An Ensemble Multi-Label Feature Selection Algorithm Based on Information Entropy

Shining Li, Zhenhai Zhang, and Jiaqi Duan
School of Computer Science, North Western Polytechnical University, China

Abstract: *In multi-label classification, feature selection is able to remove redundant and irrelevant features, which makes the classifiers faster and improves the prediction performance of the classifiers. Currently, most of feature selection algorithms in multi-label classification are dependent on the concrete classifier, which leads to high computation complexity. Hence this paper proposes an Ensemble Multi-label Feature Selection algorithm based on Information Entropy (EMFSIE), which is independent on any concrete classifiers. Its core idea consists of: Employing the information gain to evaluate the correlation between the feature and the label set, and filtering out useful features more effectively. We calculate the information gain in an ensemble framework and filter out useful features according to the threshold value determined by the effective factor. We validate EMFSIE on four datasets from two domains using four different multi-label classifiers. The experimental results and their analysis show preliminarily that EMFSIE can not only remove more than 70% of original features, which makes the classifiers faster, but also keep the prediction performance of the classifiers as good as before, even enhance the prediction performance on three datasets underthe two-tailed paired t-tests at 0.05 significance level.*

Keywords: *Data mining, ensembles, feature extraction, feature selection, information entropy, multi-label classification.*

Received October 28, 2012; accepted March 13, 2013; published online April 4, 2013

1. Introduction

Feature dimension reduction is one of the fundamental problems in the field of intelligence information processing. There are many datasets with a large number of features in most of classification applications. However, only a small number of them are relevant to label sets and irrelevant features will hurt the prediction performance of the classifiers [9]. The recent researches mainly focused on single-label classification problems [3]. In single-label classification, an instance is only associated with a single label; for example, a document can only be classified to economy or politics. However, an instance can be associated with multiple labels simultaneously in multi-label classification; for example, a document may belong to economy and politics at the same time. Multi-label classification has been applied to a variety of domains, including text categorization [7], bioinformatics [1], and music classification [8]. Owing to differences between single-label classification and multi-label classification, some of feature selection algorithms in single-label classification cannot be directly applicable to problems in multi-label classification. Thus, it is essential to investigate feature dimension reduction issues in multi-label classification.

Feature dimension reduction can be categorized as feature extraction and feature selection [3]. Currently, several approaches of feature dimension reduction have been developed in multi-label classification, but the majority focus on feature extraction. Feature extraction maps feature variables from high dimensional space to low dimensional space. It constructs new features out of original ones, which makes us hardly understand problems and can't provide a better understanding of underlying process that generated data. For instance, if we know which features are useful for the classification, it can avoid bringing so much data of irrelevant features, which will reduce workload for us extremely when collecting data. In addition, feature extraction cannot deal with discrete data. Feature selection removes irrelevant and redundant features out of original ones. Usually feature selection has a triple purpose [4]: Improving the performance of classifier, making classifiers faster, and providing a better understanding of the process that generated data. Recently a few algorithms of feature selection have been proposed [11, 21]. However, the computation complexity of them is dependent on the concrete classifiers and the feature subset varies with the classifiers.

To deal with the aforementioned problems, we proposes an Ensemble Multi-label Feature Selection algorithm based on Information Entropy (EMFSIE), which adopts the information gain to evaluate correlation between the feature and the label set. We calculate the information gain between the feature and the label set in an ensemble framework and then select useful features based on the threshold value, and finally experimental results demonstrate the effectiveness of EMFSIE. In addition, EMFSIE is independent of any classifiers. Hence, the selected

feature subset will not vary with the classifiers. The main contributions of this paper are as follows:

1. We propose the concept of the information gain between the feature and the label set. It adopts the sum of the information gain between the feature and each label in the label set as the information gain between the feature and the label set, which is used to evaluate correlation between the feature and the label set.
2. We calculate the information gain between the feature and the label set in an ensemble framework. The threshold selection strategy employs the effective factor to determine the threshold value and the simple control of the effective factor enable EMFSIE applied to different applications.

The remainder of this paper is organized as follows: Section 2 presents related work about feature dimension reduction on multi-label classification. Section 3 introduces multi-label classification and the information gain. Section 4 describes the proposed algorithm of feature selection in detail, including calculating the information gain in an ensemble framework and the threshold selection strategy. In section 5, experimental results on four datasets are presented. Finally, section 6 concludes this paper.

