一种半监督的多标签 Boosting 分类算法*

赵晨阳[a]，侣　洁[b]
（西北大学 a. 数学系；b. 信息科学与技术学院，西安 710069）

摘　要：针对标记数据不足的多标签分类问题，提出一种新的半监督 Boosting 算法，即基于函数梯度下降方法给出一种半监督 Boosting 多标签分类的框架，并将非标记数据的条件熵作为一个正则化项引入分类模型。实验结果表明，对于多标签分类问题，新的半监督 Boosting 算法的分类效果随着非标记数据数量的增加而显著提高，在各方面都优于传统的监督 Boosting 算法。

关键词：Boosting 算法；半监督学习；多标签分类
中图分类号：TP301.6　　**文献标志码**：A　　**文章编号**：1001-3695(2012)09-3266-03
doi:10.3969/j.issn.1001-3695.2012.09.017

Semi-supervised multi-label Boosting algorithm

ZHAO Chen-yang[a], SI Jie[b]
(*a. Dept. of Mathematics, b. School of Information & Technology, Northwest University, Xi'an* 710069, *China*)

Abstract: For multi-label classification problem without enough labeled data, this paper proposed a new semi-supervised Boosting algorithm. It provided a semi-supervised general multi-label Boosting framework by using functional gradient descent method. It also used the conditional entropy as a regularization term on unlabeled data in classification model. Experimental result shows that the performance of the new semi-supervised Boosting algorithm can be improved by increasing unlabeled data; it also has a better result than traditional supervised Boosting algorithm by different measures.

Key words: Boosting algorithm; semi-supervised learning; multi-label classification

0　引言

半监督学习[1]是一种机器学习方法，它同时利用标记数据和非标记数据的信息，大多数般情况下是少量的标记数据和大量的非标记数据来建立有效的模型，从而改善监督学习中标记数据不足的问题。Zhu 等人[2]讨论了半监督学习中的一些常用方法，包括产生式模型、自学习模型、协同训练、信息理论正则化模型和基于图的模型等。然而这些分类算法只关注单标签、二类或多类的分类问题。

多标签分类区别于单标签分类在于同一个实例可以被归属于多个类别。近年来，多标签分类问题引起了人们越来越多的学习研究和应用需求。Tsoumakas 等人[3]总结了关于多标签分类的一些常见的应用，如文本分类、音乐分类、图和声音信号的分类以及生物医学应用等。

AdaBoost.MH 算法[4,5]是一种 AdaBoost 算法[6]在多标签分类问题中的扩展。通过将一个多标签分类问题转换为多个二类分类问题，并且在一个单独迭代过程中同时处理这些二类问题，而不是将其分别处理，有效地解决了多标签分类问题。与其他的监督学习方法一样，AdaBoost.MH 算法在分类过程中没有利用非标记数据的信息，因此分类效果会受到标记数据不足问题的限制。

本文通过改进 AdaBoost.MH 算法，利用条件熵作为一种非标记数据的信息度量，提出了一种半监督的 Boosting 算法来解决多标签的分类问题，称为 SemiBoost.MH 算法。将 SemiBoost.MH 算法看做是一个优化问题，用函数梯度下降方法[7]去优化包含标记数据以及非标记数据信息的目标函数，促使算法去寻找假定的标签，从而减少标签变量的非确定程度。

1　AdaBoost.MH 算法

用 X 表示数据空间，用 Y 表示所有标签的有限集合，$|Y| = K$。在多标签的分类问题中，每一个数据 $x \in X$ 属于一个标签集合 $y \subset Y$，因此，每一个标记数据以及其标签记做 (x,y)。当 $|y| = 1$ 时，多标签的分类问题则简化为单标签的分类问题。

对于每一个标签 $l \in Y$，定义 $y[l] = 1$，如果 $l \in y$，否则 $y[l] = -1$，那么，每一个标记数据则对应于一个 K 维的标签集合 y。多标签分类的目标是预测所有正确的标签，而不是单标签分类中仅仅一个正确的标签。

