DOI:10.3969/j.issn.1673-4785.201403064
网络出版地址:http://www.cnki.net/kcms/doi/10.3969/j.issn.16734785.201403064.html

一种多标记数据的过滤式特征选择框架

郭雨萌,李国正
(同济大学 电子与信息工程学院控制系,上海 201804)

摘 要:提出一种过滤式的多标记数据特征选择框架,并在卡方检验基础上进行实现和实验研究。该框架计算每个特征在各个类标上的卡方检验,然后通过得分的统计值计算出每个特征的最终排序情况,选取了最大、平均、最小 3 种统计值分别进行了实验比较。在 5 个评价指标、4 个常用的多标记数据集和 3 个学习器上的对比实验表明,3 种得分统计方式各有优劣,但都能提高多标记学习的效果。

关键词:特征选择;多标记;过滤式;卡方检验
中图分类号: TP391 **文献标志码**:A **文章编号**:1673-4785(2014)03-0292-06

中文引用格式:郭雨萌,李国正. 一种多标记数据的过滤式特征选择框架[J]. 智能系统学报, 2014, 9(3): 292-297.
英文引用格式:GUO Yumeng, LI Guozheng. A filter framework of the multi-label feature selection[J]. CAAI Transactions on Intelligent Systems, 2014, 9(3): 292-297.

A filtering framework for the multi-label feature selection

GUO Yumeng, LI Guozheng
(School of Electronic and Information Engineering, Tongji University, Shanghai 201804, China)

Abstract: The researchers of multi-label learning mainly focus on the classifier performance, regardless of the influence of the dataset feature. This paper proposes a filter framework of the multi-labeled data feature selection. The algorithm implementation and experiment were carried out based on the Chi-square test. This framework calculates the CHI-square test for each feature on each label, and then the ranking order of each feature is computed by the statistics of the score. This paper considers three different types of statistical data (average, maximum, minimum) for the experimental comparisons. The contrasting experiments with the four common multi-label datasets with three classifiers and five evaluation criteria show that these three score statistical methods share both superior and inferior characteristics, but still improve the performance for multi-label learning problems.

Keywords: feature selection; multi-label; filter; CHI-square test

多标记数据[1]中每个样本可以同时带有多个类标,并且广泛地出现在不同的应用领域,比如文本分类、媒体标注、信息检索、生物信息学等。对于这种数据的分析需要利用多标记学习技术[2-3]。由于大量不同的多标记学习技术被提出,所以该技术仍是研究热点,目前可以分为问题转化和算法适应 2 种类型。在问题转化类型中,BR(binary relevance),CC(classifier chain)和 RAkEL(random k-labelsets)分类器是典型代表。而在算法适应类型中,MLkNN(multi-label k nearest neighbor)、AdaBoost.MH(adaboost multi-class hamming trees)和 RankSVM(rank support vector machine)属于将一些先进的单标记分类器转化为多标记分类器的一类。LEAD(multi-label learning by exploiting label dependency)和 LIFT(multi-label learning with label-specific features)分类器则更进一步,考虑到特征子集和利用类标的层级

收稿日期:2014-03-25. 网络出版日期:2014-06-14.
基金项目:国家自然科学基金资助项目(61273305).
通信作者:李国正. E-mail:gzli@tongji.edu.cn.

结构去进行学习分类的一类。多标记学习技术发展的动力来自于实际应用问题,很具有研究价值。

虽然多标记学习技术还需要许多研究工作,但是很少的科研工作者将目光转向数据集中一些不相关或冗余的特征。减少这些特征会在一定程度上提高多标记学习器的分类能力,因此对数据集进行特征选择预处理是很有必要的。特征选择[4-5]的目的是在高维数据中降低子集维度,主要有过滤式、包装式和嵌入式等 3 种不同形式。过滤式与目标学习器无关,具有计算简单,效率高的优势[6-7]。本文提出一种过滤式多标记特征选择的框架,并以卡方检验[8]为特征评价的准则。

