《电子技术应用》 www.ChinaAET.com

针对弱标记的多标记数据集成学习分类方法

李凤英,李 宏,李 培
(中南大学 信息科学与工程学院,湖南 长沙 410083)

摘 要：提出一种针对弱标记的多标记数据集成学习分类方法，它通过采用基于相似性成对约束投影的方法来处理数据,更好地利用了弱标记样本的特征,从而提高了分类性能。

关键词：分类;多标记数据;集成学习;弱标记数据
中图分类号：TP399 **文献标识码**：A **文章编号**：1674-7720(2012)13-0073-03

Multi-label classification method for weak labeling

Li Fengying, Li Hong, Li Pei
(School of Information Science and Engineering, Central South University, Changsha 410083, China)

Abstract: This paper presents a multi-label classification method for weak labeling, which deals with data by using the way of pairs constraints projection based on similarity. The method can make better use of characteristics of the sample with weak label and improve the classification performance.

Key words: classification; multi-label data; ensemble learning; weak labeling

数据挖掘技术随着现代技术的飞速发展变得越来越重要了。分类是数据挖掘中的一个重要研究领域,目前分类算法有很多,经典的有决策树、贝叶斯模型、支持向量机等。在很多现实生活的分类问题中,一个样本往往同时属于多个不同的类别,比如:一幅画同时拥有“素描”、“人物”、“运动”等多个标记。多标记学习就是一种针对多标记样本进行学习的重要技术。对多标记数据进行正确的分类已成为近年来机器学习和数据挖掘中的热点研究方向。

以往多标记学习的研究是在训练样本标记完整的情况下进行的。但是,在现实生活应用中,多数样本的标记不是完整的,而且为每个样本提供完整的标记非常困难。在此,一个弱标记样本包含其对应所有标记中的部分标记。现有的多数多标记学习方法,由于不能对这种弱标记样本进行有效地学习，可能会给训练集引入大量的噪声。为了有效地利用这些弱标记样本进行学习,本文提出一种针对弱标记的多标记数据集成学习分类方法。

1 研究现状

目前,对多标记数据分类做了很多研究。最典型的多标记算法是 ML-KNN 算法。该算法是对已有 K 近邻算法的改进。传统的 K 近邻算法是基于向量的空间距离来选取近邻,但有的分类处理中要用到向量的夹角,所以广凯和潘金贵提出一种基于向量夹角的 K 近邻多标记分类算法。Sapozhnikova 等人提出了使用 ART(Adaptive Resonance Theory)神经网络的方法解决多标记分类问题。段震等人提出了基于覆盖的多标记学习方法等。但是,目前针对弱标记数据的多标记分类方法比较少。孔祥南等人提出了一种针对弱标记的直推式多标记分类方法。直推式学习是利用未标记数据学习的主流技术之一。

集成学习是近年来机器学习领域中研究热点之一。经典的两个集成算法是 Bagging 和 Boosting。张燕平等人提出了一种新的决策树选择性集成学习方法,杨长盛等人提出了基于成对差异性度量的选择性集成方法等。目前的集成学习研究集中于传统的单标记学习,此前 Zhang 等人已在单标记分类中引入成对约束建立基分类器,李平在多标记分类中引入了软成对约束建立基分类器。受此启发,本文在针对弱标记数据分类中引入了基于相似性成对约束投影的多标记集成学习方法。

2 多标记集成学习算法

2.1 算法的引入

集成学习方法可以提高总体的分类准确率,但针对

《电子技术应用》 www.ChinaAET.com

弱标记的多标记集成学习算法几乎没有。本文首次将集成学习引入到针对弱标记的多标记学习中。此前，李平首次将集成学习引入到多标记分类中。软成对约束指的是：若两个样本的标记相同数大于等于预先设定的阈值，则将样本放到 M 集合中，否则放到 C 中[1]。但是，当样本的标记不是完整的时候，这个方法容易导致本该放到 M 集合中的样本对却放到了 C 中。因此，本文针对这个问题提出了基于相似性成对约束投影的多标记集成学习方法 RPCME。