Schapire 等人[4,5]在 1999 年提出了处理多标签问题的 AdaBoost.MH 算法，用 Hamming 损失作为衡量多标签分类的一个准则，定义如下：

$$L_H(F) = \frac{1}{mK}\sum_{i=1}^{m}\sum_{l=1}^{K} I(\operatorname{sign}(F(x_i,l)) \neq y_i[l]) \tag{1}$$

其中：F 是一个从 X 到 R^K 的映射，m 为标记数据的数量。

算法的目标是在标记数据 $(x_i,y_i)(i=1,\cdots,m)$ 上训练一个映射 $F_T = \alpha_1 h_1(x,l) + \cdots + \alpha_T h_T(x,l)$，使得 Hamming 损失达到最小。其中：$h_t(x,l)(t=1,\cdots,T)$ 表示一个弱假设试图预

收稿日期：2012-02-13；**修回日期**：2012-03-19　　**基金项目**：国家自然科学基金资助项目(10771169)

作者简介：赵晨阳(1984-)，女，博士研究生，主要研究方向为数据挖掘(starstaryang@163.com)；侣洁 (1990-)，女，本科，主要研究方向为人工智能.

测数据 x 的标签为 l,α_t 表示 $h_t(x,l)$ 的权重。由于 Hamming 损失是不连续且不可导的,因此 AdaBoost. MH 的优化过程通过最小化指数损失函数:

$$\sum_{i=1}^{m}\sum_{l=1}^{K}\exp(-y_i[l]F_t(x_i,l)) \tag{2}$$

作为 Hamming 损失的上界来实现的。

具体算法流程如下:

算法 1

输入:标记数据$(x_i,y_i)(i=1,\cdots,m)$。

输出:$F_T(x,l)=\sum_{t=1}^{T}\alpha_t h_t(x,l)$。

初始化:数据分布 $D_i(l)=1/(mK)$,$i=1,\cdots,m$,$l\in y$,$t=1,\cdots,T$。

a)选择 h_i 以及 α_t 使得

$$\sum_{i=1}^{m}\sum_{l=1}^{K}D_i(l)\exp(-y_i[l]\alpha_t h_t(x_i,l))$$

达到最小。

b)更新 $D_i(l)=D_i(l)\exp(-y_i[l]\alpha_t h_t(x_i,l))$。

2　SemiBoost.MH 算法

监督学习算法往往需要大量的标记数据,但在现实世界中标记数据通常不容易得到。当标记数据的数量很小时,AdaBoost. MH 算法的分类效果就会受到限制。本文提出的 SemiBoost. MH 算法通过最小化非标记数据的条件熵[8],避免了 AdaBoost. MH 算法在标记数据不足时的局限性,从而很好地解决了多标签的分类问题。

2.1　算法原理

由函数梯度下降的角度来看,SemiBoost. MH 算法可以看做是用梯度下降的方法来优化一个基于标记数据和非标记数据的关于 F 损失函数 $R(F)$。可以定义替代的损失函数如下:

$$R(F_t)=\sum_{i=1}^{m}\sum_{l=1}^{K}L_L(F_t(x_i,l))+\gamma\sum_{i=m+1}^{n}\sum_{l=1}^{K}L_U(F_t(x_i,l)) \tag{3}$$

其中:$x_i(i=1,\cdots,m)$表示标记数据,$x_i(i=m+1,\cdots,n)$表示非标记数据;L_L 表示标记数据的替代损失函数,通常采用一个光滑的凸函数;L_U 表示非标记数据的替代损失函数;参数 γ 用来控制非标记数据的影响权重。