1　过滤式多标记特征选择框架

过滤式方法的基本思想是使用一种独立于分类器的评价指标来衡量某个特征的好坏,即选择该特征优先级。过滤式方法在计算效率上往往优于其他 2 种特征选择方法。

卡方检验可以用来度量特征 t 和类标 c 之间的相关程度。假设 t 和 c 之间符合具有一阶自由度的 CHI 分布。t 和 c 的 CHI 值由式(1)计算:

$$\chi^2(t,c) = \frac{N(AD - BC)^2}{(A + C)(B + D)(A + B)(C + D)} \tag{1}$$

式中:χ^2 值表示 CHI 值,N 表示数据集中样本的总个数;A 表示包含 t 且属于分类 c 的样本数;B 为包含 t 但是不属于 c 类的样本数;C 表示属于 c 类但是不包含 t 的样本数;D 表示既不属于 c 也不包含 t 的样本数。可以看出 N 固定不变,$A + C$ 为属于 c 类的样本数,$B + D$ 为不属于 c 类的样本数,所以式(1)可以简化为

$$\chi^2(t,c) = \frac{(AD - BC)^2}{(A + B)(C + D)}$$

当特征和类标相互独立时,$\chi^2(t,c) = 0$。$\chi^2(t,c)$ 的值越大,特征 t 和类标 c 越相关。

本文提出的过滤式多标记特征选择框架的基本思想是:首先单独计算每个特征 t 与各个类标 c 的 CHI 值,然后再根据得分统计方式决定每个特征的最终得分,最后将特征按照最终得分进行降序排列,并进行前向搜索得到特征子集。

下面为通过计算每个特征 t 与各个类标 c 的 CHI 值,并根据得分统计方式得到最终得分的公式:

$$\chi_{\text{avg}}{}^2(t) = \frac{1}{m}\sum_{i=1}^{m}\chi^2(t,c_i) \tag{2}$$

$$\chi_{\max}{}^2(t) = \max_{1\leqslant i\leqslant m}\{\chi^2(t,c_i)\} \tag{3}$$

$$\chi_{\min}{}^2(t) = \min_{1\leqslant i\leqslant m}\{\chi^2(t,c_i)\} \tag{4}$$

式中 m 为类标个数。式(2)表示特征与各类标的平均 CHI 值作为该特征的最终得分;式(3)表示选取特征与各类标 CHI 值中的最大值作为该特征的最终得分统计;式(4)表示选取特征与各类标 CHI 值中的最小值作为该特征的最终得分统计。

实验数据来自于 MULAN 网站上公开的多标记数据集,数据集相关信息如表 1 所示。

表 1　实验数据集相关信息

Table 1　The characteristics of datasets

数据集名称	所属领域	样本个数	属性个数	类标个数
emotions	音乐	593	72	6
medical	文本	978	217	20
Scene	图像	2407	249	6
yeast	生物	2417	103	14

实验采用 5 种常用的多标记学习评价指标[9],对多标记数据特征选择之后的分类性能进行评价:排名损失、汉明损失、差一错误、覆盖范围、平均查准率。以上 5 种评价指标中,前 4 种评价指标的值越小,最后 1 种评价指标的值越大,表明性能越好。

实验采用 10 轮 10 倍交叉验证方法,即将实验数据随机平均分成 10 份,每次将 1 份作为验证集,其余 9 份整体作为训练集,不重复进行 10 次实验,统计其平均结果,作为实验最终结果。

通过将预处理后的多标记数据集利用卡方检验准则,可以分别得到每个特征 t 对应的各个类标 c 的 CHI 值。然后,按照不同的得分统计方式得到每个特征的最终得分,最后根据每个特征的最终得分,将全体特征做降序排列,使用前向搜索依次选取前 n 个特征(n =1,2,…)作为特征子集。

