

# 基于噪声概率的多核集成算法

冯 杰, 武德安, 吴 磊

(电子科技大学 数学科学学院, 四川 成都 611731)

**摘 要:** 为解决多核集成(Multiple Kernel Boosting, MKBoost)算法对噪声敏感的问题, 考虑理想分类器区别对待正常样本与噪声样本的特点, 提出了一种适用于被噪声污染数据集的多核集成算法. 采用 KNN(K 最近邻)方法与 logistic 回归的融合, 构造了样本噪声概率函数, 计算出每个样本是噪声的概率, 根据噪声概率构造了新的损失函数, 利用加法模型得到每轮迭代的基分类器系数. UCI 数据集上的实验结果表明, 该算法可以有效降低多核集成算法对噪声的敏感程度, 提高了鲁棒性.

**关键词:** 多核学习; 提升方法; K 最近邻法; logistic 回归; 支持向量机; 鲁棒性

**中图分类号:** TP391

**文献标识码:** A

**文章编号:** 1000-7180(2017)09-0063-04

## A Noise-Probability Based Multiple Kernel Boosting Algorithm

FENG Jie, WU De-an, WU Lei

(School of Mathematics, University of Electronic Science and Technology of China, Chengdu 611731, China)

**Abstract:** To solve the problem that the Multiple Kernel Boosting (MKBoost) algorithm was sensitive to noise, the characteristic that noise data and normal data were treated differently by the optimal classifier was considered, and a MKBoost algorithm applied to datasets polluted by noise data was proposed. The fuse of KNN (K-nearest neighbor) algorithm and logistic regression was used for the construction of noise probability function, and calculating the probability of each instance being noise. A new loss function was constructed via the noise probability function, and additive model was utilized for computing the coefficients corresponded to base classifier in each iteration. Experiments on the datasets from UCI show that the proposed method could efficiently reduce the sensitivity of MKBoost algorithm to noise, and improved the robustness

**Key words:** multiple kernel learning; boosting method; KNN; logistic regression; SVM; robustness

### 1 引言

在机器学习领域, 多核学习(Multiple Kernel Learning)越来越受到关注, 因为相比单核学习, 多核学习可以克服样本特征中规模巨大、异构信息、多维数据不规则及数据在高维特征空间分布不平坦的现象<sup>[1]</sup>. 近年来出现了多种有效的多核学习理论与方法, 如 2011 年陶剑文与王士同提出了多核局部领域适应学习方法<sup>[2]</sup> (Local Learning-based Domain Adaptation). 2012 年 Hao Xia, Steven 提出了集成多核学习的算法框架 MKBoost<sup>[3-4]</sup>, 该算法大大降

低了计算量, 也具有较高的精度, 但 Boosting 思想的引入, 也带来了对噪声敏感的问题<sup>[5]</sup>. 本文受加法模型和 logistic 回归的启发, 从考虑 Adaboost 算法对噪声样本敏感的角度, 提出了样本噪声程度概率化的多核集成算法 (Noise-Probability based MK-Boost, MKB\_NP), 为解决多核集成算法对噪声敏感的问题提供了一种可行方法.

### 2 MKB\_NP 算法

针对 Adaboost 算法对噪声敏感的问题, 本文引入噪声概率函数, 在新的损失函数基础上, 修改每

轮迭代中的样本权值,提出了 MKB\_NP 算法,满足了理想分类器两大要求<sup>[6]</sup>:(1)理想分类器尽量将非噪声样本分类正确;(2)理想分类器尽量将噪声样本分类错误。

## 2.1 Adaboost 算法简介

本文所提的方法是 AdaBoost 的拓展,所以首先简单介绍 Adaboost 算法:设实例空间为  $X$ ,训练数据集为  $\text{Train} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ ,迭代次数为  $T$ ,在二分类问题中,Adaboost 算法流程如下:

算法 1. AdaBoost algorithm

1:初始化训练数据的权值分布  $D_1 = \{1/n, 1/n, \dots, 1/n\}$

2:Do for  $t = 1; T$

① 基于样本权值分布  $D_t$  训练得到基分类器  $f_t(x)$

② 计算  $f_t(x)$  在训练数据集上的分类错误率

$$\begin{aligned}\epsilon_t &= \sum_{i=1}^n D_t[f_t(x) \neq y_i][f_t(x) \neq y_i] \\ &= \begin{cases} 1, & f_t(x) \neq y_i \\ 0, & f_t(x) = y_i \end{cases}\end{aligned}$$

