www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 11 November 2017, Page No. 23194-23198
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i11.19

REAL: Multi-Label Learning with Relevant Feature for Each Label

Jianjun Yan[1], Rui Guo[2]

[1] Center for Mechatronics Engineering, East China University of Science and Technology
130 Meilong Road, Shanghai 200237, China
[2] Center for TCM Information Science and Technology, Shanghai University of Traditional Chinese Medicine,
1200 Cailun Road, Shanghai 201203, China

Abstract:
Multi-label learning problems have become a key topic in machine learning research in recent years. However, most approaches have focused on exploiting the interdependences between labels, whereas the correlations between the original features and each group of possible class labels have been rarely examined. The association degree of a selected feature is biased toward each discriminate class label. With the aim of addressing the gaps in previous studies, the current paper proposes a novel framework called multi-label learning with Relevant fEature for eAch Label. Using this mechanism, a classification model to deal with enron and medical data sets is established. The experimental results demonstrate the effectiveness and competitive performance of the proposed scheme which outperformed other multi-label classification methods significantly**.**

Keywords: Feature selection, Multi-label learning, Text classification, REAL.

1. Introduction

In traditional data mining tasks, *single-label* classification is commonly used. It is known as disjoint *multi-class* classification which assigns an object to exactly one class. An instance x is associated with a single label λ from a set of mutually exclusive labels L, $|L|>1$. A single-label data set is denoted by $D=\left\{\left(\vec{x}_i,\lambda_i\right),i=1...N\right\}$. In contrast, *multi-label* classification in modern applications is more general and complex. It is known as unrestricted *multi-class* classification which collects simultaneously a set of labels $Y\subseteq L$ with each instance x. A multi-label data set is denoted by $D=\left\{\left(\vec{x}_i,Y_i\right),i=1...N\right\}$ [1].

Multi-label learning tasks are originally applied in text categorization in which each document may belong to several predefined topics, such as *government* and *health*, or *rock* and *blues* [2, 3]. Aside from text categorization, multi-label learning tasks are also used widely in other real-world problems. For instance, in a classification for natural scenery, each image may belong to several image types at the same time, such as sea and sunset [4]. In a *music-emotion* classification, music may simultaneously evoke more than one emotion such as *relaxed* and *sad* [5]. In an automated video annotation, each video clip may belong to a number of semantic classes such as urban and building [6]. In functional genomics, each gene may be associated with a set of functional classes such as *metabolism*, *transcription*, and *protein synthesis* [7].

Multi-label classification algorithms can be divided into two general categories [8] (i) *problem transformation* methods and (ii) *algorithm adaptation* methods.

The algorithms in the first group are self-determined. They transform multi-label classification tasks into one or more single-label classifications, regression, or ranking tasks. Binary relevance [9], a straightforward problem transformation approach, predicts positively the label sets of an unknown instance by N binary classifiers. Independent classifiers are commonly individual selection [10, 11] and fusing selection [12]. This *one-against-all* strategy has been criticized to ignore the correlations among labels [13]. Label Powerset (LP), a ranking problem transformation approach, outputs the probability distribution of each label of a new instance. A ranking of the labels is produced by a specific threshold (e.g., 0.5). The independent classifiers are generally individual selection. This one-against-one strategy has the advantage of taking label correlations into account, but it suffers from a large number of label subsets. Majority of the label subsets are associated with very few examples. The random k-label sets (RAkEL) method avoids the aforementioned problems of LP in [14]. It constructs an ensemble of LP classifiers. Each LP classifier is trained using a different, small, random subset from the set of labels. The ranking of labels is accomplished by setting the threshold of the average zero-one decisions of each model per considered label.

The second group includes methods that extend specific learning algorithms in order to handle multi-label data directly. C 4.5 is an adaptation algorithm in which entropy calculation is modified, and multiple labels are allowed [15]. AdaBoost.MH and AdaBoost.MR [3] are two decomposing adaptation algorithms applied on weak classifiers. BP-MLL [16] introduces a new error function that captures the characteristics of multiple labels without reducing the trivial time cost of a neural network. ML-kNN [17] uses the maximum a posteriori principle based on the prior and posterior probabilities for each label frequency within the k nearest neighbors. Ranking SVM [7] attempts to minimize ranking loss while maintaining a large margin. MMAC [18] deals with the construction of classification rule sets using association rule mining. The labels of each instance are ranked according to the support of the corresponding multi-label rule.

