th feature from the set $\{F - S_{i-1}\}$, where S_{i-1} is a feature subset with $i-1$ features, is performed by identifying the f_i that maximizes the value of following the relevance criterion [16,19,21,23].

$$\max_{f_i \in \{F - S_{i-1}\}} [Rel(f_i) - Red(f_i)] \tag{1}$$

where $Rel(f_i)$ and $Red(f_i)$ denote the dependency of f_i to L and the dependency between f_i and the already selected features of S_{i-1}, respectively. Thus, the task can be solved by scoring each feature based on Eq. (1), and then including the top-ranked feature at each iteration [13,17,32]. In the multi-label feature selection method that employs Eq. (1), the algorithm attempts to avoid the selection of features carrying the same information that is given by already selected features based on $Red(f_i)$.

In previous studies, $Rel(f_i)$ is calculated as a large value because it is computed by adding the dependency values between f_i and each label. On the other hand, when the number of involved labels is large, $Red(f_i)$ implies too small values compared with $Rel(f_i)$ [21,22], or incurs

* Corresponding author.
E-mail address: dwkim@cau.ac.kr (D.-W. Kim).

http://dx.doi.org/10.1016/j.patcog.2017.01.014
Received 12 August 2016; Received in revised form 13 December 2016; Accepted 9 January 2017
Available online 10 January 2017
0031-3203/

erroneous calculation because of repetitive dependency computations along with each label [16,23]. As a result, irrelevant features can be included in the final feature subset because of inaccurate relevance evaluation. In this paper, we propose an effective multi-label feature selection method based on a new $Red(f_i)$ function that considers $Rel(f_i)$ within its calculation while avoiding erroneous dependency calculations, resulting in the improvement of multi-label learning accuracy.

2. Related work

In multi-label feature selection studies, one of the major trends includes the application of a feature selection method for single-label problems after transforming label sets into a single label [29,33]. In addition to the merits from the immediate use of conventional methods and their side effects [34], an algorithm adaptation strategy that directly manages multi-label problems has also been considered [36]. 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 [10,27], $l_{2,1}-$norm function optimization [26], label ranking error [9], Hilbert-Schmidt independence criterion [14], F-statistics [13], and label-specific feature selection [43]. In this paper, we focus on a mutual information-based multi-label feature selection method because its theoretical background and advantage have been thoroughly discussed in previous studies [6,16,17,21–23].

When mutual information is employed in its original form for evaluating feature relevance, the algorithm inevitably faces the problem of high-dimensional joint probability estimation caused by multiple labels in L [16,21]. Because the process often becomes impractical given the insufficient patterns and characteristics of particular labels [1,3], researchers have attempted to circumvent this difficulty by focusing on dependency among variable subsets [19,33], resulting in variations that provide each with a unique advantage against the characteristics of datasets and evaluation measures [18].

In the work of [16], the authors demonstrated that mutual information can be decomposed into the sum of dependencies among all possible variable subsets across S and L. To circumvent the intractable calculations, the dependency between features and the label set is approximated by considering each label and label pairs. In addition, the dependency among features is determined by adding the dependency of all combinations composed of two features and one label. Finally, the relevance of the feature is calculated by subtracting the dependency among the features from that to labels. A similar approach was also employed in the pieces of works of [19,23]. Thus, relevance evaluation is commonly based on calculations that involve repetitive dependency estimates for too many variable subsets along with given labels. As a result, the relevance evaluation process is unscalable to the number of labels because possible errors caused from the dependency estimation for each variable subset will be cumulated to the final relevance score.

If two features are mutually independent, these two features certainly have a different nature in terms of dependency on L. Based on this property, a criterion for relevance evaluation can be derived without incurring repetitive dependency calculations [21]. The same score function was employed for a quadratically programmed objective function for considering the global perspective of a selected feature subset [22]. However, these two methods commonly suffer from an incompact feature subset because feature dependency is determined regardless of the amount of dependency on labels, resulting in underestimation of feature dependency compared with the dependency on labels when a large number of labels are involved.