max 指的是选取利用卡方检验准则得到的每个特征对应各个类标所有 CHI 值的最大值,作为该特征的最终得分,进行特征排序。

avg 指的是选取利用卡方检验准则得到的每个特征对应各个类标所有 CHI 值的平均值,作为该特征的最终得分,进行特征排序。

min 指的是选取利用卡方检验准则得到的每个特征对应各个类标所有 CHI 值的最小值,作为该特征的最终得分,进行特征排序。

在将处理好的特征进行排序后,多标记分类器将利用搜索到的特征子集去完成分类任务。为了更加客

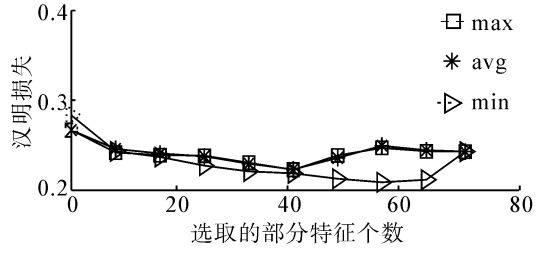
观地测试特征子集的分类效果,实验选取了 3 个多标记分类器,分别是 BR[10]、CC[11]和 MLkNN[12]。

3 实验结果及分析

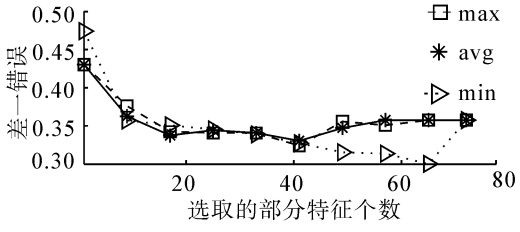
按照上节的实验设置,在 4 个公开数据集上先进行特征选择,再分类,实验结果做如下分析。

3.1 Emotions 数据集上的实验结果分析

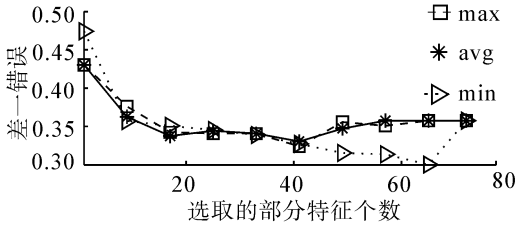
如图 1(其中横轴坐标表示特征子集所含有的特征个数,纵轴坐标表示特征子集在相应指标下的实验结果数值,之后分析相同)和表 2 所示,在 BR 分类器下,随着特征个数增多到最后阶段 3 种得分统计方式搜索到的特征子集性能较差。虽然开始在 min 下搜索到的特征子集相比于其他 2 种方式,在 5 种评价指标下性能较差,但是随着特征个数的增加,min 下的实验结果渐渐超过 avg 和 max,最终达到全局最优,得到最优特征子集。而且 avg 和 max 下搜索得到的特征子集除了在差一错误评价指标下的实验结果存在较明显差异,在其余 4 种评价指标下预测结果差异较小。同时,可以看出在 CC 分类器下,整体趋势与 BR 分类器下相似,但是后期波动较小。在 MLkNN 分类器下,整体趋势与 BR 分类器下相似,但是后期波动较大。



