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 $L = \{l_1, \ldots, l_t\}$, 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/