The general scheme utilized by previous methods is the identification of all feature representations of the instances in original data sets. The aspect of output space is particularly emphasized [19]. The systemic dependence between the original features and multiple class labels is taken into account effectively. However, the original features are utilized directly to predict the class labels from the perspective of input space. The collection of features, which has a strong association with each group of *possible class labels*, is only a subset of the original features. That is, the irrelevance and redundancies of the overall feature space may suffer from low prediction accuracy. For example, in *image* classification, *sky* and *desert* are supposed to be two possible class labels. Specific *color*-based features are preferred in differentiating sky and non-sky images, *texture*-based characteristics are preferred in differentiating desert and non-desert images, whereas both color- and *texture-based* features might be useful in differentiating other labels [18] In text classification, features such as *source*, *open*, and *kernel* may be more associated with the class label Linux, whereas attributes such as *anonymous*, *reader*, and *apple* may be more related to the class label *mobile* [19] .

The above-mentioned perspectives imply that each group of possible class labels has its own feature subset. In order to determine the dependence of features and class labels in multi-label learning tasks, a novel framework called multi-label learning with Relevant *fEature* for *eAch Label* (REAL) is proposed. This scheme is carried out by an ensemble of multiple kNN classifiers. Prior and posterior probabilities, as well as the posteriori maximum principle in [16] , are adopted to determine the label sets of the test instance. The feature selection methods are combined efficiently in the training and test periods of our framework.

The rest of the current paper is organized as follows. Section 2 presents the details of the REAL scheme. Section 3 shows the experimental results of comparative research. Finally, Section 4 presents the conclusions and future research directions.

2. The REAL Approach

Most traditional multi-label classification approaches in vector spaces are used based on the assumption that the instances should have a same set of features in the input space for each label. But, for specific labels, not all the features have strong correlations with those. Therefore, we are look for an approach assume that the instances have different set of features in the input space for each label hope to Hope to eliminate the interference characteristics as far as possible. In the REAL algorithm, we extract the best feature subset correlated with a certain label as its input space, and then search for the K nearest neighbors and calculate the posterior probability combined with the ML-KNN algorithm. Fig.1 simply profiles the basic idea of our framework.

REAL algorithm consists of 3 main steps: At the 1st step, REAL algorithm extract the best feature subset for each label using feature selection methods, such as *CMIM* [20], *MIFS* [21], *MRMR* [22], *JMI* [23], and *MIM* [24, 25]. At the 2nd step, in order to adapt to the assumption that the instances have different set of features in the input space for each label, we had improved the ML-KNN algorithm. When searching for the K nearest neighbors, the distance between two training instances for each label is calculated in the corresponding feature subspace instead of the whole feature space in witch is adopted by the ML-KNN algorithm. Then we could calculate the posterior probability with the K nearest neighbors and further the confidential threshold value. At the final step, for each test instance, search for the K nearest neighbors in corresponding feature subspace of a particular label, and identify if the instance belong to the label by estimate the posterior probabilities. Take the same operation to each label, and determine the label subset in which the test instance belongs to.

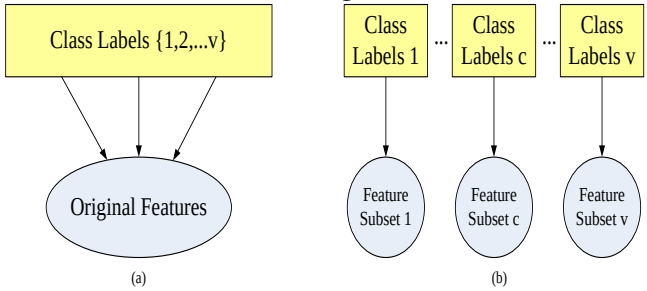

Figure 1: The simplified structures used to encode the dependencies of features and class labels. (a) The relevance between the original features and multiple class labels is not separated; (b) The relevance between each feature subset and its corresponding group of *possible class labels* is separated.