2. Related Works

Feature dimension reduction can be categorized as feature extraction and feature selection. There are some researches for feature extraction in multi-label classification in recent years. Zhang *et al.* [19] proposed an algorithm for multi-label classification, into which integrated the process of feature extraction and feature selection. Firstly, it employed feature extraction based on principal component analysis to remove irrelevant features. Afterwards it adopted feature selection based on genetic algorithm to select the most useful subset of features, whose fitness function integrated both hamming loss and ranking loss. Zhang and Zhou [22] proposed a feature extraction algorithm based on dependence maximization in multi-label classification. Based on the Hilbert-Schmidt Independence Criterion it projected original features into a lower dimensional feature space by maximizing the dependence between the original features and the associated labels. Park and Lee [12] extended linear discriminant analysis which is originally focused on the single-label problem to deal with dimensionality reduction in multi-label classification. Yu *et al.* [18] proposed a supervised approach called multi-label informed latent semantic indexing for multi-label dimensionality reduction based on Latent Semantic Indexing, which is an unsupervised approach in single-label classification. It captured the correlations of labels by solving an optimization problem for linear projection whose objective function incorporated both features and labels. Ji and Ye [5] proposed a linear dimensionality reduction approach for multi-label classification. It learned a linear function for each label by solving an optimization problem using the least square loss. Feature extraction constructs new features out of original ones, and therefore it is difficult to understand which features are useful among original features.

There are also a few works about feature selection in the context of multi-label classification. Li *et al.* [11] proposed a semi-supervised multi-label learning algorithm. It performed an embedded feature selection algorithm to remove the irrelevant features, which employed the prediction risk criterion to evaluate the feature subset. Zhang *et al.* [21] proposed a feature selection approach based on simulated annealing for multi-label data. It employed the simulated annealing algorithm to search the optimal subset, which used average precision to evaluate the feature subset. However, the above approaches for feature selection are associated with concrete classifiers, hence not only the computation complexity of them is dependent on the concrete classifiers, but also the feature subsets are different when different classifiers are employed in the same algorithm. Lee *et al.* [10] proposed a multi-label feature selection algorithm, which adopted approximating mutual information to evaluate the correlation between the feature and the label set. They derived approximated joint mutual information using Shearer's inequality, and maximized conditional mutual information together with incremental search for selecting feature subset. but they didn't give the threshold selection strategy to filter out useful features, and hence the proposed algorithm can't be applied directly in applications.

3. Multi-Label Classification and Information Gain

3.1. Multi-Label Classification Description

In this section, we introduce the formal notations of multi-label classification used throughout this paper.

- *Definition 1:* Let a vector of n attribute values $X = (x_1, x_2, ..., x_n)$ represent an instance, $xi \in R$ and $L = \{l_1, l_2, ..., l_m\}$ is a finite set of labels. The label set with, which an instance is associated is the subset of L, and it is represented by a label vector $Y = (y_1, y_2, ..., y_m)$. If an instance is associated with the label l_i, the corresponding y_i is 1, otherwise 0. Then the multi-label dataset containing t instances is denoted by:

$$D = \{(X_i, Y_i) / 1 \le i \le t, X_i \in R^n\} \quad (1)$$

3.2. Information Gain Description

Let the set $A = \{a_1, a_2, ..., a_m\}$, the set $B = \{b_1, b_2, ...,$

$b_n\}$, and then the information gain between A and B is defined as:

$$IG(B, A) = H(B) - H(B/A) \quad (2)$$

The information gain is used to measure the degree of correlation between A and B. The larger the information gain value is, the stronger the correlation between A and B is. According to the information theory, it can be educed that:

$$H(B/A) \geq 0 \quad (3)$$

Hence,

$$IG(B; A) \leq H(B) \quad (4)$$

For the same reason:

$$IG(B; A) \leq H(A) \quad (5)$$

The another equivalent form of information gain can be rewritten as:

$$IG(B; A) = H(A) + H(B) - H(AB) \quad (6)$$

We will prove that the information gain between the feature and the label can discriminate the useless features and the useful features by the following two theorems. The proofs of two theorems are provided in the appendixes.

- *Theorem 1:* If the set A and the set B are independent on each other, the information gain value is minimum.
- *Theorem 2:* If the set B is dependent on the set A completely, the information gain value is maximum.

