基于 kNN 的多标签分类预处理方法

徐晓丹[1,2] 姚明海[1] 刘华文[2] 郑忠龙[2]

(浙江工业大学信息工程学院 杭州 310023)[1] (浙江师范大学数理与信息工程学院 金华 321004)[2]

摘 要 多标签学习已成为当前机器学习的研究热点。为了提高分类性能,对训练集中的噪声数据进行预处理,提出一种基于 k 近邻(kNN)的多标签分类去噪方法:对现有的多标签数据集进行分析后获得近似正态分布的特征,通过将噪声标记改为其 k 近邻标记的方法,滤去部分噪声信息,从而得到相对高质量的数据集。在 MULAN 平台上使用多个数据集对 6 种多标签分类算法进行了噪声去除前后的对比测试,实验结果表明,多标签的预处理方法有效提高了分类器的性能。此方法对于分布特征明显的数据集具有较好的适用性。

关键词 多标签,分类,正态分布,预处理,kNN

中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.5.021

Pre-processing Method of Multi-label Classification Based on kNN

XU Xiao-dan[1,2] YAO Ming-hai[1] LIU Hua-wen[2] ZHENG Zhong-long[2]

(College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China)[1]

(College of Mathematics, Physics and Information Engineering, Zhejiang Normal University, Jinhua 321004, China)[2]

Abstract Multi-label learning is a new field in machine learning. In order to improve the multi-label classification precision, a new kNN method was used to remove the noise labels. First, a normal distribution is discovered by analyzing the characteristics of multi-label datasets, and then the high quality datasets are generated by changing the value of noisy labels to their k-Nearest Neighbors. In the experiments, six kinds of multi-label classification methods were tested on MULAN with new datasets. Compared to the primal datasets, the classification precision based on new datasets is better. Research results show this method is suitable for the data set which has a regular distribution.

Keywords Multi-label, Classification, Normal distribution, Pretreatment, kNN

1 引言

信息化程度的提高使得网络数据呈现指数级增长,如何有效地利用这些数据成为当前研究者普遍关注的问题。特别是在数据量不断增大的情况下,数据的标注结构愈发复杂,往往表现为多标签的形式,由此而兴起的多标签数据挖掘[1,2]的研究正日益凸显出其价值。同时该技术的应用领域也从文本分类、生物信息学拓展到场景分类、图像视频的语义标注[3,4]、音乐情感分类[5-7]、营销指导[8]等。

多标签分类问题可以描述为:给定一个集合 X 和一个标签集合 Y,其中 X 为 d 维的特征空间向量 R^d,Y 是包含 q 个标签的集合 $Y=\{y_1,y_2,y_3,\cdots,y_q\}$。包含 m 个数据的多标签数据集 D 可以表示为 $D=\{(x_i,Y_i)\mid 1\leqslant i\leqslant m\}$,其中 $x_i\in X$ 是一个 d 维的特征向量 $(x_{i1},x_{i2},\cdots,x_{id})^{\mathrm{T}}$,$Y_i\in Y$ 是 x_i 对应一个标签集合。多标签分类就是从多标签数据集 D 中获得一个学习模型 $h:\rightarrow 2^Y$,对于给定的任意一个实例 $x_i\in X$,模型 h 输出该实例对应的标签集合 Y_i。模型 h 是一个 X 到 Y 的映射函数,能给出未知类别数据样本的所有对应的标签。

当前的多标签分类的研究方法主要归为两类:1)将多标签的学习转化为多个单标签的学习,然后利用已有的成熟的单标签学习解决方案来进行[9],典型的代表有 BR[10](Binary Relevance),RPC[11](Ranking by Pairwise Comparison),LP[12](Label Powerset)等;2)修改已有的单标签学习算法,使之能适应多标签的学习[13,14],典型的方法有 BP-MLL[15](Back-Propagation for Multi-Label Learning),SVM(Support Vector Machine),MLkNN[16](Multi-Label k-Nearest Neighbor)以及多个方法的综合[17]等。在各种分类器算法中,如何提高分类精度是普遍关注的一个问题,由于训练样本集的质量是影响多标签分类算法精度的一个重要因素,本文从分析多标签数据集的自身特征出发,根据数据集体现出来的正态分布特征,利用 kNN 方法将噪声数据的标签值改为其最近邻的值,以提高分类器效果。