To evaluate objectively the intersections among feature subsets, a Coincidence Degree formula is used:

$$CDFS_k = \frac{1}{(L-1)^2 N} \sum_{i=1}^{L} \sum_{j=1}^{L} \sum_{k=1}^{N} (1 - |A_{i,j}(k)|) \quad i \neq j \tag{1}$$

$$A_{i,j}(k) = \begin{cases} \dfrac{Pos_i(f_i(k)) - Pos_j(f_i(k))}{N} & f_i(k) \in f_j \\ 1 & otherwise \end{cases}$$

where L is the number of possible class labels, N is the number of selected features, and Pos_j is the position of the specific feature in the *jth* feature subset. $f_i(k)$ refers to the *kth* feature in the *ith* feature subset, and $A_{i,j}(k)$ refers to the distance between Pos_i and Pos_j. The maximum coincidence degree is equal to 1 when the index and position of feature subsets are exactly the same.

Apart from the *first-order* approaches which break down multi-label learning tasks into multiple binary classification problems [26], the REAL framework can be applied to other multi-label approaches.

3. Experiments

In this section, a series of experiments is carried out to evaluate the effectiveness of the proposed method. A brief description of the two real-world data sets and the evaluation criteria is given. The experimental results are then presented and discussed.

3.1 Dataset Description and Configuration

Two common multi-label text data sets are used in the experiment:

- *Enron Data Set. Enron* is a subset of the Enron e-mail corpus . It is already labeled with a hierarchical set of categories developed by the UC Berkeley Enron E-mail Analysis Project2. The label categories take the form of a checklist in which there is high cardinality. Figure 3 shows the characteristics of a small sample of the Enron data set.
- *Medical Data Set. Medical* is medical-text data set compiled for the Computational Medicine Centers 2007 Medical Natural Language Processing Challenge3. It is already labeled with insurance codes. Each sample document includes a

brief free-text summary of patients' symptom history and their prognosis. For example, in radiology reports, ICD-9-CM codes serve as indications that a certain procedure will be performed. There are official guidelines for radiology ICD-9-CM coding. One guideline is that every disease code should have a minimum number of digits before reimbursement will occur. A definite diagnosis should always be coded whenever possible, whereas an uncertain one should never be coded. Symptoms must also never be coded unless there is a definite diagnosis [27].

Let $|S|$, $dim(S)$, $L(S)$ respectively denote the number of *instances*, number of *features*, number of *possible class labels*. In addition, several other multi-label properties are denoted as:

(a) $LCard(S) = \frac{1}{T}\sum_{i=1}^{T}|Y_i|$: label cardinality which is used to determined the average number of labels per example;

(b) $LDen(|S|) = \frac{LCard(S)}{L(S)}$: label density which normalizes *LCard* (*S*) by the number of possible labels;

(c) $DL(S) = |\{Y \mid \exists x : (x,Y) \in S\}|$: distinct label set which counts the number of distinct label combinations appeared in the data set;

(d) $PDL(S) = \frac{DL(S)}{|S|}$: proportion of distinct label set which normalizes *DL*(*S*) by the number of examples.

All statistic characteristics of the two multi-label data sets are shown in Table 1.

Table 1: Characteristics of the experimental data sets.

Data set	$\|S\|$	$dim(S)$	$L(S)$	$LCard(S)$	$LDen(S)$	$DL(S)$	$PDL(S)$
enron	1702	1001	53	3.378	0.064	753	0.442
medical	978	1449	45	1.245	0.028	94	0.096

3.2 Evaluation Measures

Given a test set $\Gamma = \{(x_i, Y_i) \mid 1 \le i \le m\}$, the following evaluation metrics designed specifically for multi-label learning are used in [3]:

(1) *Average precision*: evaluates the average fraction of labels ranked above a particular label y ∈ Y. The performance is perfect when $avgprec_{\Gamma}(f) = 1$.

$$avgprec_{\Gamma}(f) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{|Y_i|}\sum_{y \in Y_i}\frac{\left|\{y' \mid rank_f(x_i, y') \le rank_f(x_i, y), y' \in Y_i\}\right|}{rank_f(x_i, y)} \quad (2)$$