Theorem 1 and theorem 2 reveal that the information gain has ability to evaluate the degree of correlation between A and B. Therefore this paper investigates the correlation between the feature and the label set based on the information gain between the feature and the label.

4. Feature Selection Based on Ensemble Method and Information Entropy

4.1. The Correlation between the Feature and the Label Set

- *Definition 2:* For the feature x and the label set $L = \{l_1, l_2, \ldots, l_n\}$, let $IG(l_i; x)$ denotes the information gain between the feature x and the label l_i, and then the information gain between the feature x and the label set L is represented as:

$$IGS(L; x) = \sum_{i=1}^{n} IG(l_i; x) \quad (7)$$

There is:

$$IGS(L; x) \geq 0 \quad (8)$$

Because of:

$$IG(l_i / x) \geq 0 \quad (9)$$

It can be derived that:

$$IGS(L; x) \leq \sum_{i=1}^{n} H(l_i) \quad (10)$$

According to the information theory. We will prove that the information gain between the feature and the label set can discriminate the useless features and the useful features by the following two theorems. The proofs of two theorems are described in the appendixes.

- *Theorem 3:* If a feature is independent on each label of the label set $L = \{l_1, l_2, \ldots, l_n\}$, the information gain between x and L is minimum.
- *Theorem 4:* If a feature is dependent on each label of the label set $L = \{l_1, l_2, \ldots, l_n\}$, the information gain between x and L is maximum.

It can be seen that the useless features will be removed by employing *IGS (L; x)* according to theorem 3. Theorem 3 and theorem 4 reveal that it is possible to remove features irrelevant to the label set L by setting a reasonable threshold value for the information gain. We firstly normalize *IG* before calculating *IGS* in order to evaluate the information gain between each feature and the label in the same range. Yu and Liu [17] has proved that the information gain is symmetrical, that is:

$$IG(B; A) = IG(A; B) \quad (11)$$

Then:

$$IG(B; A) \leq H(B) \quad (12)$$

$$IG(A; B) \leq H(A) \quad (13)$$

We use:

$$SU(A, B) = 2\left[\frac{IG(B; A)}{H(A) + H(B)}\right] \quad (14)$$

To normalize *IG (B; A)*. It ensure that *SU (A, B)* $\in$ *[0, 1]*. If there is:

$$SU(A, B) = 0 \quad (15)$$

It indicates that A and B are independent on each other. If there is:

$$SU(A, B) = 1 \quad (16)$$

It implies that one of them is dependent on the other one.

Algorithm 1: Calculate the Information Gain between the Feature and the Label
Input: Multi-label Dataset D
Output: The Set of the Information Gain between the Feature and the Label Set IGS = {IGS$_1$, IGS$_2$, IGS$_n$} IGS = CalculateIE (D)

1. *For each feature $x_i \in X$*
2. *For each label $l_i \in L$*
   Calculate the information gain IG (l_j; x_i) between the feature x_i and the label l_j according to the formula (6)
3. *s_j can be obtained by normalizing the IG(l_j; x_i) according to the equation 14,*
4. *End*

5. *Based on the set of the information gain between the feature and the label $S = \{s_1, s_2, \ldots, s_p\}$, calculate the information gain IGS_i between the feature x_i and the label set L according to the equation 7*
6. *End*

Algorithm 1 calculates the information gain IGS_i between each feature and the label set in the dataset, and then return $IGS = \{IGS_1, IGS_2, IGS_n\}$, the set of the information gain between the feature and the label set. Step 2 to 5 calculates the information gain $IG(l_j; x_i)$ between the current feature x_i and each label l_j of the label set L, which is employed to measure the degree of the correlation between the current feature x_i and the label set L. The larger the information gain value is, the stronger the correlation between the current feature x_i and the label set L is. Here the label set L is defined in the definition 1.