到稿日期:2014-02-18 返修日期:2014-04-15 本文受浙江省教育厅项目(Y201328291),浙江省自然科学基金项目(LZ14F030001,LY14F020012)资助。

徐晓丹(1978-),女,博士生,讲师,主要研究方向为数据挖掘、机器学习,E-mail:xuxiaodan@zjnu.cn;姚明海(1963-),男,博士,教授,主要研究方向为机器学习与智能;刘华文(1978-),男,博士,副教授,主要研究方向为数据挖掘、特征学习;郑忠龙(1976-),男,博士,教授,主要研究方向为模式分类、机器学习。

2 多标签数据集特征分析及预处理方法

2.1 多标签数据集特征分析

对于一个多标签分类器来说，其分类性能主要取决于两个因素：1)所采用的分类方法，在这方面学者们已经开展了广泛的研究，并通过不断改进，使得分类精度得到了很大的提高；2)训练样本集的质量，训练样本集类别越准确，内容越全面，得到的分类器质量就越高。而在实际应用中，训练集中往往包含着噪声，这主要有两方面的原因：1)是数据集本身的冗余性使得不一致样本的存在成为可能；2)人工对训练数据集进行标注时出现误标，这在大规模数据集中经常出现。这些噪声的存在将对分类结果产生重要的影响。对于一个分类器来说，在给定分类方法的情况下，数据集的质量便成为影响分类器性能的决定性因素。

从数据中学习是机器学习的主要特点，故得到数据分布的准确估计对学习效果有着至关重要的作用。尽管在实际问题中数据的精确分布往往很难得到，但若是能得到近似的分布特征，则对这些数据集的分布进行研究将会是一个提高分类效果的有效途径。

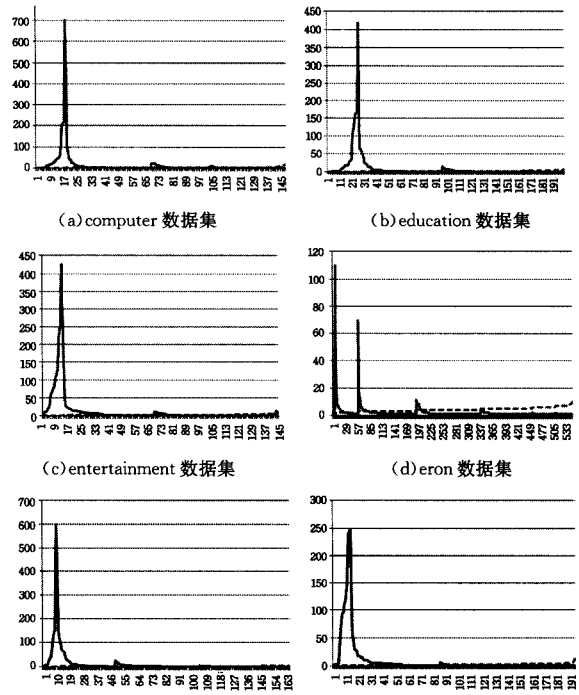
鉴于此，本文对 MULAN[18]、SourceForge[19] 上提供的近 20 个数据集进行了统计分析，得到的部分数据如表 1 所列。其中，domain 给出了数据源的信息，instances 栏代表样本的个数，labels 表示标签的个数，cardinality 表示数据集样本标签数的平均值，density 表示标签的密度。在表 1 中，rcvlv2 数据集是平均了 5 个子集的统计结果。

表 1 多标签数据集统计分析

name	domain	instances	labels	cardinality	density
bibtex	text	7395	159	2.402	0.015
birds	audio	645	19	1.014	0.053
corel5k	images	5000	374	3.522	0.009
delicious	text(web)	16105	983	19.02	0.019
emotions	music	593	6	1.869	0.311
enron	text	1702	53	3.378	0.064
genbase	biology	662	27	1.252	0.046
mediamill	video	43907	101	4.376	0.043
medical	text	978	45	1.245	0.028
rcv1v2(avg)	text	6000	101	2.65	0.026
scene	image	2407	6	1.074	0.179
yeast	biology	2417	14	4.237	0.303