(2) *Coverage*: evaluates how far we need to go down a list of labels on the average, in order to cover all proper labels of the instance. It is related loosely to precision at the level of perfect recall.

$$coverage_{\Gamma}(f) = \frac{1}{m}\sum_{i=1}^{m}\max_{y \in Y_i} rank_f(x_i, y) - 1 \quad (3)$$

(3) *Hamming loss*: evaluates how many times instance-label pairs are misclassified, i.e., a label not belonging to the instance is predicted or a label belonging to the instance is not predicted.

$$hloss_{\Gamma}(f) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n}|f(x_i) \Delta Y_i| \quad (4)$$

where Δ stands for the symmetric difference between two sets.

(4) *One-error*: evaluates how many times the top-ranked label is not in the set of proper labels of the instance. The performance is perfect when $one-error_{\Gamma}(f) = 0$.

$$one-error_{\Gamma}(f) = \frac{1}{m}\sum_{i=1}^{m}\left\langle [\arg\max_{y \in Y} f(x_i, y)] \notin Y_i \right\rangle \quad (5)$$

For any predicate π, $\langle \pi \rangle$ equals 1 if π holds and 0 otherwise. Note that for single-label classification problems, the *one-error* is identical to the ordinary classification error.

(5) *Ranking loss*: evaluates the average fraction of label pairs that are reversely ordered for the instance. The performance is perfect when $rloss_{\Gamma}(f) = 0$.

$$rloss_{\Gamma}(f) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{|Y_i|\overline{|Y_i|}}\left|\{(y_1, y_2) \mid f(x_i, y_1) \le f(x_i, y_2), (y_1, y_2) \in Y_i \times \overline{Y_i}\}\right| \quad (6)$$

where $\overline{Y}$ denotes the complementary set of Y_i in Y.

Remarks:

The *higher* the value of $avgprec_{\Gamma}(f)$, the *better* the performance. The *smaller* the values of $coverage_{\Gamma}(f)$, $hloss_{\Gamma}(f)$, $one-error_{\Gamma}(f)$, and $rloss_{\Gamma}(f)$, the *better* the performance.

3.3 Comparison and Analysis

Based on the REAL scheme, the top n features relevant to each group of possible class labels are selected using CMIM, where n = [5, 10, 50, 100, 500] is divided to enron and medical data sets. Table 2 shows the comparative results.

In the experiments, the symbol "↓" indicates "the smaller the better," whereas "↑" indicates "the higher the better." The best results are represented in bold. Ten-fold cross-validation is employed on both data sets to ensure a reliable prediction. As shown in Table 2, the accuracy is reduced with the development of the dimensionality of features in each feature subset. The highest level of accuracy is achieved when the number n for the *enron* data set is approximately equal to 50, and that for the medical data set is equal to 5. The main features relevant to the target group of possible class labels can better predict unknown instances.

Furthermore, when L = 53 and N = 50, $CDFS_{enron}$ = 0.068; when L = 45 and N = 5, $CDFS_{medical}$ = 0.017. The result reveals that the intersection of selected features in each feature subset is weak. Dividing the original features into a specific feature subset according to the number of possible class labels is crucial to the success of the framework.

Based on the optimal dimension of features in each feature subset, the data sets are evaluated further using different *filter* selection mechanisms. Table 3 reports the experimental results in detail.

There are not that many differences in the results of using various *filter* selection mechanisms except for the MIFS method. This finding indicates the robustness of our REAL algorithm. The configurable parameter β, which is only set randomly, may have caused the fluctuations in the results of the MIFS method.

Finally, to verify the superiority of our framework, it is compared with four other popular multi-label learning algorithms: ML-kNN, BSVM, RAkEL, and ECC. For REAL approach, the experimental results of *enron* and *medical* data sets are based on 50 features and 5 features selected with CMIM criterion, respectively. For REAL and ML-kNN, the number of nearest neighbors is fixed at 10 in [16]. For BSVM, the models are obtained through cross-training strategy in [4]. For RAkEL, the parameter of the random subset of size k is incremented from 2 (the minimum value), the threshold is set to 0.5, and the

ensemble iterations are set to 10 [28]. For ECC, the kernel type is chosen for Linear in LIBSVM [29], and ensemble iterations are set to 10 in [1]. Table 4 shows the experimental results in detail.