4.2. Ensemble Method

In classification problems, if all classifiers disagree with one another, then the ensemble method will work better than an individual classifier [16]. Since the ensemble method can overcome the errors some of the classifiers introduce. This study integrates the ensemble method into the process of feature selection in order to filter out more effectively. First of all, it partitions the multi-label dataset into several subsets, and then calculates the information gain between each feature and the label set on every subset, and finally sums up the information gain based on every subset. In this process we need to solve two problems: How to divide multi-label dataset into several subsets, and how many subsets dataset should be partition into. For the first problem, this study adopts the cluster algorithm to partition the dataset. It employs the classical cluster algorithm kmeans [15] to partition multi-label dataset into several subsets based on the feature variants so that the subsets are different from one another. Before solving the second problem, we first introduce label cardinality [3]:

$$Label\ \ cardinality\,(LC) = \frac{1}{t}\sum_{i=1}^{t}|Y_i| \tag{17}$$

Which is used to describe the average number of labels of an instance in multi-label dataset. For the second problem, we round up the label cardinality to get the number of subsets. That is $Subset\_Size = \lceil LC \rceil$.

Algorithm 2: Calculate the Information Gain based on the Ensemble Method
Input: Multi-label Dataset D
Output: The Ensemble Set of the Information Gain $AG\_IGS = \{AG\_IGS_1, AG\_IGS_2, \ldots, AG\_IGS_n\}$, $AG\_IGS = GetAggregationIE\ (D)$

1. *Calculate the label cardinality of dataset D according to the formula (17), and obtain the number of subsets $Subset\_Size = \lceil LC \rceil$*
2. *Use kmeans to cluster the dataset D based on the feature variants with Subset_Size as the parameter and then obtain the set of the subsets $KD = \{KD_1, KD_2, \ldots, KD_{Subset\_Size}\}$*
3. *For each subset $KD_i \in KD$*
4. *$IGS = CaculateIE\ (KD_i)$*
5. *For each $IGS_i \in IGS$*
6. *$AG\_IGS_i = AG\_IGS_i + IGS_i$*
7. *End*
8. *End*

Algorithm 2 integrates the ensemble method into the process of calculating the information gain between each feature and the label set based on algorithm 1. Step 1 obtains the number of the subsets by calculating the Label Cardinality (LC) of the dataset D and rounding up the LC. Step 2 employs kmeans to divide the dataset D into several subsets and obtains the set of the subsets $KD = \{KD_1, KD_2, \ldots, KD_{Subset\_Size}\}$. Here, there is $KD_1 \cap KD_2 \cap \ldots \cap KD_{Subset\_Size} = \varnothing$ and $D = KD_1 \cup KD_2 \cup \ldots \cup KD_{Subset\_Size}$. Step 5 to 7 aggregates the information gain on all the subsets.

4.3. Threshold Selection

AG_IGS, the set of the information gain between the feature and the label set, has been obtained in section 4.2. We need to filter out the useful features by setting a reasonable threshold value based on *AG_IGS*.

- *Definition 3:* Let the sum of all elements of *AG_IGS* be S, $\delta \in [0, 1]$, and then the product of S and δ, $S * \delta$, is called the amount of the effective information, and δ is the effective factor.

We assume that the features are independent on one another. According to theorem 3 and theorem 4, it can be known that the information gain between the feature and the label set represents the amount of information the feature contains. The larger the information gain is, the more important the feature is for the label set. The most important feature is taken precedence to be filtered out. We can simply tune the effective factor δ to enable EMFSIE to be applied to different practical applications.

Algorithm 3: Filter Out Features
Input: Multi-label Dataset D, the Set of the Information Gain between the Feature and the Label Set AG_IGS, the Effective Factor δ
Output: The Dataset After Dimension Reduction D', $D' = SelectFeatures\ (D, AG\_IGS, \delta)$

1. *Sum up all the elements of AG_IGS to be S*
2. *Extract the feature set $F = \{F_1, F_2, \ldots, F_n\}$ from dataset D*
3. *Sort AG_IGS by descending to get the sorted set of the information gain between the feature and the label set $AG\_IGS\_SORT = \{AG\_IGS\_SORT_1, AG\_IGS\_SORT_2, \ldots, AG\_IGS\_SORT_n\}$ and the corresponding set of features $F\_SORT = \{F\_SORT_1, F\_SORT_2, \ldots, F\_SORT_n\}$*
4. *SelectedFeatures = {}*
5. *EffectInfoGain = S * δ*

6. *ReservedInfoGain = 0*
7. *For i = 0:1:|AG_IGS_SORT|*
8. *If ReservedInfoGain > EffectInfoGain*
9. *Break*
10. *Else*
11. *ReservedInfoGain = ReservedInfoGain + $AG\_IGS\_SORT_i$*
12. *Add the corresponding feature F_SORT_i to the selected set SelectedFeatures*
13. *End*
14. *End*
15. *Remove the unselected features from dataset D according to the selected set SelectedFeatures, and then return the dataset after dimension reduction D'*