2.1.1　弱假设 h_t 的选择

为了使损失函数式(3)达到最小,在每一次迭代中应更新组合的分类器 F 为 $F_t=F_{t-1}-\alpha\nabla R(F_{t-1})$。其中:$\nabla R(F)$表示 R 在 F 的梯度,α 表示步长。由于 F 是一个弱假设的组合,在第 t 次迭代中 F 的更新只能局限为 $F_t(x,l)=F_{t-1}(x,l)+\alpha_t h_t(x,l)$。因此,问题转换为寻找一个最接近负梯度方向的弱假设,需要寻找一个弱假设 h_t,使得它和 $-\nabla R(F)$ 的内积 $-\nabla R(F)\cdot h_t$ 达到最大。那么问题等价于选择 h_t 使

$$\nabla R(F)\cdot h_t=\sum_{i=1}^{m}\frac{\partial L_L(F_t)}{\partial F_t(x_i,l)}h_t(x_i,l)+\gamma\sum_{i=m+1}^{n}\frac{\partial L_U(F_t)}{\partial F_t(x_i,l)}h_t(x_i,l) \tag{4}$$

达到最小。

2.1.2　步长 α_t 的选择

在选择了最优的$\hat{h}_t$ 后,接下来的问题就是如何选择步长 α_t,以使得

$$R(\alpha_t,\hat{h}_t)=\sum_{i=1}^{m}\sum_{l=1}^{K}L_L(F_t(x_i,l)+\alpha_t\hat{h}_t(x_i,l))+\gamma\sum_{i=m+1}^{n}\sum_{l=1}^{K}L_U(F_t(x_i,l)+\alpha_t\hat{h}_t(x_i,l)) \tag{5}$$

达到最小。可以采用多种优化方法来完成,如使用 L-BFGS[9] 算法。

2.2　SemiBoost.MH 算法框架

基于以上算法介绍,算法框架归纳如下:

算法 2　目标:最小化 $R(F)$。

输入:标记数据$(x_i,y_i)(i=1,\cdots,m)$以及非标记数据 $x_i(i=m+1,\cdots,n)$。

输出:$F_T(x,l)=\sum_{t=1}^{T}\alpha_t h_t(x,l)$。

初始化:$F_0\equiv 0$。

对于 $t=1,\cdots,T$:

a)对于每一个标签 $l\in Y$,选择 $h_t(x,l)$使得$\nabla R(F)\cdot h_t$ 达到最小。

b)选择步长 α_t,以使得式(5)达到最小。

c)更新 $F_t(x,l)=F_{t-1}(x,l)+\alpha_t h_t(x,l)$。

2.3　SemiBoost.MH 算法中的替代损失函数

在 SemiBoost. MH 算法框架中,标记数据以及非标记数据的损失函数采用替代的损失函数来表示。本节将讨论算法中关于标记数据以及非标记数据具体的损失函数。

2.3.1　标记数据的对数损失函数

由于 Hamming 损失是非凸的,且不是连续的,因此很难对其直接优化。本文采用对数损失函数作为标记数据的替代损失,其形式如下:

$$L_L(F(x_i,l))=\log(1+\exp(-y_i[l]F(x_i,l))) \tag{6}$$

最小化对数损失函数意味着最大化标记数据的条件似然函数 $\sum_{i=1}^{m}\sum_{l=1}^{k}\log p(y_i[l]\mid x_i)$,这里 $\log p(y_i[l]\mid x_i)=\frac{\exp(-y_i[l]F(x_i,l))}{\sum_y\exp(-yF(x_i,l))}$。因此,对于标记数据以及每一标签选择弱假设 $h_t(x,l)$的目的在于最小化

$$\sum_{i=1}^{m}\frac{\partial L_L(F_t)}{\partial F_t(x_i,l)}h_t(x_i,l)=\sum_{i=1}^{m}\frac{-\exp(-y_i[l]F_t(x_i,l))}{1+\exp(-y_i[l]F_t(x_i,l))}y_i[l]h_t(x_i,l)-\sum_{i=1}^{m}D_i(l)y_i[l]h_t(x_i,l) \tag{7}$$

其中:$D_i(l)=\frac{\exp(-y_i[l]F_t(x_i,l))}{1+\exp(-y_i[l]F_t(x_i,l))}$可以看做是每一标记数据对于每一标签的权重。因此,最小化式(7)相当于最小化加权的 Hamming 损失。