③ 计算基分类器  $f_t(x)$  的系数

$$a_t = \frac{1}{2} \log \frac{1 - \epsilon_t}{\epsilon_t}$$

④ 更新训练样本的权值分布,其中  $Z_t$  为归一化因子

$$D_{t+1} = \begin{cases} D_t e^{-a_t} / Z_t & f_t(x) = y_i \\ D_t e^{a_t} / Z_t & f_t(x) \neq y_i \end{cases}$$

3:循环结束,得到最终强分类器

$$F(x) = \text{sign}\left(\sum_{t=1}^M a_t f_t(x)\right)$$

## 2.2 构造噪声概率函数

### 2.2.1 K 近邻算法

K 近邻(KNN)算法以其简单性、有效性而成为经典的非参数分类算法,在机器学习中得到广泛的应用。分类时,对新的实例,根据其  $K$  个最近邻训练实例的类别,通过多数表决等方式进行预测。文献<sup>[6]</sup>采用 KNN 算法对数据集进行噪声探测,主要利用了目标样本的邻域信息,假设样本  $(x_i, y_i)$  的  $K$  个最近邻点集合为  $Z_i$ ,基分类器为  $f_t(x)$ ,真实类别为  $y_j$ ,则噪声探测结果  $u^{KNN}$  如式(1),若集合  $Z_i$  中分类错误的样本越多,则样本  $(x_i, y_i)$  是噪声的可能性越大。

$$u^{KNN}(x_i, y_i) = \sum_{(x_j, y_j) \in Z_i} I(f_t(x_j) \neq y_j) / K \quad (1)$$

### 2.2.2 噪声概率函数

Sam Kwong 等人提出了基于噪声探测器的 ND\_Adaboost (Noise-Detection Adaboost) 算法<sup>[6]</sup>,

该算法将训练集噪声探测结果的平均值  $u$  作为阈值,进行二值化处理,噪声探测函数  $p(x)$  如式(2)。若  $u^{KNN} > u$ ,则代表  $(x_i, y_i)$  为噪声样本。

$$p(x_i) = \text{sign}(u - u^{KNN}(x_i, y_i)),$$

$$u = \sum_{i=1}^n u^{KNN}(x_i, y_i) / n \quad (2)$$

如果能精确地识别噪声样本,探测函数  $p(x)$  二值化处理无疑是最理想的方法,既简单又高效。但由于训练集自身的噪声率不可忽视,“精确识别”这个条件是极为苛刻的,在实际中很难达到。本文引入样本噪声概率函数  $p^*(x)$ ,如式(3),由于  $(u - u^{KNN})$  的值域是  $(-1, 1)$  的子集,对应的 logistic 回归函数的值域区间过窄,所以我们引入参数  $b$ ,使得噪声概率函数  $p^*(x)$  更有效地区分噪声样本与正常样本。式(3)主要利用 logistic 回归模型预测噪声概率,与式(2)中的  $p(x)$  相比,区别在于将噪声探测结果映射到  $[0, 1]$ ,而不是  $\{-1, +1\}$ 。这种处理容错率更高,更符合实际情况。

$$p^*(x_i) = \frac{1}{1 + \exp[-b(u - u^{KNN})]} \quad (3)$$

### 2.3 损失函数

损失函数是描述系统在不同参数值之下的损失,Adaboost 算法是在所有基分类器的线性组合空间上最小化损失函数,其损失函数为式(4)。通过 Adaboost 自适应性的分析可以发现,那些较难被基分类器判别正确的样本将在下一轮迭代中获得相对更高的权重,而噪声样本由于背离数据的正常分布,难以被基分类器准确识别,其权值极有可能呈指数增长,从而导致 Adaboost 对噪声敏感,影响其分类器泛化能力。

$$L_{ada}(y, f(x)) = \sum_{i=1}^n \exp(-y_i f(x_i)) \quad (4)$$

鉴于以上不足,基于噪声概率函数  $p^*(x)$ ,提出了新的损失函数形式,如式(5),当  $p^*(x)$  趋近于 0,即该样本是噪声的概率接近 0,则新损失函数  $L(y, f_t(x))$  与  $L_{ada}(y, f_t(x))$  一致,正常样本分类正确会减小损失函数,满足理想分类器的第一个要求;反之  $p^*(x)$  趋近于 1 时,即该样本为噪声的概率越接近 1,噪声样本分类错误会减小损失函数,满足理想分类器的第二个要求。

$$L(y, f_t(x)) = \sum_{i=1}^n [(1 - p^*(x_i)) \exp(-y_i f_t(x_i)) + p^*(x_i) \exp(y_i f_t(x_i))] \quad (5)$$

### 2.4 算法推导与算法流程

Adaboost 算法是模型为加法模型、损失函数为

指数函数、学习算法为前向分布算法的分类学习算法<sup>[7]</sup>. 所以针对式(5)中的损失函数,推导出本文的 MKB\_NP 算法.