Algorithm 3 filters out the useful features according the threshold value based on *AG_IGS* obtained in algorithm 2. Step 1 calculates the sum of all the elements of *AG_IGS*. Step 3 sorts *AG_IGS* to ensure that the feature with large value will be taken precedence to be filtered out. *F_SORT*, the feature set with the same order as *AG_IGS* is obtained. Steps 7 to 14 filters out the useful features according to the threshold value determined by the product of *S* and δ. Step 15 performs dimension reduction on the dataset *D* according to the selected feature set to remove irrelevant features.

4.4. Complexity Analysis

According to definition 1, for algorithm 1 step 3 calculates the information gain between the feature and the label, and the complexity is $O(t)$; step 2 to 5 calculates the information between the current feature and each label of*L*, and the complexity is $O(m \times t)$; therefore the complexity of algorithm 1 is $O(n \times m \times t)$. For algorithm 2, step 2 uses kmeans to divide dataset into several subsets, and the complexity is $O(Subset\_Size \times t \times r)$(here *r* is the number of iterations); the complexity of step 3 to 8 is $O(Subset\_Size \times n \times m \times t)$; therefore the complexity of algorithm 2 is $O(Subset\_Size \times n \times m \times t)$. For algorithm 3, step 3 uses bubble sort algorithm to sort *AG_IGS*, and the complexity is $O(n^2)$; the complexity of step 7 to 14 is $O(n)$ and the complexity of step 15 is $O(t)$; therefore the complexity of algorithm 3 is $O(n^2)$.

From the aforementioned analysis, we can derive that the complexity of EMFSIE is $O(Subset\_Size \times n \times m \times t)$. Usually the value of *Subset_Size* is small, for instance*Subset_Size* = 3.

5. Experimentation and Result Analysis

We have implemented EMLFSIE algorithm based on Mulan [2] which is built on the open source Weka [14] library. We firstly performed EMLFSIE algorithm on the datasets, and then employed the classifiers to make prediction using the 10-fold cross-validation. Experiments are conducted to prove that EMLFSIE can not only make the classifiers faster but also keep the prediction performance as good as before or even better. EMLFSIE is evaluated on four classifiers: BR [3] (Binary Relevance), LP [3] (Label Powerset), MLkNN [20] (Multi-Label k-Nearest Neighbor) and CC [6] (Classifier Chains for Multi-Label Learning). The parameters of all the classifiers are set as default values in the Mulan. The decision tree algorithm C4.5 [13] is adopted as the base classifier of BR and LP. The values of the effective factor δ are required to be determined according to different application domains, which is acquired by repeating experiments with δ varying from 0.05 to 0.95 with step size of 0.05. In experiments, δ is 0.35 in the text domain, and 0.9 in the biological domain.

5.1. Datasets

Experiments were conducted on four multi-label datasets[1]: Enron, Medical, Bibtex and Genbase, which are from two kinds of tasks: text analysis and gene function prediction. Table 1 displays four multi-label datasets from two kinds of domains and their associated statistics.

Table 1. Multi-label datasets and associated statistics.

Name	Domain	Instances	Features	Labels	Cardinality
Enron	Text	1702	1001	53	3.378
Medical	Text	978	1449	45	1.245
Bibtex	Text	7395	1836	159	2.402
Genbase	Biology	662	1186	27	1.252

5.2. Evaluation Measures

The objectives of feature selection are to make the classifiers faster and improve the performance of the classifiers. Micro Fmeasure and HammingLoss [3] are adopted to evaluate the prediction performance of the classifiers. Dimensionality rate is employed to measure the effectiveness of dimension reduction. Reduction of features can make the classifiers faster. Let the predicting label set of a classifier predicting an unseen instance be *Z*, and the evaluation measures are defined as follows according to definition 1. Micro FMeasure is defined as:

$$Micro\ FMeasure = \frac{2 \times tp}{2 \times tp + fp + fn} \tag{18}$$

Here *tp*, *fp* and *fn* represent the number of true positives, false positives and false negatives respectively. The larger Micro FMeasure is, the better the prediction performance is. HammingLoss is defined as:

$$Hamming Loss = \frac{1}{t}\sum_{i=1}^{t}\frac{|Y_i \Delta Z_i|}{m} \tag{19}$$

Here Δ stands for the symmetrical difference of two label sets. The smaller HammingLoss is, the better the