2.3.2　非标记数据的替代损失函数

最小化条件熵 $H(p(y|x))$意味着增大条件分布 $p(y|x)$的确定程度,从而更容易去估计非标记数据的标签。用条件熵 $H(p(y|x_i))$作为非标记数据的替代损失函数,其形式如下:

$$L_U(F_t(x_i,l))=H(p(y|x_i))=-\sum_y p(y|x_i)\log p(y|x_i)=\sum_y\frac{\log(1+\exp(-yF_t(x_i,l)))}{1+\exp(-yF_t(x_i,l))} \tag{8}$$

如果定义非标记数据的权重为 $D_i(l,y)=\frac{\exp(-yF_{t-1}(x_i,l))}{1+\exp(-yF_{t-1}(x_i,l))}$,那么式(8)可以表示为

$$\nabla R(F)\cdot h_t=\sum_{i=1}^{m}\frac{\partial L_L(F_t)}{\partial F_t(x_i,l)}h_t(x_i,l)+\gamma\sum_{i=m+1}^{n}\frac{\partial L_U(F_t)}{\partial F_t(x_i,l)}h_t(x_i,l)=$$

$$-\sum_{i=1}^{m} D_t(l)y_i[l]h_t(x_i,l) +$$
$$\gamma \sum_{i=m+1}^{n}\sum_{y}\frac{\log(1+\exp(-yF_{t-1}(x_i,l)))-1}{1+\exp(-yF_{t-1}(x_i,l))}D_t(l,y)yh_t(x_i,l) \quad (9)$$

需要注意的是,这里的目标函数式(9)是非凸的,具体的解释可以参见文献[10]。

3 实验

本文采用一个真实的多标签数据(Emotions 数据集)来验证 SemiBoost. MH 算法在多标签分类中的应用效果,并将结果与 AdaBoost. MH 算法相比较。其目的是为了说明利用非标记数据的信息能够有效地提高监督的 Boosting 算法的效果,并且随着非标记数据数量的增加,分类算法的效果可以随之提高。

Emotions 数据来自于 Mulan[10] 网站。它是一个音乐分类数据集,每一条数据代表了一首歌曲,标签代表了歌曲的类型。更详细的信息可以参 Mulan 网站。

3.1 评价标准

Hamming 损失是一种最常用的评估多标签分类效果的标准。此外,介绍几种其他的标准,这些标准也用于文献[3]。

$$\text{accuracy} = \sum_{i=1}^{M}\frac{|y_i=1 \wedge \text{sign}(F(x_i)=1)|}{|y_i=1 \vee \text{sign}(F(x_i)=1)|}$$

$$\text{precision} = \sum_{i=1}^{M}\frac{|y_i=1 \wedge \text{sign}(F(x_i)=1)|}{|\text{sign}(F(x_i)=1)|}$$

$$\text{recall} = \sum_{i=1}^{M}\frac{|y_i=1 \wedge \text{sign}(F(x_i)=1)|}{|y_i=1|}$$

其中:M 表示测试数据的数量,y_i 表示测试数据真实的标签。准确率(accuracy)、精确度(precision)和召回率(recall)越大,表示多标签分类的结果越好。

3.2 实验结果

从 Emotions 数据集中随机选取 40 条数据作为标记数据,其余的数据作为非标记数据、交叉验证数据和测试数据。其中交叉验证数据是为了选择最优的参数 γ。为了体现增加非标记数据数量的作用,变化非标记数据的数量从 40、80、120、160 到 200。表 1 为 Adaboost. MH 与 SemiBoost. MH 算法关于多个评价标准的比较结果。可以明显看出,对于所有这些评价标准来说,SemiBoost. MH 算法均优于 Adaboost. MH 算法。

表 1 AdaBoost. MH 与 SemiBoost. MH 算法结果比较

非标记数据数量	Hamming	accuracy	precision	recall
0	28.22	40.89	53.30	49.01
40	27.31	41.67	54.79	49.75
80	26.57	42.33	54.52	51.07
120	26.24	43.42	56.68	53.22
160	26.29	43.49	58.42	53.22
200	26.16	43.64	57.51	53.47