假设经过  $t-1$  轮训练,已经得到  $F_{t-1}(x)$ ,在第  $t$  轮训练中得到  $a_t, f_t(x)$  和  $F_t(x)$ ,则有  $F_t(x) = F_{t-1}(x) + a_t f_t(x)$ ,最小化式(4)中的损失函数,即求得系数  $a_t$  与基分类器  $f_t(x)$ ,使得强分类器  $F_t(x)$  在训练数据集上的损失函数最小,则有

$$\begin{aligned} (a_t, f_t) &= \operatorname{argmin}_{a, f(x)} \sum_{i=1}^n (1 - p^*(x_i)) \\ &\quad \exp(-y_i (F_{t-1}(x_i) + a f(x_i))) + \\ &\quad p^*(x_i) \exp(y_i (F_{t-1}(x_i) + a f(x_i))) \\ &= \operatorname{argmin}_{a, f(x)} \sum_{i=1}^n (W_{1i}^t \exp(-y_i a f(x_i)) + \\ &\quad W_{2i}^t \exp(y_i a f(x_i))) \end{aligned} \quad (6)$$

式(6)中,  $W_{1i}^t, W_{2i}^t$  形式如式(7),显然  $W_{1i}^t, W_{2i}^t$  只与  $F_{t-1}(x)$  相关,故两者均与  $a_t, f_t(x)$  相互独立.

$$\begin{aligned} W_{1i}^t &= (1 - p^*(x_i)) \exp(-y_i F_{t-1}(x_i)), \\ W_{2i}^t &= p^*(x_i) \exp(y_i F_{t-1}(x_i)) \end{aligned} \quad (7)$$

首先求解第  $t$  轮训练的基分类器  $f_t(x)$ ,对任意的  $a > 0$  有

$$\begin{aligned} &\sum_{i=1}^n (W_{1i}^t \exp(-y_i a f(x_i)) + W_{2i}^t \exp(y_i a f(x_i))) \\ &= \sum_{f(x_i)=y_i} (W_{1i}^t e^{-a} + W_{2i}^t e^a) + \\ &\quad \sum_{f(x_i) \neq y_i} (W_{1i}^t e^a + W_{2i}^t e^{-a}) \\ &= (e^a - e^{-a}) \sum_{f(x_i) \neq y_i} (W_{1i}^t - W_{2i}^t) + \\ &\quad \sum_{i=1}^n (W_{1i}^t e^{-a} + W_{2i}^t e^a) \end{aligned} \quad (8)$$

式(8)中与  $f_t(x)$  有关的,只有左边连加符号部分,故要使得损失函数最小,则有

$$f_t(x) = \operatorname{argmin}_{f(x)} \sum_{f(x_i) \neq y_i} (W_{1i}^t - W_{2i}^t) \quad (9)$$

然后求解第  $t$  轮训练的分类器系数  $a_t$ ,式(8)中对  $a$  求导,令导数为 0,则有

$$a_t = \frac{1}{2} \log \left[ \frac{\sum_{f(x_i) \neq y_i} W_{2i}^t + \sum_{f(x_i) = y_i} W_{1i}^t}{\sum_{f(x_i) \neq y_i} W_{1i}^t + \sum_{f(x_i) = y_i} W_{2i}^t} \right] \quad (10)$$

根据式(9)与式(10)得到的  $a_t, f_t(x)$ ,更新  $W_{1i}^t, W_{2i}^t$ ,如式(11)所示,最后完成对样本权重的更新  $D^{t+1}$ ,如式(12)所示.

$$\begin{aligned} W_{1i}^{t+1} &= W_{1i}^t \exp(-y_i a_t f(x_i)), \\ W_{2i}^{t+1} &= W_{2i}^t \exp(y_i a_t f(x_i)) \end{aligned} \quad (11)$$

$$D_i^{t+1} = \frac{|W_{1i}^{t+1} - W_{2i}^{t+1}|}{\sum_{i=1}^n |W_{1i}^{t+1} - W_{2i}^{t+1}|} \quad (12)$$

