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/