[1]Available at http://mulan.sourceforge.net/datasets.html.

predicting performance is. Dimension rate is defined as:

$$DR = \frac{N_b - N_a}{N_b} \quad (20)$$

Which is used to measure the effectiveness of dimension reduction. $DR \in [0, 1]$. The larger DR is, the better the effectiveness of dimension reduction is. N_b represents the number of features before dimension reduction. N_a represents the number of features after dimension reduction.

5.3. Results Analysis

The experimental results are reported in the form of average ± standard deviation. The two-tailed paired t-tests at 0.05 significance level is employed to analyze the experimental results in order to examine whether the results have a statistical significant advantage. In the following tables, the results in bold represents that the prediction performance of the classifiers are improved significantly and the others represent that there is no significant difference after dimension reduction under the analysis of the two-tailed paired t-tests.

Table 2 displays the Micro Fmeasure values of four different classifiers on four datasets both before and after dimension reduction. Table 3 shows the HammingLoss values of four different classifiers on five data sets both before and after dimension reduction. Table 4 lists the Dimension Rate values on five data sets.

In the text domain, there are three datasets: Enron, Medical and Bibtex in the experiments. From the results of Tables 2 and 3, it can be seen that after performing dimension reduction a better prediction performance can be achieved for four classifiers in most situations. We further analyze the results using the two-tailed paired t-tests at 0.05 significance level, and it shows that LP and MLkNN outperform than before significantly, while there are no significant differences for BR and CC. Therefore it indicates that the prediction performance of four classifiers is as good as before and even better than before significantly for LP and MLkNN. Table 4 shows that the Dimension rate is 72.6 % on Enron and 92.2 % on Medical, and 80.4 % on Bibtex. It indicates that EMLFSIE can remove the irrelevant features greatly and make the classifiers faster.

There is a dataset named genbase belonging to the biological domain. Tables 2 and 3 shows that four classifiers perform better than before in most situations. We further analyze the results using the two-tailed paired t-tests at 0.05 significance level, and it reveals that there are no significant differences as before for four classifers. The above analysis indicates that the prediction performance of four classifiers is as good as before. Table 4 shows that the Dimension Rate is 96.3% on Genbase. It indicates that EMLFSIE can remove the irrelevant features greatly, which makes the classifiers faster.

The above analysis indicates that EMLFSIE demonstrates its great effectiveness on text and biological datasets, which not only removes the irrelevant features greatly but also keeps the prediction performance of the classifiers as good as before.

6. Conclusions

This paper proposes an ensemble multi-label feature selection algorithm based on information entropy, which is independent on any concrete classifier. On the basis of information gain between the feature and the label, the sum of the information gain between the feature and each label in the label set as the information gain between the feature and the label set is used to evaluate correlation between the feature and the label set. The information gain between the feature and the label set is calculated in an ensemble framework in order to filter out useful features more effectively. Finally, useful features are filter out according to the threshold value determined by the effective factor. The experimental results show that EMFSIE can not only remove irrelevant feature greatly but also keep the prediction performance of the classifiers as good as before or even better and make the classifiers faster.

Table 2. Micro Fmeasure of the classifiers before and after dimension reduction.

	Enron		Medical		Bibtex		Genbase	
	Before	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**
BR	0.5481±0.0251	0.5511±0.0285	0.8091±0.0282	0.8172±0.0275	0.4084±0.0096	0.4119±0.0151	0.9880±0.0058	0.9881±0.0067
LP	0.4308±0.0226	**0.4596±0.0184**	0.7526±0.0270	**0.7874±0.0280**	0.3017±0.0116	**0.3218±0.0096**	0.9801±0.0176	0.9815±0.0195
MLkNN	0.4778±0.0226	**0.4950±0.0236**	0.6800±0.0401	**0.7550±0.0344**	0.2218±0.0157	**0.3254±0.0189**	0.9462±0.0319	0.9561±0.0202
CC	0.5363±0.0308	0.5396±0.0221	0.8115±0.0324	0.8132±0.0271	0.3993±0.0105	**0.4055±0.0140**	0.9880±0.0058	0.9881±0.0067

Table 3. HammingLossof the classifiers before and after dimension reduction.