From the *enron* data set, REAL ranks first place in terms of *average precision*, *coverage*, *hamming loss*, and *ranking loss*. However, it ranks second in the *one-error* criterion. For the *medical* data set, REAL ranks first place in terms of *average precision*, *coverage*, and *ranking loss*. However, it ranks second in the *hamming loss* and *one-error* criteria. The inferior ranks of the REAL scheme did not appear in the comparative experiment. These results indicate that our approach outperforms other multi-label learning schemes.

Table 2: Experimental results of REAL on the *enron*/ *medical* data set (mean±std).

Evaluation criterion	Data set	Dimension of features in each feature subset				
		n=5	*n*=10	*n*=50	*n*=100	*n*=500
Average Precision↑	*enron*	0.680±0.022	0.693±0.018	**0.694±0.021**	0.690±0.023	0.652±0.026
	medical	**0.886±0.018**	0.875±0.039	0.856±0.021	0.833±0.019	0.805±0.018
Coverage↓	*enron*	0.213±0.016	0.208±0.013	**0.211±0.019**	0.214±0.019	0.233±0.015
	medical	**0.041±0.007**	0.050±0.017	0.059±0.019	0.063±0.017	0.064±0.018
Hamming Loss↓	*enron*	0.050±0.003	0.050±0.002	**0.047±0.002**	0.048±0.002	0.050±0.003
	medical	**0.012±0.001**	0.013±0.002	0.017±0.002	0.022±0.001	0.025±0.009
One-error↓	*enron*	0.249±0.025	**0.237±0.035**	0.253±0.034	0.247±0.035	0.282±0.049
	medical	**0.150±0.035**	0.166±0.058	0.170±0.030	0.198±0.021	0.246±0.033
Ranking Loss↓	*enron*	0.074±0.008	**0.071±0.006**	0.072±0.008	0.074±0.008	0.085±0.009
	medical	**0.028±0.005**	0.032±0.015	0.041±0.012	0.044±0.011	0.045±0.013

Table 3: The experimental results of REAL using different filter selection approaches (mean±std) on the enron/ medical data set (50 features for enron / 5 features for medical).

Evaluation criterion	Data set	*Filter* selection approaches in *REAL*				
		CMIM	MIM	MRMR	MIFS	JMI
Average Precision↑	*enron*	**0.694±0.021**	0.686±0.022	0.677±0.021	0.648±0.026	**0.694±0.013**
	medical	0.886±0.018	0.880±0.021	0.884±0.027	**0.899±0.031**	0.885±0.017
Coverage↓	*enron*	**0.211±0.019**	0.214±0.014	0.233±0.027	0.249±0.019	0.213±0.018
	medical	0.041±0.007	0.042±0.010	0.046±0.013	0.045±0.013	**0.040±0.009**
Hamming Loss↓	*enron*	**0.047±0.002**	0.049±0.002	0.049±0.003	0.051±0.002	0.049±0.002
	medical	**0.012±0.001**	**0.012±0.001**	**0.012±0.001**	**0.012±0.002**	**0.012±0.001**
One-error↓	*enron*	0.253±0.034	**0.240±0.045**	0.254±0.040	0.287±0.037	0.246±0.028
	medical	0.150±0.035	0.162±0.035	0.151±0.045	**0.128±0.049**	0.153±0.030
Ranking Loss↓	*enron*	**0.072±0.008**	0.074±0.007	0.082±0.010	0.090±0.009	0.073±0.007
	medical	0.028±0.005	0.028±0.006	0.030±0.008	0.029±0.010	**0.027±0.006**

Table 4: The experimental results of each multi-label learning algorithm (mean±std) on the *enron*/ *medical* data set**.**