从表 1 可以看到，每个数据集的标签密度值 density 都很低，基本分布在 0～0.1 之间，说明多标签数据集的数据分布具有较大的稀疏性；同时可以得到，虽然不同数据集的标签数目差别很大(最小为 6，最大为 983)，但每个数据集自身的平均标签值 cardinality 都在(1，5)范围内(除个别数据集外)。由此可以考虑每个数据集的标签是否有一定的分布规律。为此，对于某个数据集对应的标签集合 $Y=\{y_1,y_2,y_3,\cdots,y_q\}$，统计 Y_i 中的标签个数(称之为长度)以及 Y_i 在整个集合中的出现次数，得到长度和次数的关系特征。对每个数据集都做相应的统计，图 1 给出了部分数据集特征分布结果。

在图 1 中，y 轴表示某个标签集合 $Y_i=\{0,1\}^m$ 出现的次数，x 轴表示整个数据集合中不同标签集合的个数。用实线绘制的曲线表示不同标签的出现次数，虚线表示的曲线是标签长度。图 1 显示不同数据标签的出现次数呈现近似的正态分布，很好地验证了前面提出的数据集分布规律的特点。



(a) computer 数据集　(b) education 数据集

(c) entertainment 数据集　(d) eron 数据集

(e) health 数据集　(f) recreation 数据集

图 1 各数据集分布图

根据图 1 给出的分布规律，可以做如下假设：在正态分布曲线中，分布在两端的出现次数为 1 的标签集合 $Y_r=\{0,1\}^m$，其内部标记有可能部分错误，若是能将这个集合进行有效的修正，则可去除学习过程中的噪声信息，提高分类器效果。修正样本 x_r 所对应的标记集合 $Y_r=\{0,1\}^m$ 比较合理的一种方法是将其标记集合改为与它相关联的某个标记集合。由此，可以使用 kNN 方法来得到与 $Y_r=\{0,1\}^m$ 相关联的标签集合，用最近邻的标签集合 $Y_k=\{0,1\}^m$ 值来替代当前的标签集合。实验表明这是一种有效的方法，在接下来的第 2.2 小节中具体描述该方法的实现。

2.2 基于 kNN 的特征去噪

k 近邻方法(kNN)是目前理论上比较成熟的方法，主要用于分类。其基本思想是：如果 1 个样本的 k 个最近的邻居大部分属于某个类别，则该样本也属于这个类别，即根据最邻近的 1 个或几个样本的类别来决定待分类样本的类别。在 k 近邻算法中，所选择的邻居都是已经正确分类的对象。

为了测试方法的有效性，对于 k 近邻的选择，本文分别使用两种方法。

(1)样本 X 的近邻：根据样本 X 的距离来选择当前样本的 k 个邻居，在 k 个邻居中使用投票法选择某一个邻居，将当前样本的标签 y 改为邻居的标签值。如果 k 个邻居所对应的标签 y 都不相同，则取标签在整个样本集中出现次数最多的一个 y 的值作为当前样本的标签。

(2)标签 y 的近邻：y 的 k 个邻居都是不相同的，根据这 k 个邻居在样本集中的出现次数来选择当前标签要更改为哪个邻居的标签。同时做如下改进：在选择 k 个邻居进行距离排序时，如果距离值相同，则选择二进制值上接近的这个值排在前面。例 101010，与它距离值相等的有 101011、101110，则选择 101011 排在前面，这是基于如下假设：如果样本属于某个类的标记是被错误标记的，那么越靠近后面出错率越高。

本文使用上述方法得到新的标签集合 Y，然后将预处理

后的数据集合在 mulan 平台上使用已有的分类方法进行测试，并与原数据集进行对照分析，得到比较结果。本文分别取 k 的值为 1 和 3 进行了测试，第 3.3 节中表 5 的数据给出的是 $k=3$ 的实验结果。

3 实验及结果分析

3.1 实验数据集

实验使用通用数据集，这些数据可以在 mulan 网站[11]下载获得。本文使用 6 个数据集，其数据集的规模如表 2 所列。

表 2 实验数据集属性

数据集	样本个数	样本特征个数	标签个数
computer	5000	681	33
entertainment	5000	640	21
eron	1702	1001	53
health	5000	612	32
recreation	5000	606	22
yeast	2417	103	14

3.2 多标签分类评价指标

常用的多标签分类指标主要有以下几个。

(1)Hamming Loss[1]