另外,SemiBoost. MH 算法的分类效果可以随着更多非标记数据的加入而提高。图 1 ~ 4 分别为 AdaBoost. MH 与 SemiBoost. MH 算法的 Hamming 损失、准确率、精确度、召回率。由图 1 ~ 4 可以看出,随着非标记数据数量的增加,Hamming 损失有下降的趋势,并且准确度、精确率和召回率有上升的趋势。

4 结束语

SemiBoost. MH 算法主要用于解决多标签的分类问题,它不仅仅是 AdaBoost. MH 算法的一种半监督的扩展,并且提供了一种更灵活的算法框架,允许引入不同的标记数据以及非标记数据的损失函数。实验结果表明,SemiBoost. MH 算法明显优于 AdaBoost. MH 算法,并且效果随着非标记数据数量的增加而提高。

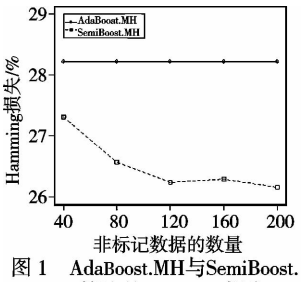

图 1 AdaBoost.MH与SemiBoost.MH 算法的Hamming损失

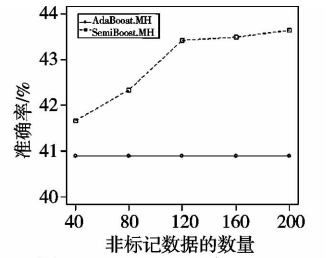

图 2 AdaBoost.MH与SemiBoost.MH算法的准确率

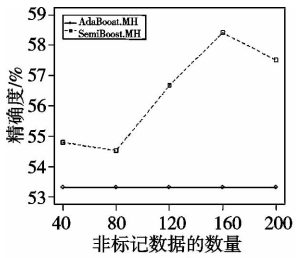

图 3 AdaBoost.MH与SemiBoost.MH算法的精确度

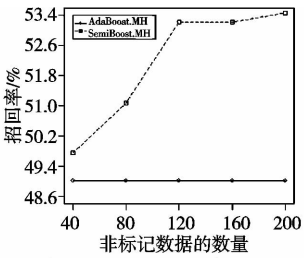

图 4 AdaBoost.MH与SemiBoost.MH算法的召回率

参考文献:

[1] CHAPELLE O, SCHOLKOPF B, ZIEN A. Semi-supervised learning [M]. Cambridge: The MIT Press, 2006: 213-304.

[2] ZHU Xiao-jin, GOLDBERG A. Introduction to semi-supervised learning [J]. Synthesis Lectures on Artificial Intelligence and Machine Learning, 2009, 3(1): 1-130.

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

[4] SCHAPIRE R E, SINGER Y. Improved boosting algorithms using confidence-rated predictions[J]. Machine Learning, 1999, 37(3): 297-336.

[5] SCHAPIRE R E, SINGER Y. BoosTexter: A boosting-based system for text categorization [J]. Machine Learning, 2000, 39(2/3): 135-168.

[6] FREUND Y, SCHAPIRE R E. A decision-theoretic generalization of on-line learning and an application to boosting [J]. Journal of computer and System Sciences, 1997, 55(1): 119-139.

[7] MASON L, BAXTER J, BARTLETT P, *et al*. Boosting algorithms as gradient descent [C]//Advances in Neural Information Processing Systems. 1999: 512-518.

[8] GRANDVALET Y, BENGIO Y. Semi-supervised learning by entropy minimization [C]//Advances in Neural Information Processing Systems. 2005: 529-536.

[9] LIU D C, NOCEDAL J. On the limited memory BFGS method for large scale optimization [J]. Mathematical Programming, 1989, 45(3): 503-528.

[10] BOYD S, BANDENBERHE L. Convex optimization [M]. Cambridge: Cambridge University Press, 2004: 151-189.

[11] TSOUMAKS G, SPYROMITRO-XIOUFIS E, VILCEK J, *et al*. Mulan: a Java library for multi-label learning [J]. Journal of Machine Learning Research, 2011, 12(7): 2411-2414.