文章编号: 1006-2475(2016) 03-0016-03

一种基于成对标签的 Rakel 算法改进

周恩波,叶荣华,张微微,周子涵

(浙江师范大学数理与信息工程学院,浙江 金华 321004)

摘要: Rakel(Random k-labelsets) 算法从原始标签集中随机选择一部分标签子集,并且使用 LP(Label Powerset) 算法训练相应的多标签子分类器。由于随机选择标签的原因,导致 LP 子分类器预测性能不好。本文基于标签的共现关系选择成对标签来训练 LP 分类器,提出 PwRakel(Pairwise Random k-labelsets) 算法。该算法通过挖掘标签相关性扩展训练集,有效提高分类性能。实验结果表明,所提出的算法与 Rakel 算法以及其他算法对比,分类准确度更高。

关键词: 多标签分类; 标签相关性; PwRakel

中图分类号: TP181 **文献标识码:** A **doi:** 10.3969/j.issn.1006-2475.2016.03.004

An Improved Rakel Approach Based on Label Pairwise

ZHOU En-bo, YE Rong-hua, ZHANG Wei-wei, ZHOU Zi-han

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

Abstract: Rakel (Random k-labelsets) randomly selects a number of label subsets from the original set of labels and uses the LP (Label Powerset) method to train the corresponding multi-label classifiers. But the models maybe have a poor performance because of randomization nature. Thus in this paper we firstly capture some pairwise relationships based on label co-occurrence between the labels to training LP classifier by PwRakel (Pairwise Random k-labelsets) algorithm. The method extends the training set by exploiting label correlations to improve classification performance effectively. The experimental results indicate that the proposed method improves multi-label classification accuracy compared with the Rakel algorithm and to other state-of-the-art algorithms.

Key words: multilabel classification; label correlation; PwRakel

0 引 言

传统的单标签分类[1]中,每个实例只有一个单标签,然而在实际应用领域中,对象通常具有多种标签。例如在音频情感分类[2]中,一个音频文本可以有多种情感类型。因此多标签分类的重要性日益凸显。尤其是在文本分类、音频分类、基因分类等领域中应用更为广泛。

文献中将多标签学习分为 2 种[3]: 1) 算法适应方法。主要思想是扩展特定的学习算法使其可以直接处理多标签数据。例如 C4.5 被用来改造处理多标签数据,以及结合贝叶斯理论和 k 近邻算法的 ML-KNN 算法等。2) 问题转换方法。主要思想是将多标签数据的学习转换为一个或多个单标签学习。代表性算法有二值相关 Binary Relevance(BR) 算法、Classifier Chain(CC) 方法[4]、Random k-labelsets(Rakel) 方法。

上述算法中 Rakel 算法通过随机选择标签构建子标签集合进行训练,从而在这个角度上考虑标签的相关关系[5-6]。Rakel 算法在预测准确度上有着很好的性能,但是集成过程中训练得到的子模型性能不够理想,导致对预测准确度有一定的影响。因此本文提出的 Pairwise Random k-labelsets(PwRakel) 算法从标签之间的共现性考虑,找到相关性高的成对标签,融合到原有训练集中,作为 Rakel 的随机子标签,从而得到更好的子模型,提升预测准确度。实验利用几个不同数据集上的实验结果来检验 PwRakel 算法的可行性,并将该算法与一些常用的多标签算法进行比较,通过计算几个评价系统性能的指标来说明该算法有一定的优势。

收稿日期: 2015-10-20

基金项目: 浙江省自然科学基金资助项目(y1100169)

作者简介: 周恩波(1990-) ,男,浙江宁波人,浙江师范大学数理与信息工程学院硕士研究生,研究方向: 机器学习,人工智能; 通信作者: 叶荣华(1971-) ,男,浙江绍兴人,教授,博士,研究方向: 语义 Web 服务,Agent 技术; 张微微(1990-) ,女,安徽安庆人,硕士研究生,研究方向: 人工智能,机器学习。