Evaluation criterion	Data set	Algorithm				
		REAL	ML-kNN	BSVM	RA*k*EL	ECC
Average Precision↑	*enron*	**0.694±0.021**	0.631±0.015	0.591±0.018	0.616±0.028	0.638±0.023
	medical	**0.886±0.018**	0.806±0.033	0.865±0.042	0.769±0.031	0.872±0.032
Coverage↓	*enron*	**0.211±0.019**	0.248±0.014	0.428±0.025	0.474±0.025	0.388±0.023
	medical	**0.041±0.007**	0.046±0.008	0.046±0.016	0.066±0.016	0.070±0.019
Hamming Loss↓	*enron*	**0.047±0.002**	0.052±0.002	0.060±0.002	0.097±0.069	0.056±0.004
	medical	0.012±0.001	0.022±0.001	0.011±0.002	0.036±0.026	**0.010±0.002**
One-error↓	*enron*	0.253±0.034	0.305±0.028	0.302±0.037	0.282±0.054	**0.226±0.031**
	medical	0.150±0.035	0.178±0.026	0.155±0.046	0.307±0.053	**0.098±0.031**
Ranking Loss↓	*enron*	**0.072±0.008**	0.092±0.009	0.180±0.011	0.201±0.016	0.242±0.021
	medical	**0.028±0.005**	0.038±0.018	0.031±0.013	0.047±0.010	0.099±0.031

4. Conclusions

In the present work, feature filters for single-label classifiers have been integrated into multi-label learning tasks. The goal is to address the bias introduced by the interleaving of original features to each group of possible class labels. The novel REAL framework, which takes into account the correlations between the feature subset and the target class labels, presents a new way of handling multi-label learning tasks. The extensive comparative results confirm the effectiveness of our scheme.

However, our proposed mechanism may not be exactly suitable for some data sets in which the coincidence degrees are high in original features. Except for exploiting the inferred correlations, further research on the incorporation of inter-label dependences into the REAL framework is needed. In addition, we will attempt to mine a rule that can be used to determine the optimal number of features in the feature subset corresponding to each group of possible class labels.

References

1. J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier chains for multi-label classification, In: W Buntine, M Grobelnik, and J Shawe-Taylor, eds, Lecture Notes in Artificial Intelligence 5782, Springer, Berlin, pp. 254-269, 2009

