Knowledge-Based Systems 104 (2016) 52–61

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

# Multi-label learning with label-specific feature reduction

Suping Xu[a,e,f], Xibei Yang[a,b,e,f,*], Hualong Yu[a], Dong-Jun Yu[c], Jingyu Yang[c], Eric C.C. Tsang[d]

[a] School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, PR China
[b] School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, PR China
[c] Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information, Nanjing University of Science and Technology, Ministry of Education, Nanjing 210094, PR China
[d] Faculty of Information Technology, Macau University of Science and Technology, 519020, Macau
[e] Intelligent Information Processing Key Laboratory of Shanxi Province, Shanxi University, Taiyuan 030006, PR China
[f] Key Laboratory of Oceanographic Big Data Mining and Application of Zhejiang Province, Zhejiang Ocean University, Zhoushan 316022, PR China

ARTICLE INFO

*Article history:*
Received 22 December 2015
Revised 10 March 2016
Accepted 13 April 2016
Available online 25 April 2016

*Keywords:*
Feature reduction
Fuzzy rough set
Label-specific feature
Multi-label learning
Sample selection

ABSTRACT

In multi-label learning, since different labels may have some distinct characteristics of their own, multi-label learning approach with label-specific features named LIFT has been proposed. However, the construction of label-specific features may encounter the increasing of feature dimensionalities and a large amount of redundant information exists in feature space. To alleviate this problem, a multi-label learning approach FRS-LIFT is proposed, which can implement label-specific feature reduction with fuzzy rough set. Furthermore, with the idea of sample selection, another multi-label learning approach FRS-SS-LIFT is also presented, which effectively reduces the computational complexity in label-specific feature reduction. Experimental results on 10 real-world multi-label data sets show that, our methods can not only reduce the dimensionality of label-specific features when compared with LIFT, but also achieve satisfactory performance among some popular multi-label learning approaches.



## 1. Introduction

Nowadays, multi-label learning problem has received an increased attention in real-world applications. For example, in semantic annotation of images [3,16,26,49], a picture can be annotated as camel, desert and landscape. In text categorization [5,11,17,29], a document may belong to several given topics, including economics, finance or GDP. In bioinformatics [6,13,50], each gene may be associated with a set of functional classes, such as metabolism, transcription and protein synthesis. In all cases above, each sample may be associated with more than one label simultaneously and predefined labels for different samples are not mutually exclusive but may overlap. This situation is distinct from the traditional single-label learning where predefined labels are mutually exclusive, each sample only belongs to a single label.

Over the last decade, many multi-label learning approaches have been witnessed [12,28,58]. Generally, the existing methods can be grouped into two main categories [43], i.e., algorithm adaptation methods and problem transformation methods. Algorithm adaptation methods extend specific single-label learning algorithms to directly handle multi-label data by modifying some constraint conditions, such as AdaBoost.MH [40], ML-*k*NN [59], MLNB [60], and RankSVM [9]. Problem transformation methods, transform the multi-label task into one or more corresponding single-label ones and then handle them one by one through traditional methods. The well-known problem transformation methods include binary relevance (BR), label power set (LP) and pruned problem transformation (PPT). BR [3] learns a binary classifier for each label independently and predicts each of the labels separately, so it cuts up the relationship among different labels. LP [44] considers each unique set of labels that exists in a multi-label training set as a new single-label multi-value class. Though this method considers the correlations among different labels, it easily leads to a higher time consumption since the number of new classes is increased exponentially with the increasing of labels. Meanwhile, some new classes created by a few samples may lead to class unbalance problem. PPT [34] abandons the new classes associated with extremely small number of samples or assigns these samples with new labels that can create accepted classes, while some abandoned classes will lead to the loss of multi-label information. Although above methods have achieved good performance in multi-label learning, they make use of the same features to achieve

* Corresponding author at: School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, PR China.