针对二类分类问题,根据上述推导过程, MKB\_NP 算法的具体步骤如下

算法 2. NP\_MKB algorithm

输入:训练集 Train,  $M$  个核函数  $k_j, j = 1, 2, \dots, M$

输出:最终强分类器  $F(x)$

1: Do for  $i = 1; M$

以  $k_j$  为核函数,在训练集 Train 上得到弱分类器  $f_i(x)$ ,并由式(1)、(2)、(3)计算每个样本的噪声概率  $p^*(x)$ ,初始化  $W_{1i}^1, W_{2i}^1$  与初始权值分布  $D^1$  如下:

$$W_{1i}^1 = \frac{\sum_{j=1}^M (1 - p(x_i))}{M}, W_{2i}^1 = \frac{\sum_{j=1}^M p(x_i)}{M},$$

$$D_i^1 = \frac{|W_{1i}^1 - W_{2i}^1|}{\sum_{i=1}^n |W_{1i}^1 - W_{2i}^1|}$$

2: Do for  $t = 1; T$

Do for  $j = 1; M$

依照样本权值分布  $D^t$  训练得到弱分类器  $f_i^t(x)$ ,并按照式(9)选择该轮基分类器  $f_i^*(x)$ ,根据式(10)计算其对应的系数  $a_i^*$ ,如果  $a_i^* < 0$ ,则  $W_{1i}^{t+1} = W_{1i}^t, W_{2i}^{t+1} = W_{2i}^t$ ,否则按照式(11)、(12)更新  $W_{1i}^{t+1}, W_{2i}^{t+1}$  以及  $D^{t+1}$

3: 循环结束,得到最终强分类器  $F(x) = \operatorname{sign}(\sum_{i=1}^M a_i f_i(x))$

### 3 实验结果

本文随机选择了 6 个 UCI<sup>[8]</sup>数据集,分别使用新算法与文献[3]中的 MKB\_D1、MKB\_D2 算法进行对比实验,实验数据集如表 1 所示. 实验中本文分别随机修改 10%、20%、30% 的训练样本类别属性值,以得到不同噪声水平的训练集.

表 1 实验数据集信息

| Datasets             | Samples | Features | Classes |
|----------------------|---------|----------|---------|
| Balance-scale        | 567     | 4        | 2       |
| Breast-cancer        | 569     | 32       | 2       |
| Ionosphere           | 351     | 34       | 2       |
| Blood-transfusion    | 748     | 5        | 2       |
| Diabetic-retinopathy | 1 151   | 20       | 2       |
| Pima-indians         | 768     | 8        | 2       |

对于每个数据集,本文采用 8 个核函数(5 个高斯核函数,3 个多项式核函数),利用 LIBSVM 工具箱<sup>[9]</sup>,在 Matlab2013b 上进行实验. 我们在每个噪声水平下,对每个数据集重复实验 20 次,每次随机选择一半的样本作为训练集,剩下的作为测试集. 训练

样本进行零均值,单位方差化处理,测试样本按照训练样本的均值、方差进行归一化. 最大迭代次数  $T=100$ ,在样本噪声概率计算时,KNN 中  $K=11$ ,采用欧氏距离进行距离度量,噪声概率函数  $p^*(x)$  中,参数  $b=18$ . 实验结果为 20 次实验的平均值,如表 2 所示.

表 2 MKB\_NP 与 MKB\_D1、MKB\_D2 测试准确率比较

| Datasets             | Noise | MKB_D1         | MKB_D2         | MKB_NP         |
|----------------------|-------|----------------|----------------|----------------|
| Balance Scale        | 0%    | 0.885 4        | <b>0.908 0</b> | 0.877 2        |
|                      | 10%   | 0.843 4        | 0.852 4        | <b>0.901 4</b> |
|                      | 20%   | 0.801 8        | 0.863 5        | <b>0.872 6</b> |
|                      | 30%   | 0.827 3        | 0.810 6        | <b>0.856 8</b> |
| Breast Cancer        | 0%    | 0.816 4        | 0.827 8        | <b>0.837 5</b> |
|                      | 10%   | 0.803 1        | 0.797 6        | <b>0.845 2</b> |
|                      | 20%   | 0.782 3        | 0.768 1        | <b>0.812 8</b> |
|                      | 30%   | 0.751 7        | <b>0.812 8</b> | 0.807 8        |
| Ionosphere           | 0%    | 0.858 8        | <b>0.895 7</b> | 0.867 8        |
|                      | 10%   | <b>0.864 6</b> | 0.843 7        | 0.852 4        |
|                      | 20%   | 0.820 5        | 0.811 4        | <b>0.839 8</b> |
|                      | 30%   | 0.803 2        | 0.801 0        | <b>0.814 5</b> |
| Blood Transfusion    | 0%    | 0.786 6        | <b>0.805 3</b> | 0.802 4        |
|                      | 10%   | <b>0.792 3</b> | 0.753 8        | 0.774 7        |
|                      | 20%   | 0.765 8        | 0.765 8        | <b>0.796 6</b> |
|                      | 30%   | 0.740 6        | 0.752 1        | <b>0.792 3</b> |
| Diabetic Retinopathy | 0%    | 0.812 5        | 0.805 6        | <b>0.819 5</b> |
|                      | 10%   | <b>0.808 2</b> | 0.793 2        | 0.800 0        |
|                      | 20%   | 0.752 3        | <b>0.799 2</b> | 0.786 7        |
|                      | 30%   | 0.749 2        | 0.754 5        | <b>0.779 6</b> |
| Diabetic Retinopathy | 0%    | 0.712 3        | 0.717 2        | <b>0.728 7</b> |
|                      | 10%   | 0.694 3        | 0.682 8        | <b>0.740 5</b> |
|                      | 20%   | 0.674 2        | 0.675 8        | <b>0.708 7</b> |
|                      | 30%   | 0.648 5        | 0.653 4        | <b>0.685 9</b> |

