Knowledge-Based Systems 000 (2017) 1–16

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Online Multi-label Group Feature Selection

Jinghua Liu[a], Yaojin Lin[b,*], Shunxiang Wu[a], Chenxi Wang[b]

[a] Department of Automation, Xiamen University, Xiamen, 361000, PR China
[b] School of Computer Science, Minnan Normal University, Zhangzhou, 363000, PR China

ARTICLE INFO

Article history:
Received 19 June 2017
Revised 6 December 2017
Accepted 7 December 2017
Available online xxx

Keywords:
Online feature selection
Multi-label learning
Streaming feature
Group feature selection

ABSTRACT

Feature selection for multi-label learning has received intensive interest in recent years. However, traditional multi-label feature selection are incapable of considering intrinsic group structures of features and handling streaming features simultaneously. To solve this problem, we develop an algorithm called Online Multi-label Group Feature Selection (OMGFS). Our proposed method consists of two-phase: online group selection and online inter-group selection. In the group selection, we design a criterion to select feature groups which is important to label set. In the inter-group selection, we consider feature interaction and feature redundancy to select an optimal feature subset. This two-phase procedure continues until there are no more features arriving. An empirical study using a series of benchmark data sets demonstrates that the proposed method outperforms other state-of-the-art multi-label feature selection methods.

1. Introduction

Recent years have witnessed the proliferation of multi-label learning, since it has been widely applied in various real-world application scenarios (e.g., text categorization [15,31], gene function classification [4], and music emotion recognition [33]). It focuses on the research problem that each instance is associated with multiple concepts simultaneously [4,33,48,51]. For example, a document may belong to several topic categories, a gene can be associated with a set of functional targets, and a piece of music can be assigned to multiple genres. However, multi-label data always exhibits enormous amount of features (even reach tens of thousands), which puzzles the research community a lot.

To address the curse of dimensionality, feature selection is used to reduce the dimensionality by eliminating the redundant and/or irrelevant features [10,18,20,21,54–56]. To date, numerous multi-label feature selection methods have been proposed from different computing paradigms, such as large margin [22,32], fuzziness [27,59], dependency [16,46], and information metric [8,14,19,23,24,26,30]. Regarding to information metric, Lee and Kim [12] presented a feature selection algorithm for multi-label classification using multivariate mutual information, which selected an effective feature subset via maximizing the dependency between the selected features and label set. Lee and Kim [13] proposed a fast multi-label feature selection based on information-theoretic feature ranking, which employed a score function to assess the importance of each feature and analyzed it in terms of computational cost. Lin et al. [23] presented a multi-label feature selection approach based on max-dependency and min-redundancy, which considered the conditional redundancy between the candidate feature and the selected features. Moreover, Lin et al. [22] proposed a multi-label feature selection method based on neighborhood mutual information. This algorithm can generalize neighborhood entropy in single-label learning to fit multi-label learning. These methods conduct feature selection based on the assumption that the whole feature space is known in advance. However, in many real-word scenarios, the features are generated dynamically. For example, hot topics are continuously changing in the social network platform of twitter. When a hot topic appears, it is always accompanied with a set of fresh keywords (which also means a set of new features). This phenomenon illustrates that it is unrealistic to wait for a complete set of features, and it is absolutely necessary to perform feature selection when new features arrival.

On the other hand, existing multi-label feature selection methods evaluate feature individually, and ignore the underlying structure of features. For instance, in image processing, features could be generated gradually in groups, such as color, texture, and other visual information. In the circumstance of disease diagnosis, the examination results consist of groups of data including electrocardiogram, blood routine, and urinalysis. As we know, the information of group structure can be regarded as a type prior knowledge of features, and the destruction of group structure may degrade learning performance. Therefore, performing feature selection with group structure is most suitable for the actual situation. To date, several research works related to group feature

* Corresponding author.
E-mail addresses: yjlin@mnnu.edu.cn, zzlinyaojin@163.com (Y. Lin).

https://doi.org/10.1016/j.knosys.2017.12.008
0950-7051/

selection have been presented, such as group Lasso [43], and sparse group Lasso [34,40]. However, these group feature selection methods are designed for single-label learning. To the best of our knowledge, no group feature selection method for multi-label learning has been reported so far.

Indeed, many existing algorithms can effectively select a subset of features for different multi-label classification learning. However, these algorithms are not capable of considering intrinsic group structures of features and dealing with streaming features simultaneously. Motivated by these observations, we propose a novel framework for multi-label feature selection in an online manner by integrating with group structure analysis. Our work is inspired by feature relevance, feature redundancy, and feature interaction. More specifically, when a new group of features arrives, we first design a novel criterion based on information theory to determine whether the group of features is relevant. This process is called online group selection. And then, the relevant feature groups are added into a buffer pool. Once the buffer pool is full, we re-evaluate the correlation between all the selected features and the features stored in the buffer pool, i.e., we process an online inter-group feature selection. The process can be accomplished with feature interaction and feature redundancy analysis. Finally, extensive experiments show that our proposed approach is able to yield significant gains as compared with many other state-of-the-art multi-label feature selection methods.