*E-mail addresses:* supingxu@yahoo.com (S. Xu), yangxibei@hotmail.com (X. Yang), yuhualong@just.edu.cn (H. Yu), njyudj@njust.edu.cn (D.-J. Yu), yangjy@mail.njust.edu.cn (J. Yang), cctsang@must.edu.mo (E.C.C. Tsang).

http://dx.doi.org/10.1016/j.knosys.2016.04.012
0950-7051/

the learning purposes in different labels. Actually, different labels may have distinct characteristics of their own, and these characteristics are more inclined to judge whether labels belong to a specific sample. Fortunately, Zhang [61,62] has proposed the representative LIFT algorithm and validated the effectiveness of constructing label-specific features. For each label, LIFT employs clustering analysis in the positive and negative samples respectively, and then constructs label-specific features by checking the distances between the sample and all the clustering centers. (There is not any semanteme for constructed label-specific features, which can be regarded as a set of distances.) However, construction of label-specific features may encounter the increasing of feature dimensionalities, and a large amount of redundant information exists in feature space. As a result, the structure information between different samples will be disrupted, and even more be destroyed, which leads to the decreasing of the performance of multi-label learning approach. To alleviate this problem, an effective solution is to perform dimension reduction in label-specific features.

Rough set theory is a good mathematical tool for describing incomplete and uncertain data. With over 30 years of development, it has been widely applied in attribute reduction [18,30], feature selection [20,22,31,42,55], rule extraction [25,38] and uncertainty reasoning [46]. Numerous researchers [31,32] have used the various rough set models for dealing with single-label data analyses in real-world applications. Recently, some researchers [53,54,56,57] begin to attempt at carrying out multi-label classification via rough set approaches, however, all of them determine different labels in the same feature space, which contradicts the fact that different labels may have distinct characteristics of their own. In this paper, with the idea of attribute reduction based on fuzzy rough set, we will develop a multi-label learning approach with label-specific feature reduction (FRS-LIFT), which uses the approximation quality to evaluate the significance of specific dimension and takes the forward greedy search strategy. Furthermore, sample selection is an effective data compression technique, which can reduce the time and memory consumption in attribute reduction. On the basis of FRS-LIFT, another multi-label learning approach with label-specific feature reduction by sample selection (FRS-SS-LIFT) will be presented at the same time. To validate the effectiveness of FRS-LIFT and FRS-SS-LIFT, we conduct comprehensive experiments on 10 real-world multi-label data sets. Experimental study shows clear advantages of FRS-LIFT and FRS-SS-LIFT over various multi-label learning algorithms.

The rest of this paper is organized as following. Section 2 introduces the formal definition of multi-label learning's framework and LIFT approach. Section 3 provides some background materials on fuzzy rough set and sample selection, and then the details of our FRS-LIFT and FRS-SS-LIFT are presented. Section 4 describes data sets, evaluation metrics, experimental settings, and then analyzes the results of comparative studies on 10 multi-label data sets. Finally, Section 5 summarizes and sets up several issues for future work.

## 2. Multi-label learning

### 2.1. Multi-label learning's framework

Let $X=\mathbb{R}^d$ be the $d$-dimensional sample space and $L=\{l_1,l_2,\ldots,l_m\}$ be the finite set of $m$ possible labels. $T=\{(x_i,Y_i)|i=1,2,\ldots,n\}$ denotes the multi-label training set with $n$ labeled samples, where $x_i\in X$ is a $d$-dimensional feature vector such that $x_i=[x_i^1,x_i^2,\ldots,x_i^d]$, $Y_i\subseteq L$ is the set of labels associated with $x_i$.

The goal of multi-label learning is to produce a real-valued function $f:X\times P(L)\to\mathbb{R}$. In detail, for each $x_i\in X$, a prefect learning system will tend to output larger values for labels in $Y_i$ than those not in $Y_i$ [59], i.e., for any $l$, $l'\in L$, if $l\in Y_i$ and $l'\notin Y_i$, $f(x_i,l)>f(x_i,l')$ holds.