用于计算分类器预测的结果与实际结果的平均差值，Hamming Loss 的值越小，说明分类器性能越好。Δ 表示预测集合与标准集合之间对应元素的差异，公式如下所示：

$$HammingLoss=\frac{1}{t}\sum_{i=1}^{t}\frac{|Y_i\Delta Z_i|}{m}$$

(2)One-Error

用于评价样本的标签排序序列中，序列最前端的标记不属于标记集合的情况。下式中，arg min $r_i(\lambda)$ 表示样本 x_i 的标签中排在最前面的标签。

$$One\text{-}Error=\frac{1}{m}\sum_{i=1}^{m}\delta(\arg\min r_i(\lambda)),\lambda\in Y_i$$

(3)Coverage

用于评价样本的标记排序序列中，覆盖属于样本的所有标记所需的搜索深度情况，该值越小，分类性能越好。max $r_i(\lambda)$ 表示样本 x_i 的标签中，排在最后面的标签。

$$Coverage=\frac{1}{m}\sum_{i=1}^{m}\delta(\max r_i(\lambda)-1),\lambda\in Y_i$$

(4)Ranking Loss

用于评价样本标记的排序中出现错误排序的情况，该值越小，分类效果越佳。

$$RankingLoss=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{|Y_i||\overline{Y}_i|}|\{(\lambda_a,\lambda_b):r_i(\lambda_a)>r_i(\lambda_b),(\lambda_a,\lambda_b)\in Y_i\times\overline{Y}_i\}|$$

(5)Average Precision

用于评价在样本的预测排序中，排在该样本标记之前的标记仍属于样本标记集合的情况，该值越大，性能越佳。

$$AvgPre=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{|Y_i|}\sum_{\lambda\in Y_i}\frac{|\{\lambda'\in Y_i:r_i(\lambda')\leqslant r_i(\lambda)\}|}{r_i(\lambda)}$$

3.3 实验结果分析

为了评测提出的预处理方法的有效性，使用 2.2 节的基于 kNN 的方法对数据集进行了去噪预处理，然后将预处理后的数据分别应用于常见的分类算法：CLR(Calibrated Label Ranking)、RAkEL(Random-k Labelsets)、LP(Label Powerset)、MLkNN(Multi-label k-nearest neighbor)、IncludeLabels、MLStacking(Multi-Label Stacking)，在 Mulan[11]平台上得到 eron 数据集的前后比较结果。表 3 给出了 k 值分别取 1 和 3 时，在 eron 数据集上 CLR 分类方法的比较结果。表 4 给出了 eron 数据集上各个分类方法的结果数据比较，实验中 k 的取值为 3。

表 3 k 取不同值时 eron 数据集上 CLR 方法的各个指标结果

CLR 方法	预处理前	k=1	k=3
Hamming Loss	0.024	0.017	0.014
One-Error	0.062	0.070	0.062
Coverage	4.514	3.385	2.554
Ranking Loss	0.018	0.013	0.010
Average Precision	0.88	0.900	0.917

表 4 eron 数据集预处理前后分类结果的比较

		Hamming Loss	One-Error	Coverage	Ranking Loss	Average Precision
CLR	处理前	0.024	0.062	4.514	0.018	0.88
	处理后	0.014	0.062	2.554	0.010	0.917
	提高	41.96%	0.00%	43.43%	43.52%	4.16%
MLkNN	处理前	0.046	0.225	8.224	0.05	0.715
	处理后	0.037	0.2	4.711	0.028	0.766
	提高	19.57%	11.11%	42.72%	44.00%	7.13%
Include-Labels	处理前	0.023	0.069	10.46	0.059	0.819
	处理后	0.017	0.063	6.855	0.039	0.854
	提高	26.09%	8.70%	34.46%	33.90%	4.27%
RAkEL	处理前	0.02	0.048	12.389	0.081	0.864
	处理后	0.014	0.049	8.465	0.057	0.9
	提高	30.00%	−2.08%	31.67%	29.63%	4.17%
LP	处理前	0.041	0.736	35.187	0.517	0.316
	处理后	0.027	0.537	25.583	0.369	0.5
	提高	34.15%	27.04%	27.29%	28.63%	58.23%
ML-Stacking	处理前	0.047	0.271	7.991	0.053	0.701
	处理后	0.039	0.26	5.764	0.04	0.735
	提高	17.02%	4.06%	27.87%	24.53%	4.85%