	Enron		Medical		Bibtex		Genbase	
	Before	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**
BR	0.0508±0.0022	0.0500±0.0027	0.0103±0.0014	0.0100±0.0016	0.0146±0.0004	**0.0133±0.0004**	0.0011±0.0006	0.0011±0.0007
LP	0.0717±0.0028	**0.0672±0.0026**	0.0135±0.0016	**0.0115±0.0017**	0.0205±0.0006	**0.0196±0.0003**	0.0019±0.0019	0.0018±0.0021
MLkNN	0.0523±0.0022	0.0518±0.0025	0.0151±0.0018	**0.0124±0.0020**	0.0136±0.0004	**0.0128±0.0004**	0.0048±0.0030	0.0040±0.0020
CC	0.0524±0.0024	0.0521±0.0025	0.0102±0.0017	0.0102±0.0016	0.0146±0.0004	**0.0133±0.0004**	0.0011±0.0006	0.0011±0.0007

Table 4. Dimension rate of datasets.

	N_b	N_a	DR(%)
Enron	1001	274	72.6
Medical	1449	113	92.2
Bibtex	1836	360	80.4
Genbase	1186	44	96.3

Acknowledgements

This research was supported by the Fundamental Research Grant of Northwestern Polytechnical University under Grants JC20120208, and National Science and Technology Major Project of the Ministry of Science and Technology of China under Grants 2012ZX03005007.

References

[1] Celine V., Jan S., Leander S., Sao D., and Hendrik B., "Decision Trees for Hierarchical Multi-Label Classification," *Machine Learning*, vol. 73, no. 2, pp. 185-214, 2008.

[2] Grigorios T., Eleftherios S., Jozef V., and Ioannis V., "Mulan: A Java Library for Multi-Label Learning," *Journal of Machine Learning Research*, vol. 12, no. 1, pp. 2411-2414, 2011.

[3] Grigorios T., Ioannis K., and Ioannis V., "Mining Multi-Label Data," *in Proceedings of Data Mining and Knowledge Discovery Handbook*, Springer, USA, pp. 667-685, 2010.

[4] Isabelle G. and Andre E., "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, vol. 3, no. 1, pp. 1157-1182, 2003.

[5] Ji S. and Ye J., "Linear Dimensionality Reduction for Multi-Label Classification," *in Proceedings of the 21st International Joint Conference on Artificial Intelligence*, USA, pp. 1077-1082, 2009.

[6] Jesse R., Bernhard P., Geoff H., and Eibe F., "Classifier Chains for Multi-Label Classification," *in Proceedings of European Conference on Machine Learning and Knowledge Discovery in Databases*, Slovenia, pp. 254-269, 2009.

[7] Johannes F., Eyke H., Eneldo L., and Klaus B., "Multilabel Classification via Calibrated Label Ranking," *Machine Learning*, vol. 73, no. 2, pp. 133-153, 2008.

[8] Konstantinos T., Kalliris G., and Vlahavas I., "Multi-Label Classification of Music into Emotions," *in Proceedings of the 9th International Conference on Music Information Retrieval*, USA, pp. 325-330, 2008.

[9] Krishna B. and Kaliaperumal B., "Efficient Genetic-Wrapper Algorithm Based Data Mining for Feature Subset Selection in a Power Quality Pattern Recognition Application," *the International Arab Journal of Information Technology*, vol. 8, no. 4, pp. 397-405, 2011.

[10] Lee J., Lim H., and Kim D., "Approximating Mutual Information for Multi-Label Feature Selection," *Electronics Letters*, vol. 48, no. 15, pp. 129-130, 2012.

[11] Li G., You M., and Ge L., "Feature Selection for Semi-Supervised Multi-Label Learning with Application to Gene Function Analysis," *in Proceedings of the 1st ACM International Conference on Bioinformatics and Computational Biology*, USA, pp. 354-357, 2010.

[12] Park C. and Lee M., "On Applying Linear Discriminant Analysis for Multi-Labeled Problems," *Pattern Recognition Letters*, vol. 29, no. 7, pp. 878-887, 2008.

[13] Quinlan J., *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, California, 1993.