### 2.2. LIFT approach

#### 2.2.1. Construction for label-specific features

LIFT aims to improve the learning performance of multi-label learning system through generating distinguishing features which capture the specific characteristics of each label $l_k\in L$. To achieve this goal, LIFT takes into account intrinsic connection between different samples in all labels. Specifically, with respect to each label $l_k$, the training samples are divided into two categories, i.e., the set of positive training samples $P_k$ and the set of negative training samples $N_k$, such that:

$$P_k=\left\{x_i\middle|(x_i,Y_i)\in T,l_k\in Y_i\right\};\tag{1}$$

$$N_k=\left\{x_i\middle|(x_i,Y_i)\in T,l_k\notin Y_i\right\}.\tag{2}$$

In other words, the training sample $x_i$ belongs to $P_k$ if $x_i$ has label $l_k$; otherwise, $x_i$ is included in $N_k$.

To consider intrinsic connection among different samples, LIFT employs clustering analysis on $P_k$ and $N_k$, respectively. Following Zhang's research [61,62], $k$-means algorithm [21] is adopted to partition $P_k$ into $m_k^+$ disjoint clusters whose clustering centers are denoted by $\{p_1^k,p_2^k,\ldots,p_{m_k^+}^k\}$. Similarly, $N_k$ is also partitioned into $m_k^-$ disjoint clusters whose clustering centers are $\{n_1^k,n_2^k,\ldots,n_{m_k^-}^k\}$. LIFT treats clustering information gained from $P_k$ and $N_k$ as equal importance, and then the numbers of clusters on $P_k$ and $N_k$ are set to be the same, i.e., $m_k^+=m_k^-=m_k$. Specifically, the number of clusters for both positive samples and negative samples is:

$$m_k=\left\lceil\delta\cdot\min(|P_k|,|N_k|)\right\rceil,\tag{3}$$

where $|\cdot|$ represents the cardinality of a set, $\delta\in[0,1]$ is the ratio parameter for controlling the number of clusters.

The above two groups of clustering centers describe inner structures of positive samples $P_k$ and negative samples $N_k$, on this basis, label-specific features can be constructed in the form of:

$$\varphi_k(x_i)=\left[d(x_i,p_1^k),\ldots,d(x_i,p_{m_k}^k),d(x_i,n_1^k),\ldots,d(x_i,n_{m_k}^k)\right],\tag{4}$$

where $d(\cdot,\cdot)$ represents the distance between two samples. In literatures [61,62], Euclidean metric is used to calculate sample distance. Actually, $\varphi_k$ is a mapping from the original $d$-dimensional sample space $X$ to a new $2m_k$-dimensional label-specific feature space $LIFT_k$, i.e., $\varphi_k$: $X\to LIFT_k$.

#### 2.2.2. Induction for classification models

LIFT induces a family of $m$ classification models $\{f_1,f_2,\ldots,f_m\}$ in the constructed label-specific feature spaces $LIFT_k(1\le k\le m)$. Formally, for each $l_k\in L$, a binary training set $T_k^*$ with $n$ samples is created from the training set $T$ according to the mapping $\varphi_k$, such that:

$$T_k^*=\left\{(\varphi_k(x_i),\phi(Y_i,l_k))\middle|(x_i,Y_i)\in T\right\},\tag{5}$$

where $\phi(Y_i,l_k)=+1$ if $l_k\in Y_i$; otherwise, $\phi(Y_i,l_k)=-1$. Based on the binary training set $T_k^*$, any binary learner can be employed to induce a classification model $f_k$: $LIFT_k\to\mathbb{R}$ for $l_k$.

Given an unseen sample $x'\in X$, the predicted label set for $x'$ is $Y'=\{l_k|f(\varphi_k(x'),l_k)>0,1\le k\le m\}$.

## 3. Multi-label learning with label-specific feature reduction

### 3.1. Fuzzy rough set

To fuse rough set approaches into machine learning problems, we will introduce the classification learning task instead of the