3. Proposed method

3.1. Limitations of previous studies

As Eq. (1) shows, the characteristics of the selected feature subset is strongly influenced by a relevance evaluation based on $Rel(f_i)$ and $Red(f_i)$. For example, if $Rel(f_i)$ is evaluated as too large compared with $Red(f_i)$, the influence of $Red(f_i)$ on the relevance evaluation eventually becomes small. In this case, the selected feature subset can be composed of features that are dependent on each other, resulting in low discriminating power within a fixed number of features. This undesirable situation can occur from conventional multi-label feature selection methods when the number of labels is large [21,22] because $Rel(f_i)$ increases as the number of labels grows, whereas $Red(f_i)$ is solely determined by the dependency on features in S_{i-1} [21,22].

To show this aspect more clearly, we conduct a preliminary analysis in this section. For demonstration purposes, we select a conventional multi-label feature selection method from the perspective of simplicity. In the work of [21,22], the i-th feature f_i is selected if it maximizes

$$\max_{f_i\in\{F-S_{i-1}\}}\left[\sum_{l\in L} M(f_i;l)-\sum_{f\in S_{i-1}} M(f_i;f)\right] \tag{2}$$

where $M(x;y)=H(x)-H(x,y)+H(y)$ is the mutual information between variables x and y, and $H(x)=-\sum P(x)\log P(x)$ is the joint entropy with their probability functions $P(x)$, $P(y)$, and $P(x,y)$. Eq. (2) indicates that $Rel(f_i)=\sum_{l\in L} M(f_i;l)$ and $Red(f_i)=\sum_{f\in S_{i-1}} M(f_i;f)$ are respectively implemented as the sum of mutual information terms: (1) between f_i and all the labels, and (2) between f_i and the already selected features in S_{i-1}. Thus, the number of mutual information terms considered in $Rel(f_i)$ is the same as the number of labels in L. In contrast, the number of the terms in $Red(f_i)$ is $|S_{i-1}|$. For example, let us consider $i=2$ where the algorithm selects the second feature. Because the number of mutual information terms in $Rel(f_i)$ is fixed to $|L|$, whereas that in $Red(f_i)$ is only one, the influence of $Red(f_i)$ on the relevance evaluation is small. As a result, a feature that is dependent on the already selected feature can be selected to compose S_2. The example shows that the feature subset with a small number of features can be composed of dependent features that are unhelpful to the improvement of multi-label learning accuracy. For multi-label feature selection, this is a serious problem because the number of features to be selected in n is typically set to a small number. Moreover, this example also indicates that the conventional method requires more features in order to attain multi-label learning accuracy to some extent because the dependent features could be members of the final feature subset, and non-influential on the improvement of multi-label learning accuracy. The reason for this result is that $Red(f_i)$ is determined regardless of $Rel(f_i)$, hence this problem can be solved by adjusting the influence of $Red(f_i)$ against $Rel(f_i)$, or vice versa. Although the $Red(f_i)$ term was calculated differently in the works of [16,19,23], they commonly suffer from too many complex relationships among the features and label combinations, resulting in inaccurate relevance evaluation caused by cumulative error from the probability estimation of dependency calculations.

In this paper, we propose a multi-label feature selection method based on a Scalable Criterion for Large Label Set (SCLS) aimed at identifying an effective feature subset for improving multi-label learning accuracy. The difference between SCLS and previous studies can be summarized as follows:

- SCLS is designed to use a simpler dependency calculation process. For example, in previous studies [16,19], $Red(f_i)$ is calculated as

$$Red(f_i)=\sum_{f\in S_{i-1}}\sum_{l\in L}\left(H(f_i)+H(f)+H(l)-H(f_i,f)-H(f_i,l)\right.$$
$$\left.-H(f,l)+H(f_i,f,l)\right) \tag{3}$$