1　多标签分类相关概念

1.1　多标签问题的定义

训练集 D[7] 由 N 个实例 $E_i = (X_i, Y_i)$，$i = 1, \ldots, N$ 组成。每个实例关联一组特征向量 $X_i = (x_{i1}, x_{i2}, \ldots, x_{iM})$ 以及一组多标签 Y_i，$Y_i \in L$，$L = \{y_1, y_2, \ldots, y_q\}$。

1.2　标签相关性的研究现状

现实中多个标签之间往往不是独立的[8]，而是存在一定的相关关系。大量研究者的研究表明，在算法中引入标签的相关关系能够提升算法的预测性能。目前研究者基于相关性的研究有以下几种。第 1 种是由 Homer[9] 算法中提到的考虑标签之间的距离，通过标签与聚类中心的距离来考察标签之间相关性。第 2 种是由 Label Ranking by Learning Pairwise Preferences[10] 中提到通过考虑标签的共现来考虑标签的相关性。Eyke Hüllermeier 利用标签共现性的统计，构建 $q(q-1)/2$ 种模型，每种模型预测对应标签概率，统计所有标签的平均值，然后与阈值比较预测标签出现概率。Grigorios Tsoumakas[11] 提出通过利用标签共现性来合成新的标签进行分类。

1.3　Rakel 的研究现状

Random k-labelsets[12]（简称 Rakel），建立了一个 LP 分类器的 Ensemble（集成），用标签集合的一小部分随机标签子集的数据集作为每一个 LP 分类器的训练集训练。Rakel 通过这种方式去避免 LP 中计算复杂度高、样本倾斜的缺陷。最后由多个 LP 分类器通过投票的方式集成预测。标签的 Ranking 通过每一个基分类器的 0 或 1 预测结果来获得。通过设置阈值也可以产生二值的分类结果。

针对 Rakel 算法的随机选择特点和基分类器简单投票的组合方式，研究者分别作出了不同的改进。

Lior Rokach[13] 考虑标签相关性将所有相关标签聚合，作为标签选择，同时他也提到利用动态置信值和阈值来改善 Rakel 算法。Hung-Yi Lo[14] 主要基于 AdaBoost 算法思想通过预测误差来调整模型权重，改善 Rakel 算法。

2　基于标签相关性的 PwRakel 多标签分类算法描述

Rakel 算法随机选择标签子集，然后训练子分类器，但是由于标签存在随机关系[15]，导致子分类器预测精确度较差。本文利用标签共现性来选择成对标签，将成对标签加入标签子集，提高标签之间的相关关系，进而提高子分类器的模型预测准确度。

本文采用二阶策略，对于所有的标签集合 $L = \{y_1, y_2, \ldots, y_q\}$，任取 2 个标签考察其相关性，共有 $q(q-1)/2$ 种组合方式。根据训练集中已有的训练数据，构建共现矩阵 M 如表 1 所示。

表 1　共现矩阵 M

		y_1	y_2	y_3	…	y_q
ω_1	Y_1	0	1	0	…	1
ω_2	Y_2	1	0	1	…	1
ω_3	Y_3	1	0	1	…	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
ω_n	Y_n	1	1	0	…	1

在表 1 中，$y_i(1 \leqslant i \leqslant q)$ 表示训练集中的第 i 个标签，$\omega_j(1 \leqslant j \leqslant n)$ 表示矩阵 M 第 j 行的所有数据，通过公式:

$$C_{\omega u, \omega v} = \sum M_{ui} \times M_{vi} \quad (1)$$

可得到 ω_u 和 ω_v 共同出现的频数 $C_{\omega u, \omega v}$，即标签 y_u 和 y_v 的共现频数。这样处理之后得到的结果 S 构成了一个对称矩阵。在改进的 PwRakel 算法中，对于每一个标签，都要考虑它与其他标签的共现频数。通过上述方法，不仅可以从训练数据集中获取更多信息，还能衡量出任意 2 个标签间的相关性大小。然后基于频数排序，得到强相关的标签集 A。

在 Rakel 选择子标签时，从标签集 A 中随机选择一组成对标签，在 L 中选择不重复的一个随机标签组成一个标签子集，然后训练子分类器。算法的训练过程如下:

算法 1　Pairwise Random k-labelsets

输入: 实例 x，聚合 LP 分类器 h_i，k 的标签集 Y_i，标签集 L

输出: 多标签分类器结果

$R \leftarrow L^k$

for $i \leftarrow 1$ to $\min(m, |L^k|)$ do

if $(i < 4 \times \frac{p}{2})$

$Y_i \leftarrow$ a 2-labelset randomly select from A + a(k − 2)-labelset randomly select from R;

else

$Y_i \leftarrow$ a k-labelset randomly selectedf rom R;

train an LP classifier $h_i: X \rightarrow P(Y_i)$ on D;

$R \leftarrow R \setminus \{Y_i\}$;

end

3　实验结果与分析

3.1　评价指标

根据文献 [16]，实验采用 Subset Accuracy、Re-

call、F-measure 这 3 个性能评价指标。

1) 分类准确度定义如下:

$$\text{SubsetAccuracy} = \frac{1}{N}\sum_{i=1}^{N} I(Z_i = Y_i) \quad (2)$$

其中 I(true) =1 和 I(false) =0。这是一种非常严格的评价方法,因为它需要预测的标签集合与真实的标签集合完全吻合。

2) 召回率定义如下:

$$\text{Recall} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap Z_i|}{|Y_i|} \quad (3)$$

3) F-measure[17] 定义如下:

$$F_1 = \frac{1}{N}\sum_{i=1}^{N}\frac{2|Y_i \cap Z_i|}{|Z_i| + |Y_i|} \quad (4)$$

F_1 为样本的基础指标,其值是数据集样本的平均值,最好的评分为 1,最差的评分为 0。

3.2 实验数据集

表 2 为样本数据集及其内部数据相关统计信息。

表 2 样本数据集及其内部数据相关统计信息

Dataset	Domain	Instances	Labels	Distinct
Emotions	Music	593	6	27
Scene	Image	2407	6	15
Birds	Audio	645	19	133
Enron	Text	1702	53	753

将本文算法与当前较流行的 4 种算法[18]:BR、Rakel、CC 以及基于 KNN 的算法 MLKNN 进行实验对比。其中 CC 的基分类器算法使用 Weka 中支持向量机分类算法,Rakel 的基分类器算法是 LP,LP 的基分类器算法同样采用支持向量机分类算法,标签子集大小 k =3,模型个数 m =2|L|(标签数量的 2 倍),阈值 t =0.5。实验中对每个测试数据集运用了 5-fold 交叉验证方法,ML-KNN 算法中 k =10,smoothing =1。表 3 ~表 5 给出了各算法的具体实验结果。

表 3 各算法在 4 个数据集上的 Subset Accuracy

	MLKNN	BR	CC	PwRakel	Rakel
Emotions	0.2899	0.2782	0.2882	**0.3355**	0.3237
Scene	0.6253	0.5314	**0.6552**	0.6323	0.6211
Birds	0.4625	0.4813	0.4720	**0.4875**	0.4783
Enron	0.0588	0.1140	0.1252	**0.1281**	0.1234

表 4 各算法在 4 个数据集上的 Recall

	MLKNN	BR	CC	PwRakel	Rakel
Emotions	0.6087	0.5929	0.6321	**0.6765**	0.6700
Scene	0.6883	0.6547	0.7052	**0.7321**	0.7196
Birds	0.4940	0.6201	0.6115	**0.6244**	0.6200
Enron	0.364	0.5349	0.5321	**0.5526**	0.5418

表 5 各算法在 4 个数据集上的 F-measure

	MLKNN	BR	CC	PwRakel	Rakel
Emotions	0.6131	0.5922	0.6087	**0.6565**	0.6541
Scene	0.6847	0.6252	0.7072	**0.7146**	0.7013
Birds	0.5051	0.607	0.5978	**0.6158**	0.6083
Enron	0.4097	0.514	0.5168	**0.5341**	0.5276

本小节主要介绍的是 PwRakel 算法与其他 4 种同类型算法的实验比较。表 3 ~表 5 分别给出了 MLKNN 算法、BR 算法、CC 算法、Rakel 算法和 PwRakel 算法在 Emotions、Scene、Birds 和 Enron 这 4 个数据集上 3 种评价指标(即分类准确度、F-measure 以及 Recall)的值。其中每行中用黑色加粗的数据为 5 种算法中在该数据集上表现最好的那种算法对应的数据。