2. A.K. McCallum, “Multi-label text classification with a mixture model trained by EM,” In: Working notes of the AAAI’99 Workshop on Text Learning, Orlando, FL, 1999
3. R.E. Schapire, Y. Singer, “Boostexter: a boosting-based system for text categorization,” Machine Learning, 39 (2-3), pp. 135-168, 2000
4. M.R. Boutell, J. Luo, X. Shen, C. M. Brown, “Learning multi-label scene classification,” Pattern Recognition, 37(9), pp. 1757-1771, 2004.
5. K. Crammer, Y. Singer, “A family of additive online algorithms for category ranking,” In: Proc. of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 151-158, 2002
6. G.J. Qi, X.S. Hua, Y. Rui, J. Tang, T. Mei, H.J. Zhang, “Correlative multi-label video annotation,” In: Proceedings of the 15th ACM International Conference on Multimedia, Augsburg, Germany, pp. 17-26, 2007
7. A. Elisseeff, J. Weston, “A kernel method for multi-labelled classification,” In: T G Dietterich, S Becker, Z Ghahramani, Eds. Advances in Neural Information Processing Systems, Vol. 14, MIT Press, Cambridge, MA, pp. 681-687, 2002
8. G. Tsoumakas, I. Katakis, “Multi-label classification: An overview,” In: International Journal of Data Warehousing and Mining, 3(3), pp. 1-13, 2007
9. G. Nasierding, G. Tsoumakas, A.Z. Kouzani, “Clustering Based Multi-Label Classification for Image Annotation and Retrieval,” In: IEEE International Conference on Systems, Man and Cybernetics, San Antonio, TX, Vols. 1-9, pp. 4514-4519, 2009
10. T. Goncalves, P. Quaresma, “A preliminary approach to the multilabel classification problem of Portuguese juridical documents,” In: 11th Portuguese Conference on Artificial Intelligence, Beja, Portugal, Vol. 2902, pp. 435-444, 2003
11. B. Lauser, A. Hotho, “Automatic Multi-label Subject Indexing in a Multilingual Environment,” In: Proceedings of the 7th European Conference in Research and Advanced Technology for Digital Libraries, Trondheim, Norway, Vol. 2769, pp. 140-151, 2003
12. S. Diplaris, G. Tsoumakas, P.A. Mitkas, I. Vlahavas, “Protein classification with multiple algorithms,” In: 10th Panhellenic Conference on Informatics (PCI 2005), Volos, Greece, Vol. 3746, pp. 448-456, 2005
13. G. Tsoumakas, I. Vlahavas, “Random k-labelsets: An ensemble method for multilabel classification,” In: Proceedings of the 18th European Conference on Machine Learning (ECML), Warsaw, Poland, Vol. 4701, pp. 406-417, 2007
14. A. Clare, R. D. King, “Machine learning of functional class from phenotype data,” Bioinformatics, 20(1), pp. 160-166, 2002.
15. M.L. Zhang, Z.H. Zhou, “Multi-label neural networks with applications to functional genomics and text categorization,” IEEE Transactions on Knowledge and Data Engineering, 18(10), pp. 1338-1351, 2006
16. M.L. Zhang, Z.H. Zhou, “ML-kNN: A lazy learning approach to multi-label learning,” Pattern Recognition, 40 (7), pp. 2038-2048, 2007
17. F.A. Thabtah, P. Cowling, Y.H. Peng, “Multiple labels associative classification,” Knowledge and Information Systems, 9 (1), pp. 109-129, 2006
18. M.L. Zhang, “LIFT: multi-label learning with label-specific features,” In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI'11), Barcelona, Spain, in press, 2011
19. J. Read, A. Bifet, G. Holmes, B. Pfahringer, “Efficient multi-label classification for evolving data streams,” Technical Report. University of Waikato. New Zealand. URL: http://www.tsc.uc3m.es/~jesse/, 2010
20. F. Fleuret, “Fast binary feature selection with conditional mutual information,” Journal of Machine Learning Research, 5, pp. 1531-1555, 2004
21. R. Battiti, “Using mutual information for selecting features in supervised neural net learning,” IEEE Trans. Neural Networks, 5(4), pp. 537-550, 1994
22. H.C. Peng, F.H. Long, C. Ding, “Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8), pp. 1226-1238, 2005
23. H.H. Yang, J. Moody, “Data visualization and feature selection: new algorithms for nongaussian data,” In: 13th Annual Conference on Neural Information Processing Systems (NIPS), Vol. 12, pp. 687-693, 2000
24. N. Kwak, C. Choi, “Input feature selection for classification problems,” IEEE Transactions on Neural Networks, 13(1), pp. 143-159, 2002
25. G. Brown, “A new perspective for information theoretic feature selection,” In: Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS), Clearwater Beach, Florida, USA, Vol. 5 of JMLR, pp. 49-56, 2009
26. M.L. Zhang, K. Zhang, “Multi-label learning by exploiting label dependency,” In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'10), Washington DC, pp. 999-1007, 2010
27. J. Pestian, C. Brew, P. Matykiewicz, D. Hovermale, N. Johnson, K.B. Cohen, and W. Duch, “A shared task involving multi-label classification of clinical free text,” In ACL, editor, Proceedings of ACL BioNLP, Prague. Association of Computational Linguistics, 2007
28. J. Read, B. Pfahringer, G. Holmes, “Multi-label Classification using Ensembles of Pruned Sets,” In: 8th IEEE International Conference on Data Mining (ICDM), Pisa, pp. 995-1000, 2008
29. C.C. Chang, C.J. Lin, “LIBSVM: A library for support vector machines,” Software available at http://www.csie.ntu.edu.tw/˜cjlin/ libsvm, 2001

Author Profile

Jianjun Yan is an Associate Professor at Huadong University of Science and technology. He received his PhD from Huazhong University of Science and technology. From 2004 His research focus is in machine learning and biological signal processing.

Rui Guo is a Research Associate at Shanghai University of Traditional Chinese Medicine. She received her PhD from Shanghai University of Traditional Chinese Medicine. Her research focus is in machine learning.