2.2 基于相似性成对约束投影

本文研究的重点是针对弱标记样本[2]如何在多标记集成学习中合理有效地利用弱标记数据提供的成对约束信息并建立强健的集成分类器。本文的基于相似性成对约束定义为：若给定的两个数据样本的相似度大于等于预先设定的阈值，则将样本放到 M 集合中，否则放到 C 中。相似度通过式(1)计算：

$$S_{ij}=\exp\left(-\frac{\|x_i-x_j\|^2}{2\sigma^2}\right) \tag{1}$$

式中 σ 为宽度参数，本文中固定为样本间的平均距离。该方法有效地避免了弱标记引起的误差，并有效地利用了样本间的相似度。

分别计算集合 C 和 M 的散度矩阵，这两个矩阵是用成对约束信息生成的。该算法通过散度矩阵计算投影矩阵，然后通过投影矩阵将原数据映射到新的数据空间[3]。

2.3 权重更新策略

由于本文的基分类器是稳定的 MLKNN 算法，所以采用的方法是：各训练样本的初始权重均设置为 1，而当迭代训练个体分类器时[4]，上一轮中被误分的样本将增加权重，如 $(1+r)$，r 为权重因子。这种方法较为简单，且能保障个体分类器的差异性。差异性是集成学习中的重要概念，基分类器差异性的大小直接影响分类器的性能。因此，为了提高分类器的差异性[5]，在每次的训练过程中，权重因子都要更新为不同的值。

2.4 多标记数据基分类器的集成

对于多个不同的基分类器组成的多标记集成分类器，通常用以下两种方法对基分类器进行集成：多数投票和加权投票。本文采用的方法是选择性多数投票方法。即在集成基分类器时，为了提高分类精度，要丢弃一些准确率比较低的分类器。本文设置准确率的阈值为 0.7，即基分类器的准确率大于 0.7 时参加集成，否则不参加集成，然后采用多数投票的方法。

2.5 RPCME 算法描述

RPCME 算法首先采用基于相似性成对约束投影建立基分类器，然后对训练样本进行分类，对错误分类的数据样本增加权重，最后对多标记集成分类器进行组合。

3 实验

3.1 数据集

本文选取了参考文献[1]中所使用的多媒体多标记数据集 scene。为了模拟弱标记样本，本文在训练集中模拟生成弱标记，并在测试集评价算法性能时使用测试样本的完整真实标记来进行性能评估。

3.2 评价指标

假设 $y_i\subseteq L$（L 为标记总集）是对应样本 d_i 的预测标记，真实标记集为 s_i。那么对于测试数据集合 D'，本文选择以下 3 个指标对多值多标记的分类结果进行评价。此外，为了对正确率 p_i 和召回率 r_i 有一个综合的度量，文中选择了 F1 度量：

$$\text{汉明距离：}Hloss(D')=1-\frac{1}{|D'|}\sum_{i=1}^{|D'|}\frac{|s_i\oplus y_i|}{|L|} \tag{2}$$

$$\text{正确率：}Acc(D')=\frac{1}{|D'|}\sum_{i=1}^{|D'|}\frac{|s_i\cap y_i|}{|s_i\cup y_i|} \tag{3}$$

$$\text{F1 度量：}F_1(D')=\frac{1}{|D'|}\sum_{i=1}^{|D'|}\frac{|2\times p_i\times r_i|}{|p_i+r_i|} \tag{4}$$

3.3 实验设计

实验中选择了 SPACME 和 MLKNN 算法作比较，所有实验都是在 Matlab R2010a 下实现的。ML-KNN 中的近邻数目设置为 10。设计以下三组实验：

第一组：设置基分类器大小 $L=5$，阈值为 0.5，对比三种算法的性能。第二组：在不同的基分类器大小下，对比三种算法的性能。第三组：在不同的阈值下，$L=5$，对比三种算法的性能（注：这里的阈值为相似性成对约束时的阈值）。

4 实验结果及分析

本节给出上述三组实验的实验结果及其相关分析。其中汉明距离越小越好，用“↓”表示，正确率和 F1 度越大越好，用“↑”表示。表 1 给出了第一组的实验结果。

表 1 三种算法的性能比较（$L=5$，$\theta=0.5$）

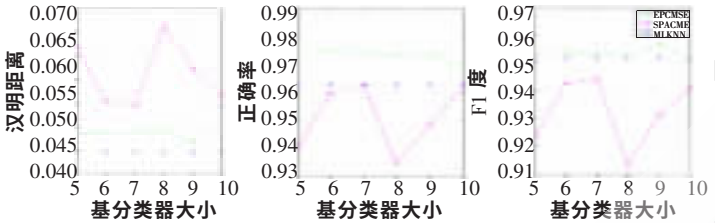
算法	Hloss ↓	Acc ↑	F1 ↑
EPCMSE	0.040 3	0.979 8	0.959 7
SPACME	0.050 8	0.961 1	0.942 6
ML-KNN	0.039 7	0.963 0	0.951 7