从表 3 ~表 5 中可以看出,PwRakel 算法在 4 个数据集的 12 个评价指标上有 11 个评价指标都是最好的,这足以说明 PwRakel 算法相对于其他同类型算法的优越性。在 3 种评价指标中,PwRakel 算法在 Recall 和 F-measure 的表现最好,在准确度上 3 个最佳,1 个仅次于最佳,说明了 PwRakel 算法在优化算法分类准确度的有效性。同时 PwRakel 算法的性能均优于 Rakel 算法,表明引入标签的相关性能有效克服 Rakel 算法选择标签不准确的缺点。

上述实验结果表明 PwRakel 算法确实能通过标签共现频数找到训练集中潜在的重要标签,并将这些重要标签结合原来的训练集形成更加完备的新的训练集,从而建立更加优化的分类器模型,提高分类的预测准确度。

4 结束语

在多标签分类中,标签之间的相关性是一个不可忽略的重要因素。为了充分利用标签之间的相关性来改善多标签分类的性能,本文提出了一种联系标签相关性的 PwRakel 算法,将 PwRakel 算法与已有的几种多标签学习算法在 4 个多标记数据集上进行比较,实验结果表明 PwRakel 算法性能较好。该算法能够在多个评价指标上都取得较好的结果,尤其在 Recall 和 F-measure 评价指标上相较 Rakel 算法具有明显的优势。但是,本算法在研究标签之间的相关性时,采用的只是简单的共现频数考虑标签的相关性,如何使用更合适的标签相关性选择方法来发现隐藏的标签关系是将来值得研究的问题。

参考文献:

[1] 陆广泉,谢扬才,刘星,等. 一种基于 KNN 的半监督分类改进算法[J]. 广西师范大学学报(自然科学版),2012,30(1):45-49.

(下转第 23 页)

参考文献:

[1] Flake G W, Tarjan R E, Tsioutsiouliklis K. Graph clustering and minimum cut trees [J]. Internet Mathematics, 2003, 1(4): 385-408.

[2] Beineke L W, Wilson R J. Topics in Algebraic Graph Theory [M]. Cambridge University Press, 2004: 276.

[3] Yang Bo, Cheung W K, Liu Jiming. Community mining from signed social networks [J]. IEEE Transactions on Knowledge & Data Engineering, 2007, 19(10): 1333-1348.

[4] 沈华伟,程学旗,陈海强,等. 基于信息瓶颈的社区发现 [J]. 计算机学报, 2008, 31(4): 677-686.

[5] 黄发良,张师超,朱晓峰. 基于多目标优化的网络社区发现方法 [J]. 软件学报, 2013, 24(9): 2062-2077.

[6] 段炼,朱欣焰. 基于社区时空主题模型的微博社区发现方法 [J]. 电子科技大学学报, 2014, 43(3): 464-469.

[7] McDaid A F, Murphy T B, Friel N, et al. Model-based clustering in networks with stochastic community finding [C] // COMPSTAT 2012 Proceedings. 2012.

[8] Chen Dongming, Dong Yanlin, Huang Xinyu, et al. A community finding method for weighted dynamic online social network based on user behavior [J]. International Journal of Distributed Sensor Networks, 2015, 2015: Article No. 97.

[9] Girvan M, Newman M E J. Community structure in social and biological networks [J]. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821-7826.

[10] Newman M E J. Fast algorithm for detecting community structure in networks [J]. Physical Review E, 2004, 69(6): 279-307.

[11] Xu Xiaowei, Yuruk N, Feng Zhidan, et al. SCAN: A structural clustering algorithm for networks [C] // Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2007: 824-833.

[12] Zhao Weizhong, Martha V, Xu Xiaowei. PSCAN: A parallel structural clustering algorithm for big networks in MapReduce [C] // Proceedings of the 27th IEEE International Conference on Advanced Information Networking and Applications (AINA). 2013: 862-869.