表 5 各个数据集预处理后提高的百分比(%)

分类方法	数据集	Hamming Loss	One-Error	Coverage	Ranking Loss	Average Precision
CLR	health	8.61	2.80	21.54	23.43	0.91
	recreation	9.67	11.17	13.45	1.43	0.45
	computer	17.17	−3.26	20.82	12.64	0.32
	yeast	11.11	11.54	4.18	13.19	0.53
	entertainment	12.05	4.24	24.54	18.75	1.09
MLkNN	health	6.98	3.26	15.30	11.05	0.91
	recreation	7.71	10.01	35.96	36.52	9.54
	computer	8.91	1.74	14.76	8.27	0.81
	yeast	6.78	7.53	5.47	8.11	1.41
	entertainment	6.55	0.72	11.92	6.52	0.26
Include-Labels	health	11.71	10.83	20.59	17.65	0.97
	recreation	10.32	−1.32	9.25	3.35	0.11
	computer	19.44	12.06	17.50	12.48	2.16
	yeast	23.53	65.00	7.58	34.83	1.49
	entertainment	16.00	7.41	17.63	13.41	1.70
RAkEL	health	16.10	4.29	9.54	6.87	0.39
	recreation	18.67	7.84	16.72	10.56	0.72
	computer	13.75	9.64	7.40	4.19	0.84
	yeast	20.54	0.00	4.93	23.81	0.13
	entertainment	7.07	5.08	6.85	1.03	0.03
LP	health	19.89	10.23	12.74	12.24	5.27
	recreation	22.03	7.78	16.17	14.64	4.73
	computer	15.73	8.40	9.68	8.10	4.56
	yeast	7.30	16.41	8.75	19.52	4.94
	entertainment	17.75	12.07	12.52	12.15	5.28
ML-Stacking	health	5.48	−1.78	7.79	1.80	0.16
	recreation	6.66	−0.80	9.18	4.63	1.18
	computer	10.19	2.34	15.53	10.94	2.90
	yeast	7.63	4.15	4.30	7.12	1.23
	entertainment	6.74	1.19	9.59	5.65	1.29

由表 4 可以得到，使用本文提出的预处理方法处理后，5

(下转第 131 页)

sys 2006). Boulder, 2006: 293-306

[17] Zheng Guo-qiang, Li Jian-dong, Zhou Zhi-li. Overview of MAC Protocols in Wireless Sensor Networks[J]. Acta Automatica Sinica, 2008, 34(3): 305-316

[18] Kulkarni S, Iyer A, Rosenberg C. An address-light, integrated MAC and routing protocol for wireless sensor networks[J]. IEEE/ACM Trans. on Networking, 2006, 14(4): 793-806

[19] Rossi M, Zorzi M. Probabilistic algorithms for cost-based integrated MAC and routing in wireless sensor networks[C]// Proc. of the 3rd Int'l Modeling and Performance Analysis of Wireless Sensor Networks (Senmetrics 2005). San Diego, 2005: 86-96

[20] Sun Li-ming, Li Jian-zhong, Chen Yu. Wireless Sensor Networks [M]. Beijing: Tsinghua University Press, 2005: 70-72

[21] Dong M, Tong L, Sadler B M. Impact of Data Retrieval Pattern on Homogeneous Signal Field Reconstruction in Dense Sensor Networks[J]. IEEE Trans. on Signal Processing, 2006, 54(11): 4352-4364

[22] Zhao Ming, Chen Zhi-gang, Zhang Lian-ming. Hybrid Spatial Correlation-based Medium Access Control for Dense Event-driven Sensor Networks[J]. Journal of System Simulation, 2007, 19(8): 1863-1868

[23] Vuran Mehmet C, Akyildiz Ian F. Spatial Correlation-Based Collaborative Medium Access Control in Wireless Sensor Networks [J]. IEEE/ACM Transactions on Networking, 2006, 14(2): 316-329

[24] Wang Hui. Principle and Application of NS2 Network Simulator [M]. Xi'an: Northwestern Polytechnical University Press, 2008

(上接第 108 页)

个评价指标均有所提高，其中提高一栏的百分比是相对于处理前的数据的。本研究对其他的 6 个数据集进行了处理，表 5 列出了其他数据集处理后的 5 个评价指标提高的百分比。