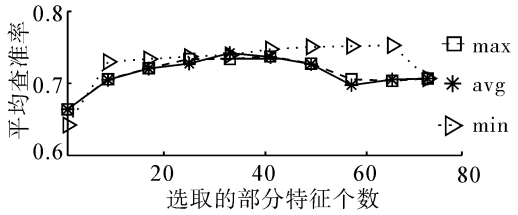
(a)在 BR 分类器下的汉明损失值



(b) 在 BR 分类器下的差一错误值



(c) 在 CC 分类器下的汉明损失值



(d) 在 MLkNN 分类器下的平均查准率值

图 1 Emotions 数据集部分实验结果

Fig.1 Partial results of the experiment on the emotions dataset

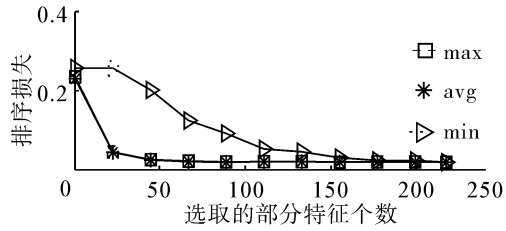
表 2 Emotions 数据集实验的最优结果比较

Table 2 Comparison of optimal results of the experiment on the emotions dataset

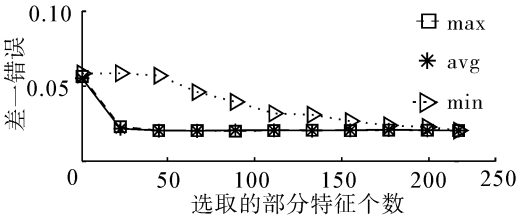
分类器	得分方式	排序损失	汉明损失	差一错误	覆盖范围	平均查准率
BR	max	0.1931	0.2211	0.3220	1.9324	0.7657
	avg	0.1939	0.2209	0.3237	1.9423	0.7656
	min	0.1738	0.2088	0.3001	1.8350	0.7849
CC	max	0.1991	0.2316	0.3304	1.9290	0.7608
	avg	0.1991	0.2305	0.3288	1.9524	0.7640
	min	0.1796	0.2105	0.3118	1.8381	0.7810
MLkNN	max	0.2260	0.2448	0.3625	2.0714	0.7395
	avg	0.2197	0.2450	0.3505	2.0529	0.7451
	min	0.2057	0.1977	0.3456	1.9716	0.7535

3.2 Medical 数据集上的实验结果分析

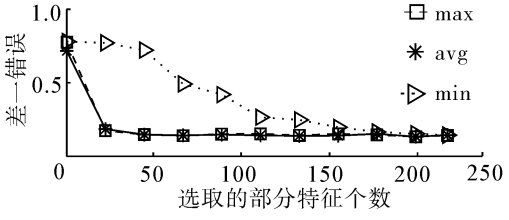
如图 2 和表 3 所示,在 BR 分类器下,avg 和 max 2 种得分统计方式搜索到的特征子集在 5 种评价指标下预测结果差异较小,几乎重叠在一起。但是从全局最优结果看,在排序损失和覆盖范围指标下,avg 和 max 都能搜到最优特征子集,而在汉明损失和差一错误指标下,avg 结果最好,在平均查准率指标下,max 结果最好。在 min 下搜索到的特征子集在 5 种评价指标下结果最差,而且收敛速度明显慢于 avg 和 max,特征选择对于分类性能提升效果较差。同时,可以看出在 CC 分类器下,整体趋势与 BR 分类器下相似。但是从全局最优结果看,在 5 种指标下,max 下搜索到最优特征子集,结果最好。在 MLkNN 分类器下,整体趋势与 BR 分类器下相似。