[13] Zhang Huijuan, Sun Shixuan. A graph clustering algorithm based on shared neighbors and connectivity [C] // Proceedings of the 8th IEEE International Conference on Computer Science & Education (ICCSE). 2013: 761-764.

[14] 吴烨,钟志农,熊伟,等. 一种高效的属性图聚类方法 [J]. 计算机学报, 2013, 36(8): 1704-1713.

[15] 冷泳林,鲁富宇. 基于 MapReduce 的 SimRank 算法在图聚类中的应用 [J]. 电子设计工程, 2015, 23(6): 9-11.

[16] Tabrizi S A, Shakery A, Asadpour M, et al. Personalized-pageRank clustering: A graph clustering algorithm based on random walks [J]. Physica A: Statistical Mechanics & Its Applications, 2013, 392(22): 5772-5785.

[17] Newman M E J. Communities, modules and large-scale structure in networks [J]. Nature Physics, 2012, 8(1): 25-31.

(上接第 18 页)

[2] 李志欣,卓亚琦,张灿龙,等. 多标记学习研究综述 [J]. 计算机应用研究, 2014, 31(6): 1601-1605.

[3] 王霄,周李威,陈耿,等. 一种基于标签相关性的多标签分类算法 [J]. 计算机应用研究, 2014, 31(9): 2609-2612.

[4] Hariharan B, Zelnik-Manor L, Vishwanathan S V N, et al. Large scale max-margin multi-label classification with priors [C] // Proceedings of the 27th International Conference on Machine Learning. 2010: 423-430.

[5] Dembczynski K, Cheng Weiwei, Hüllermeier E. Bayes optimal multilabel classification via probabilistic classifier chains [C] // Proceedings of the 27th International Conference on Machine Learning. 2010: 279-286.

[6] Li Nan, Zhou Zhi-Hua. Selective ensemble of classifier chains [M] // Multiple Classifier Systems. 2013: 146-156.

[7] Dembczyński K, Waegeman W, Hüllermeier E. An analysis of chaining in multi-label classification [C] // Proceedings of the 20th European Conference on Artificial Intelligence (ECAI 2012). 2012: 294-299.

[8] Yu Ying, Pedrycz W, Miao Duoqian. Multi-label classification by exploiting label correlations [J]. Expert Systems with Applications, 2014, 41(6): 2989-3004.

[9] Tsoumakas G, Katakis I, Vlahavas I. Effective and efficient multilabel classification in domains with large number of labels [C] // Proceedings of the 2008 ECML/PKDD Workshop on Mining Multidimensional Data (MMD'08). 2008: 30-44.

[10] Hüllermeier E, Fürnkranz J, Cheng Weiwei, et al. Label ranking by learning pairwise preferences [J]. Artificial Intelligence, 2008, 172(16-17): 1897-1916.

[11] Spolaôr N, Monard M C, Tsoumakas G, et al. Label construction for multi-label feature selection [C] // Proceedings of the 2014 Brazilian Conference on Intelligent Systems. 2014: 247-252.

[12] Tsoumakas G, Vlahavas I. Random k-labelsets: An ensemble method for multilabel classification [C] // Proceedings of the 18th European Conference on Machine Learning. 2007: 406-417.

[13] Rokach L, Schclar A, Itach E. Ensemble methods for multi-label classification [J]. Expert Systems with Applications, 2014, 41(16): 7507-7523.

[14] Lo Hung-Yi, Lin Shou-De, Wang Hsin-Min. Generalized k-labelset ensemble for multi-label classification [C] // Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing. 2012: 2061-2064.

[15] 乔健,田庆. 利用最近邻信息快速分类多标签数据 [J]. 计算机工程与应用, 2011, 47(32): 138-140.

[16] 李思男,李宁,李战怀. 多标签数据挖掘技术: 研究综述 [J]. 计算机科学, 2013, 40(4): 14-21.

[17] 李哲,王志海,何颖婧,等. 一种启发式多标记分类器选择与排序策略 [J]. 中文信息学报, 2013, 27(4): 119-126.

[18] 申超波,王志海,孙艳歌. 基于标签聚类的多标签分类算法 [J]. 软件, 2014, 35(8): 16-21.