In summary, the major contributions of the proposed method are as follows:

- Our work advances the relevance-, redundancy-, and interaction-based multi-label feature selection for managing streaming features.
- To the best of our knowledge, this is the first effort that accounts for intrinsic group structures of features and handling streaming features simultaneously in multi-label learning.
- OMGFS can be used to perform group- and single- feature selection, simultaneously.
- We validate the superiority of our proposed algorithm via comparing with other state-of-the-art multi-label feature selection algorithms from different performance views.

The rest of this paper is organized as follows. After discussing related work in Section 2, we introduce the concepts of multi-label learning and multi-label neighborhood mutual information in Section 3. In Section 4, we propose the framework of Online Multi-label Group Feature Selection and give our algorithm. Then, the experiments on benchmark data sets are demonstrated in Section 5. At last, our conclusions and future work are given in Section 6.

2. Related work

Feature selection, as an effective technology to handle with high dimensionality, has received considerable attention both in statistics and in machine learning. Existing feature selection algorithms are designed in various ways. According to whether the label information are available, feature selection methods can be categorized into unsupervised and supervised. Supervised feature selection can be broadly classified into single-label feature selection and multi-label feature selection. In which, the main difference between the two is that each instance is associated with one label or multiple labels simultaneously.

In single-label learning, feature selection can be conducted in two types, namely offline manner and online manner. For offline manner, it assumes that the global feature space on the training data has been achieved in advance and all features are examined at each round to select the best feature. A great variety of feature selection algorithms in an offline manner have been developed and proven to be effective in improving predictive accuracy for classification [9,25,28,36–38,57]. However, in real-world scenarios, features are actually generated dynamically. It is very time-consuming to wait for the calculation of all the features. Contrast to the offline manner, online feature selection assumes that the candidate features are generated dynamically and obtains an optimal feature subset from the features seen so far by processing features upon its arrival [7,18,20,39,44]. According to whether the feature arrive one by one or group by group, online feature selection can be roughly divided into online individual feature selection and online group feature selection. Some representative works about online individual feature selection have been proposed, including Grafting [29], OSFS [39], Fast-OSFS [39], SAOLA [44], and FRSA-IFS-HIS [58]. In which, FRSA-IFS-HIS include the FRSA-IFS-HIS(AA) and FRSA-IFS-HIS(AD) algorithms according to the features change (adding or deleting) dynamically. On the other hand, many algorithms of online group feature selection have been designed, such as OGFS [35], group-SAOLA [44], and GFSSF [17]. In this paper, we focus more on the scenario of online group feature selection with streaming features. In the context of dynamic features, the above mentioned feature selection algorithms perform well when dealing with high dimensional data, and they are of great practical significance.

Different with traditional single-label feature selection, there are many challenges that derive from feature space and label space in multi-label feature selection. In label space, the challenges include label correlation [6,11], missing label [45,53], weak label [42], label-specific features [51], and class-imbalance [52]. These above mentioned challenges share one common assumption that the whole label space are available in advance, and they ignore a not uncommon scenario: the label set size is unknown, or even infinite, rather than fixed a priori. Hence, Lin et al. [24] proposed multi-label feature selection with streaming labels, which assumed that the labels arrive one at a time, and the learning task is to rank features iteratively when a new label arrives. In feature space, multi-label data is also subjected to the curse of dimensionality. To date, numerous traditional multi-label feature selection methods have been developed, such as MLNB [47], MDDM [49], MDMR [23], NFNMI [22], PMU [12], FIMF [13], and RF-ML [32]. All these algorithms evaluate features based on the assumption that the feature space is known beforehand to execute feature selection. However, in real-world applications, it is difficult to obtain sufficient prior information about the structure of the feature space in an online manner. In addition, existing multi-label feature selection algorithms evaluate feature individually and generally ignore the group structure information of features, which may decrease classification performance. Therefore, we propose an online group feature selection framework, which including online group selection and online inter-group selection. Under this framework, a novel Online Multi-label Group Feature Selection algorithm is presented in this paper. Our proposed method, by comparison, makes an additional effort to manage the problem of multi-label feature selection in an online manner.

3. Preliminaries

3.1. Multi-label learning

Let $\mathcal{X} = \mathcal{R}^{n\times d}$ be the input space and $\mathcal{Y} = \{-1, +1\}^{|L|}$ be the label space with $|L|$ possible class labels, where $|L|$ denotes the number of label set L. Given a set of n multi-label training data set $\mathcal{D} = \{(\mathbf{x}_i, Y_i) | 1 \le i \le n\}$, where $\mathbf{x}_i \in \mathcal{X}$ denotes a d-dimensional feature vector $(x_{i1}, x_{i2}, \cdots, x_{id})$, and $Y_i \subseteq \mathcal{Y}$ denotes the relevant label vector $(Y_{i1}, Y_{i2}, \cdots, Y_{i|L|})$ of data point $\mathbf{x}_i$. The task of multi-label learning is to learn a function $f : \mathcal{X} \longrightarrow \mathcal{Y}$.

In multi-label classification learning, the performance evaluation functions are more complicated than the traditional