黑体数据表示在相同噪声水平下该算法的平均测试准确率最高.

由表 2 可知,在无噪声的情况下,三种算法测试准确率相当,但随着噪声水平的提高,MKB\_D1 与 MKB\_D2 测试准确率波动较大,而 MKB\_NP 算法测试准确率波动较小. 当噪声水平为 10% 时, MKB\_NP 算法在 Balance-scale、Breast Cancer、Pima-indians 这三个数据集中表现优于其他两个算法,占比为 50%;当噪声水平为 20% 时,新算法在除了 Diabetic Retinopathy 外的其他 5 数据集中测试准确率最高,占比为 83%;当噪声水平为 30% 时,MKB\_NP 算法在除了 Breast-cancer 外的其他 5 个数据集中表现最为优异,占比为 83%. 综上所述,MKB\_NP 算法在 6 个数据集上的表现优于 MKB\_D1 与 MKB\_D2 算法,且在较高噪声水平的数据分类中,对噪声数据敏感性更低,训练误差更小,鲁棒性更好.

4 结束语

针对传统多核集成算法中对噪声敏感的问题,本文利用了集成多核学习 MKBoost 的框架,引入样本噪声概率函数,构造了新的损失函数,推导出 MKB\_NP 算法,实验表明,MKB\_NP 算法有效降低了噪声敏感程度,提高了鲁棒性. MKB\_NP 算法在一定程度上解决了噪声干扰的问题,但也存在待优化之处:(1) 由于采用了 KNN 算法,所以对  $K$  值较敏感,且参数  $b$  的取值大小也会影响样本噪声概率,因此需要对该算法进行更深入的研究,探索参数对算法的影响;(2) 算法关键在于噪声概率的计算,所以可以改进噪声概率的计算方法,使得噪声识别更精确.

参考文献:

[1] 王洪桥,孙富春,蔡艳宁,等. 多核学习方法[J]. 自动化学报,2010,36(8):1038-1050.

[2] 陶剑文,王士同. 多核局部领域适应学习[J]. 软件学报,2012,23(9):2297-2310.

[3] Xia H, Hoi S C H. Mkboost: A framework of multiple kernel boosting[J]. IEEE Trans. on Knowledge and Data Engineering, 2013, 25(7): 1574-1586.

[4] Xia H, Hoi S C H. Mkboost: A framework of multiple kernel boosting[C]// Proc. 11th SIAM Int'l Conf. Data Mining(SDM). SIAM,2011: 199-210.

[5] Masnadi-Shirazi H, Vasconcelos N. Cost-sensitive boosting [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence ,2010, 33(2) :294-309.

[6] Jingjing Cao, Sam Kwong, Ran Wang. A noise-detection based Adaboost algorithm for mislabeled data[J]. Pattern Recognition,2012,45: 4451-4465.

[7] 李航. 统计学习方法[M]. 北京:清华大学出版社, 2012;143-146.

[8] Frank A, Asuncion A. UCI machinelearning repository [EB/OL]. [2016-12-18]. <http://archive.ics.uci.edu/ml>.

[9] Chang C C , Lin C J . LIBSVM: A library for support vector machines [EB/OL]. [2016-12-18]. <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

作者简介:

冯 杰 男,(1989-),硕士研究生. 研究方向为数据挖掘与计算统计. E-mail:1661798458@qq.com.

武德安 男,(1971-),博士,副教授. 研究方向为模式识别、数据挖掘.

吴 磊 男,(1978-),博士,副教授. 研究方向为模式识别、数据挖掘.