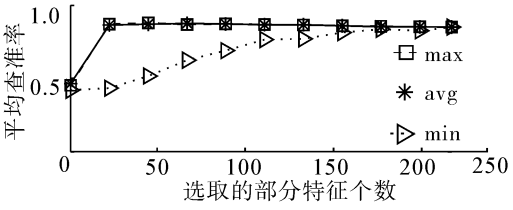
(a)在 BR 分类器下的排序损失值



(b) 在 BR 分类器下的汉明损失值



(c) 在 CC 分类器下的差一错误值



(d) 在 MLkNN 分类器下的平均查准率值

图 2　Medical 数据集部分实验结果

Fig.2　Partial results of the experiment on the medical dataset

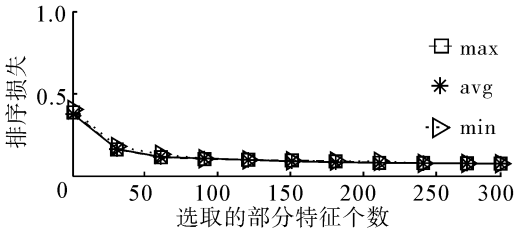
表 3　Medical 数据集实验的最优结果比较

Table 3　Comparison of optimal results of the experiment on the medical dataset

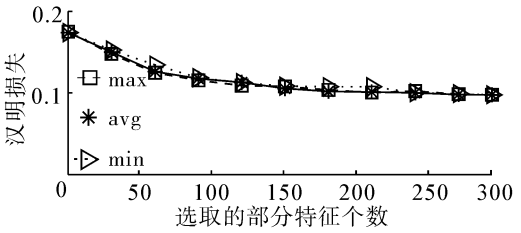
分类器	得分方式	排序损失	汉明损失	差一错误	覆盖范围	平均查准率
BR	max	0.0187	0.0196	0.1244	0.5697	0.9174
	avg	0.0187	0.0195	0.1170	0.5697	0.9165
	min	0.0188	0.0202	0.1298	0.5707	0.9148
CC	max	0.0207	0.0186	0.1266	0.6219	0.9147
	avg	0.0229	0.0188	0.1277	0.6482	0.9119
	min	0.0244	0.0199	0.1373	0.6851	0.9067
MLkNN	max	0.0309	0.0204	0.1360	0.8051	0.9014
	avg	0.0359	0.0225	0.1645	0.9189	0.8806
	min	0.0417	0.0294	0.2088	1.0344	0.8535

3.3　Scene 数据集上的实验结果分析

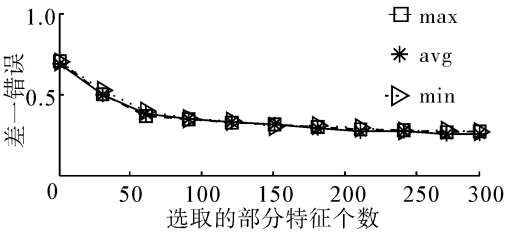
如图 3 和表 4 所示,在 BR 分类器下,3 种得分统计方式搜索到的特征子集在 5 种评价指标下预测结果差异较小,几乎重叠在一起。但是从全局最优结果看,在排序损失指标下,3 种得分统计方式达到相同结果,在汉明损失,覆盖范围和差一错误指标下,min 结果最好,在平均查准率指标下,max 结果最好。同时,可以看出在 CC 分类器下,整体趋势与 BR 分类器下相似。但是从全局最优结果看,在 5 种指标下,avg 下搜索到最优特征子集,结果最好。在 MLkNN 分类器下,整体趋势与 BR 分类器相似。但是从全局最优结果看,在 5 种指标下,min 下搜索到最优特征子集结果最好。



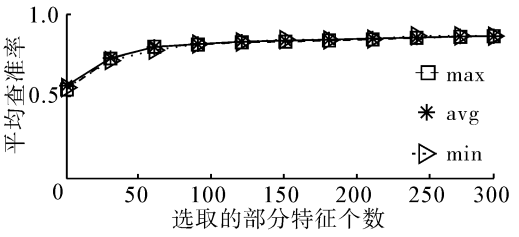
(a)在 BR 分类器下的排序损失值



(b) 在 BR 分类器下的汉明损失值



(c) 在 CC 分类器下的差一错误值



(d) 在 MLkNN 分类器下的平均查准率值

图 3　Medical 数据集部分实验结果

Fig.3　Partial results of the experiment on the medical dataset

表 4　Scene 数据集实验的最优结果比较

Table 4　Comparison of optimal results of the experiment on the scene dataset

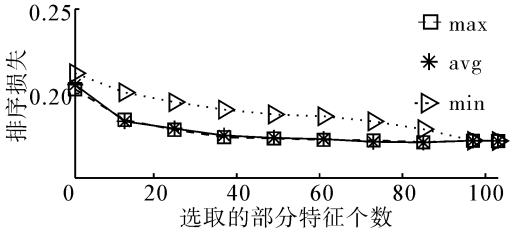
分类器	得分方式	排序损失	汉明损失	差一错误	覆盖范围	平均查准率
BR	max	0.0752	0.0970	0.2302	0.4595	0.8642
	avg	0.0752	0.0973	0.2326	0.4595	0.8629
	min	0.0752	0.0968	0.2088	0..4590	0.8629