从表 1 可以看出，EPCMSE 算法在 3 个性能指标下都优于 SPACME 算法，在正确率和 F1 下优于 MLKNN 算法，只在汉明距离这个指标下的性能略低于 MLKNN。总体来看 EPCMSE 算法的性能优于其他两个算法。

从图 1 得知 EPCMSE 算法较 SPACME 受基分类器大小的影响小，图 1(a)、(b)、(c)分别是在汉明距离、正确率和 F1 度性能指标下三种算法的性能曲线。从中可知 EPCMSE 总体性能比其他两种算法的性能好。SPACME 在大小不同的基分类器下，性能变化较大。当

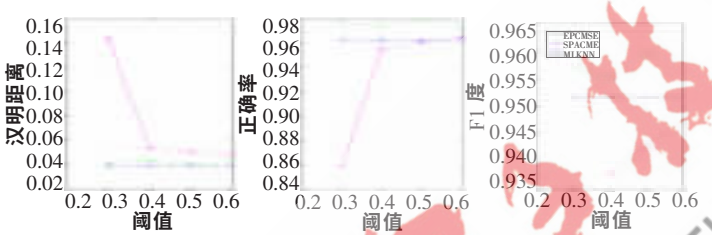
L=9 时，EPCMSE 性能达到了最好，但运行时间较长。当 L=5 时，EPCMSE 性能也比较好，且时间较短。

从图 2 可知 EPCMSE 算法较 SPACME 受阈值的影响小。图 2(a)、(b)、(c)分别表示在汉明距离、正确率和 F1 度性能指标下三种算法的性能曲线。从中可知 EPCMSE 算法的总体性能比 SPACME 和 ML-KNN 算法的性能好。当阈值等于 0.5 的时候，EPCMSE 算法达到了最好。



(a)不同基分类器下算法的汉明距离比较　(b)不同基分类器下算法的正确率比较　(c)不同基分类器下算法的 F1 度比较

图 1　不同的基分类器大小下，三种算法的性能比较曲线



(a)不同阈值下算法的汉明距离比较　(b)不同阈值下算法的正确率比较　(c)不同阈值下算法的 F1 度比较

图 2　不同的阈值下，三种算法的性能比较曲线

本文针对多标记学习任务中仅能获得弱标记数据的情况，提出了一种针对弱标记的多标记集成学习方法 EPCMSE。从实验结果中可知，通过相似性成对约束投影建立基分类器，在场景图像分类任务中，该方法在弱标记情况下，具有良好的健壮性，获得较好的分类性能。在少量的弱标记数据的情况下，如何进一步提高分类性能，将需要更多的研究。

参考文献

[1] 李平. 多标记分类中的半监督降维和集成学习 [D].长沙：中南大学，2010.

[2] 孔祥南，黎铭，姜远，等.一种针对弱标记的直推式多标记分类方法[J].计算机研究与发展，2010，47(8)：1392-1399.

[3] ZHANG D Q, CHEN S C, ZHOU Z H, et al. Constraint projections for ensemble learning[C]. In: Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI'08), Chicago, 2008.

[4] VANESSA G V, JERONIMO A G, AN1IBAL F V. Committees of Adaboost ensembles with modified emphasis functions[J]. Neurocomputing, 2010, 73: 1289-1292.

[5] 张宏达，王晓丹，等.分类器集成差异性研究[J].系统工程与电子技术，2009，31(12)：3007-3012.

(收稿日期：2012-03-07)

作者简介：

李凤英，女，1986 年生，硕士，主要研究方向：数据挖掘技术。

李宏，男，1966 年生，教授，主要研究方向：数据挖掘、图像识别技术。

李培，男，1985 年生，硕士，主要研究方向：地震勘探及数据处理方向。