表 4 和表 5 的数据显示，这种去噪方法在 5 个指标上都有所提升。其中 One-Error 指标上有个别数据出现了较小的负值，显示去噪声后指标值略有下降，其部分原因是前面在处理时，对某几个数据集来说个别非噪声数据被当成了噪声数据来处理。

多个数据集的实验结果验证了所提出的预处理方法的有效性。本文提出的方法简单可行，只要数据集的大部分数据呈现近似的正态分布，此方法就可以用来去除噪声，提高多标签的分类效果。

结束语 本文介绍了一种正态分布的数据集去噪声的方法，并在 Mulan 平台上对多个数据集、多种分类算法都进行了验证，提供了实验分析数据，表明了该方法的有效性。此方法简单方便、不受领域限制，能较好地适用于符合近似正态分布的数据集。

参 考 文 献

[1] Zhang Min-ling, Zhou Zhi-hua. ML-KNN: A lazy learning approach to multi-label learning [J]. Pattern Recognition, 2007, 7(40): 2038-2048

[2] Tsoumakas G, Katakis I, Vlahavas I. Mining multi-label data [M]// Data Mining and Knowledge Discovery Handbook. New York: Springer US, 2010

[3] Xu Xin-shun, Jiang Yuan, Peng Liang, et al. Ensemble approach based on conditional random field for multi-labels image and video annotation[C]// Proceedings of the 19th ACM international conference on Multimedia. Scottsdale, Arizona, USA, 2011: 1377-1380

[4] Wang Jing-dong, Zhao Ying-hai, Wu Xiu-qing, et al. A transductive multi-label learning approach for video concept detection [J]. Pattern Recognition, 2011, 44(10/11): 2274-2286

[5] Sanden C, Zhang J Z. Enhancing multi-label music genre techniques [C]// Proceedings of the 34th International ACM SIGIR Conference on Research and Development in information Retrieval(SIGIR'11). New York, USA, 2011: 705-714

[6] Wieczorkowska A, Synak P, Ras Z. Multi-label classification of emotions in music[C]// Proceeding of the 2006 International Conference on Intelligent Information Proceeding and Web Mining(IIPWM). 2006: 307-315

[7] Trohidis K, Tsoumakas G, Kalliris G, et al. Multi label classification of music into emotions[C]// Proceeding of 9th International Conference on Music Information Retrieval (ISMIR). Philadelphia, PA, USA, 2008: 69-75

[8] Zhang Yi, Burer S, Street W N. Ensemble pruning via semi-definite programming [J]. Journal of Machine Learning Research, 2006(7): 1315-1338

[9] Read J, Pfahringer B, Holmes G, et al. Classifier Chains for Multi-label Classification[J]. Machine Learning, 2011, 85(3): 333-359

[10] Shen X, Boutell M, Luo J, et al. Multi-label machine learning and its application to semantic scene classification[C]// Proceedings of the 2004 International Symposium on Electronic Imaging. San Jose, California, USA, 2004: 18-22

[11] Hullermeier E, Furnkranz J, Cheng W, et al. Label ranking by learning pairwise preferences[J]. Artificial Intelligence, 2008 (16): 1897-1916

[12] Read J. A pruned problem transformation method for multi-label classification[C]// Proceeding of the New Zealand Computer Science Research Student Conference. New Zealand, 2008: 143-150

[13] Tsoumakas G, Katakis I. Multi-label classification: An overview [J]. International Journal of Data Warehousing and Mining, 2007, 3(3): 1-13

[14] Tsoumakas G, Vlahavas I. Random k-Labelsets: An ensemble method for multi-label classification[C]// Proceedings of the ECML. Warsaw, Poland, 2007: 406-417

[15] Zhang Min-ling, Zhou Zhi-hua. Multi-label neural networks with applications to functional genomics and text categorization[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(10): 1338-1351

[16] Zhang Min-lin, Zhou Zhi-hua. A k-nearest neighbor based algorithm for multi-label classification[C]// Proceedings of the IEEE International Conference on Granular Computing. Beijing, China, 2005, 2: 718-721

[17] Tsoumakas G, Dimon A, Spyromitros E, et al. Correlation based pruning of stacked binary relevance models for multi-label learning[C]// Proceedings of the ECML/PKDD. Slovenia, 2009: 101-113

[18] http://mulan.sourseforge.net/datasets.html

[19] http://meka.sourceforge.net/#download