续表 1

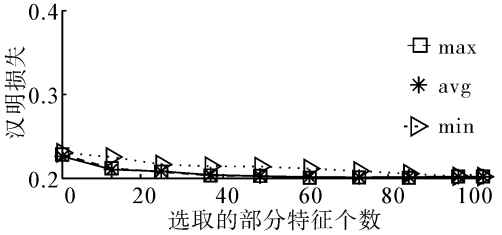
分类器	得分方式	排序损失	汉明损失	差一错误	覆盖范围	平均查准率
CC	max	0.0871	0.0984	0.2584	0.5184	0.8461
	avg	0.0839	0.0973	0.2555	0.5010	0.8493
	min	0.0856	0.0984	0.2584	0.5114	0.8462
MLkNN	max	0.0769	0.0847	0.2235	0.4711	0.8669
	avg	0.0766	0.0843	0.2231	0.4707	0.8674
	min	0.0749	0.0834	0.2119	0.4611	0.8716

3.4 Yeast 数据集上的实验结果分析

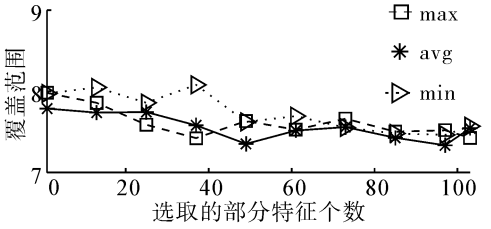
Yeast 数据集部分实验结果如图 4 所示。



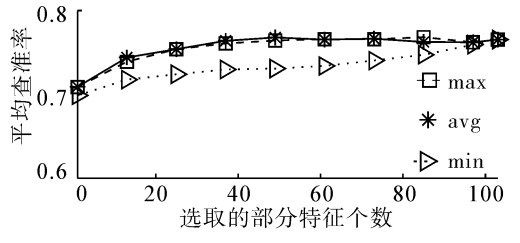
(a)在 BR 分类器下的排序损失值



(b) 在 BR 分类器下的汉明损失值



(c) 在 CC 分类器下的覆盖范围值



(d) 在 MLkNN 分类器下的平均查准率值

图 4 Yeast 数据集部分实验结果

Fig.4 Partial results of the experiment on the yeast dataset

在 BR 分类器下,avg 和 max 两种得分统计方式搜索到的特征子集在排序损失、汉明损失和平均查准率指标下预测结果差异较小,几乎重叠在一起,但是在差一错误和覆盖范围指标下,都出现不同程度的小幅震荡。在 min 下搜索到的特征子集在 5 种评价指标下结果最差,而且收敛速度明显慢于 avg 和 max,特征选择对于分类性能提升效果较差。从全局实验结果看,avg 下搜索到的特征子集,达到最优结果。同时,可以看出在 CC 分类器下,3 种取值方式搜索到的特征子集,在 5 种评价指标下的结果,都呈现出震荡的形式,尤其是在差一错误指标下,震荡幅度最大。虽然在震荡中,但是随着特征个数的增加,结果逐渐改善,说明特征选择起到了很好的提高分类性能的作用。从全局实验结果看,在排序损失和平均查准率指标下,avg 下搜索到的特征子集表现最好,而且其余 3 种评价指标下,max 下搜索到的特征子集表现最好。在 MLkNN 分类器下,整体趋势与在 BR 分类器下相似。从全局实验结果看,除了在排序损失和差一错误指标下,avg 与 max 下搜索到的特征子集,达到相同最优结果,其余 3 种评价指标下,max 的结果最好。Scene 数据集实验的最优结果比较如表 5 所示。

表 5 Scene 数据集实验的最优结果比较

Table 5 Comparison of optimal results of the experiment on the scene dataset

分类器	得分方式	排序损失	汉明损失	差一错误	覆盖范围	平均查准率
BR	max	0.0752	0.0970	0.2302	0.4595	0.8642
	avg	0.0752	0.0973	0.2326	0.4595	0.8629
	min	0.0752	0.0968	0.2088	0..4590	0.8629
CC	max	0.0871	0.0984	0.2584	0.5184	0.8461
	avg	0.0839	0.0973	0.2555	0.5010	0.8493
	min	0.0856	0.0984	0.2584	0.5114	0.8462
MLkNN	max	0.0769	0.0847	0.2235	0.4711	0.8669
	avg	0.0766	0.0843	0.2231	0.4707	0.8674
	min	0.0749	0.0834	0.2119	0.4611	0.8716

