Expert Systems with Applications 42 (2015) 2013–2025

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Mutual Information-based multi-label feature selection using interaction information

Jaesung Lee, Dae-Won Kim *

School of Computer Science and Engineering, Chung-Ang University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Available online 18 October 2014

Keywords:
Multi-label feature selection
Multivariate feature selection
Interaction information
Feature dependency

ABSTRACT

Multi-label feature selection is regarded as one of the most promising techniques that can be used to maximize the efficacy and efficiency of multi-label classification. However, because multi-label feature selection algorithms must consider multiple labels concurrently, the task is more difficult than single-label feature selection tasks. In this paper, we propose the Mutual Information-based multi-label feature selection method using interaction information. This method is naturally able to measure dependencies among multiple variables. To develop an efficient multi-label feature selection method, we derive theoretical bounds for the interaction information. Empirical studies indicate that our proposed multi-label feature selection method discovers effective feature subsets for multi-label classification problems.

1. Introduction

Multi-label classification is a challenging problem that arises in modern application areas such as text categorization, gene function classification, and semantic annotation of images (Schapire & Singer, 2000; Sebastiani, 2002; Lewis, Yang, Rose, & Li, 2004; Diplaris, Tsoumakas, Mitkas, & Vlahavas, 2005; Boutell, Luo, Shen, & Brown, 2004). As in single-label classification, the performance of multi-label classification can suffer from an excessive number of input features; too many features and training patterns with limited sizes lead to inaccurate prediction of output labels (Guyon & Elisseeff, 2003; Saeys, Inza, & Larrañaga, 2007). Theoretically, a pattern may lose its distinguishment due to irrelevant or redundant features because the similarity of each pair of patterns in same class can decrease (Watanabe, 1969). These features can cause additional problems by confusing the learning algorithm, leading to poor classification performance. As a result, feature selection is considered one of the most promising techniques to alleviate these problems (Yang & Pedersen, 1997; Chen, Yan, Zhang, Chen, & Yang, 2007; Doquire & Verleysen, 2013; Trohidis, Tsoumakas, Kalliris, & Vlahavas, 2008; Lee & Kim, 2013; Kong & Yu, 2013).

To perform feature selection for multi-label data sets, multi-label feature selection algorithms have to consider multiple labels concurrently. The simplest way of treating multi-label learning is to convert the multi-label problems into a single-label multi-class problem, and then evaluating each feature subset in terms of dependency on the transformed new single-label (Chen et al., 2007; Trohidis et al., 2008). However, this also results in subsequent problems; because the representation of multiple labels is compressed to a single-label, the newly created label contains too many classes, leading to learning difficulties (Read, 2008). To circumvent this problem, we can consider the algorithm adaptation approach (Tsoumakas & Vlahavas, 2007), in which the algorithm does not perform any transformation of multiple labels; instead, the algorithm is internally adapted to multi-label problems (Yang & Pedersen, 1997; Nie, Huang, Cai, & Ding, 2010; Kong & Yu, 2013). This can be done by incorporating a score function that is able to measure dependency between selected features and multiple labels. In the work of Lee and Kim (2013), we proposed a Mutual Information (MI)-based score function that is able to measure the importance of selected feature subset without resorting problem transformation. To overcome the difficulty involved in estimating high-dimensional joint probability from training patterns of limited sizes, we approximated the score function by using second-order interaction.

The motivation of this work is to resolve four issues that were not investigated in the study of Lee and Kim (2013). Accordingly, the contributions of this work can be summarized as follows.

- We propose a new score function that is able to consider an any-degree interaction for selecting relevant features. In contrast, the score function proposed in the study of Lee and Kim (2013) was fixed to consider only second-order interaction. Moreover, the derived score function can be more naturally

* Corresponding author. Tel.: +82 2 820 5304.
E-mail addresses: jslee.cau@gmail.com (J. Lee), dwkim@cau.ac.kr (D.-W. Kim).

http://dx.doi.org/10.1016/j.eswa.2014.09.063
0957-4174/