[14] Remco R., Eibe F., Mark A., Geoffrey H., Bernhard P., Peter R., and Lan H., "Weka-Experiences with a Java Open-Source Project," *Journal of Machine Learning Research*, vol. 11, pp. 2533-2541, 2010.

[15] Seber G., *Multivariate Observations*, John Wiley & Sons, Hoboken, New Jersey, USA, 1984.

[16] Thomas G., "Machine-Learning Research," *AI Magazine*, vol. 18, no. 4, pp. 97-136, 1997.

[17] Yu L. and Liu H., "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," *in Proceedings of the 20th International Conference on Machine Learning*, USA, pp. 856-863, 2003.

[18] Yu K., Yu S., and Tresp V., "Multi-Label Informed Latent Semantic Indexing," *in Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Brazil, pp. 258-265, 2005.

[19] Zhang M., Pena J., and Robles V., "Feature Selection for Multi-Label Naive Bayes Classification," *Information Sciences*, vol. 179, no. 19, pp. 3218-3229, 2009.

[20] Zhang M. and Zhou Z., "ML-KNN: A Lazy Learning Approach to Multi-Label Learning," *Pattern Recognition*, vol. 40, no. 7, pp. 2038-2048, 2007.

[21] Zhang Y., You L., and Chen J., "Feature Selection for Multi-Label Data by Using Simulated Annealing," *Computer Engineering and Design*, vol. 32, no. 7, pp. 2494-2500, 2011.

[22] Zhang Y. and Zhou Z., "Multi-Label Dimensionality Reduction via Dependence Maximization," *in Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, USA, pp. 1503-1505, 2008.

Shining Li is a professor and PhD supervisor in the school of computer science, northwestern polytechnical university, China. He received his PhD degree from the school of electronic and information engineering from xi'an jiaotong university, China. His current research interests include wireless sensor networks, mobile computing, and data processing.

Zhenhai Zhang is a PhD candidate in the school of computer science and technology, northwestern polytechnical university, China. He received his MSc degree from the school of computer science from xi'an university of science and technology, China. His research interests include wireless sensor networks and data processing.

Jiaqi Duan received his BSc degree in electronics information engineering and his MSc and PhD degrees in communication engineering from Northwestern Polytechnical University, China, in 2006, 2009, and 2011, respectively. His current research interests include wireless sensor networks, internet of things and cognitive radio networks.

Appendixes

- *Proof for Theorem 1*: According to the information theory, it can be known that if the set A and the set B are independent on each other then:

$$H(AB) = H(A) + H(B) \quad (21)$$

In conjunction with the formula (6) it can be inferred that the information gain is zero in this situation. Since the information gain is non-negative, that is:

$$IG(B; A) \geq 0 \quad (22)$$

Therefore the information gain value is minimum.

- *Proof for Theorem 2*: If the set B is dependent on the set A completely, then for any $x_i \in A$ and any $y_j \in B$ if and only if one of the following two conditions is required: (1) x_i and y_j appear simultaneously. (2) x_i and y_j never appear simultaneously.

If x_i and y_j appear simultaneously, there is:

$$p(y_j/x_i) = p(x_i y_j)/p(x_i) = 1 \quad (23)$$

Otherwise:

$$p(x_i y_j) = 0 \quad (24)$$

According to the information theory, it can be educed that:

$$H(B; A) = 0 \quad (25)$$

Associated with the formula (2), it can be inferred that:

$$IG(B; A) = H(B) \quad (26)$$

That is to say, the information gain value is maximum.

- *Proof for Theorem 3*: According to the theorem 1, it can be known that if the label l_i and the feature x are independent on each other, then the information gain:

$$IG(l_i; x) = 0 \quad (27)$$

According to the definition 2, it can be deduced that:

$$IGS(L; x) = 0 \quad (28)$$

Since there is:

$$IGS(L; x) \geq 0 \quad (29)$$

and therefore the information gain between the feature x and the label set L is minimum.

- *Proof for Theorem 4*: According to the theorem 2, it can be known that if the label l_i is dependent on the feature x completely, then there is:

$$IG(l_i; x) = H(l_i) \quad (30)$$

According to the definition 2, it can be deduced that:

$$IGS(L; x) = \sum_{i=1}^{n} H(l_i) \quad (31)$$

Since there is:

$$IGS(L; x) \leq \sum_{i=1}^{n} H(l_i) \quad (32)$$

and therefore the information gain between the feature x and the label set L is maximum.