3.5 实验结果

从以上所有实验结果可以看出,针对不同类型的多标记数据集,都有其特定的得分统计方式能很快地搜索到较优的特征子集,然后趋于稳定,说明特征选择起到了很好的提高分类性能的作用。为了便于使展示图片美观易懂,画图时特征子集所含特征个数采用间隔选取再绘制(本身实验数据是全的),所有的同类型图片都采用这个方法。

4　结束语

本文提出过滤式的多标记特征选择框架,并使用卡方检验作为特征评价准则,在多个多标记数据集和分类评价准则上显示特征选择有助于提高多标记学习器的学习效果。本文通过对卡方检验得分的统计计算出每个特征的最终排序情况,选取了最大、平均、最小 3 种统计方式分别进行了实验比较。实验结果表明,利用本文框架采取不同的得分统计方式,对于不同类型的多标记数据集有不同效果。过滤式多标记特征选择框架还有一些问题有待进一步解决,比如如何在得分统计中加入衡量类标间的关系,如何采取更有效得分统计方式将提升特征子集在分类器下的分类效果等。

参考文献:

[1] TSOUMAKAS G, KATAKIS I, VLAHAVAS I. Mining Multi-label Data[R]. Data Minging and Knowledge Discovery Handbook, 2010: 667-685.

[2] TSOUMAKAS G, KATAKIS I. Multi-label classification: an overview[J]. International Journal of Data Wareh -ousing and Mining, 2007, 40(3): 1-13.

[3] ZHANG M L, ZHANG K. Multi-label learning by exploiting label dependency[C]//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC, USA, 2010: 999 -1008.

[4] YANG Y, PEDERSEN J O. A comparative study on feature selection in text categorization[C]//Machine Learning International Workshop then Conference. Philadelphia, USA, 1997: 412-420.

[5] SWATI S, GHATOL A, ASHOK C. Feature selection for medical diagnosis: Evaluation for cardiovascular diseases [J]. Expert Systems with Applications, 2013, 40(10): 4146-4153.

[6] NEWTON S, EVERTON A C, MARIA C M, et al. A comparison of multi -label feature selection methods using the problem transformation approach[J]. Electronic Notes in Theoretical Computer Science, 2013, 292: 135-151.

[7] 计智伟,胡珉,尹建新.特征选择算法综述[J].电子设计工程, 2011, 19(9): 46-51.
JI Zhiwei, HU Ming, YIN Jianxin. A survey of feature selection algorithm[J]. Electronic Design Engineering, 2011, 19(9): 46-51.

[8] 邱云飞,王威,刘大有,等.基于方差 CHI 的特征选择方法[J].计算机应用研究, 2012, 29(4): 1301-1303.
QIU Yunfei, WANG Wei, LIU Dayou, et al. CHI feature selection method based on variance[J]. Application Research of Computers, 2012, 29(4): 1301-1303.

[9] ZHANG M L, ZHOU Z H. A review on multi-label learning algorithms[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 39(10): 1-43.

[10] MATTHEW R B, LUO J B, SHEN X P, et al. Learning multi-label scene classification[J]. Pattern Recognition, 2004, 37(9): 1757-1771.

[11] READ J, PFAHRINGER B, HOLMES G, et al. Classifier chains for multi-label classification[J]. Machine Learning, 2011, 85(3): 333-359.

[12] ZHANG M L, ZHOU Z H. ML-kNN: a lazy learning approach to multi-label learning[J]. Pattern Recognition, 2007, 40(7): 2038 -2048.

作者简介:

郭雨萌,男,1989 年生,博士研究生,主要研究方向为模式识别与机器学习等。

李国正,男,1977 年生,研究员,博士生导师,博士,中国人工智能学会机器学习专业委员会常务委员,主要研究方向为模式识别和生物医学数据挖掘,在研和完成国家自然科学基金项目、上海市科委“创新行动计划”重大项目子课题等多项课题,发表学术论文 100 余篇,其中 SCI 检索 40 余篇,EI 检索 50 余篇,参与撰写专著 6 部,主持翻译专著 1 部。