第45卷第1期 中国科学技术大学学报 Vol. 45, No. 1
2015年1月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Jan. 2015

文章编号:0253-2778(2015)01-0061-08

一种基于类别不平衡数据的层次分类模型

施培蓓[1],刘贵全[2],汪 中[3],卫 兵[1]

(1.合肥师范学院公共计算机教学部,安徽合肥 230601;2.中国科学技术大学计算机学院,安徽合肥 230027;
3.中国电子科技集团公司第三十八研究所数字技术部,安徽合肥 230088)

摘要:传统的机器学习方法在处理类别不平衡数据时分类性能较低,为此提出一种基于类别不平衡数据的层次分类模型.层次分类模型采用 AdaBoost 方法为基准分类器,以分类器误报率和特征建立数学模型,并证明层次分类模型的参数可以计算得到.首先以层次分类树为结构建立模型,接着针对层次分类树的结构模型进行分类代价计算,得到模型的代价与每层特征之间的定量数学描述,然后将该分类代价转换为优化问题并给出优化问题的求解过程,同时给出层次分类模型的计算结果.在 UCI 数据集上进行大量测试,以 AUC 和 F-Measure 为评价标准,相比于现有的不平衡分类方法,层次分类模型具有更优的分类性能.

关键词:机器学习;类别不平衡;层次分类;特征;评价标准

中图分类号:TP391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2015.01.010

引用格式:Shi Peibei, Liu Guiquan, Wang Zhong, et al. A hierarchical classification model for class-imbalanced data[J]. Journal of University of Science and Technology of China, 2015,45(1):61-68.
施培蓓,刘贵全,汪中,等. 一种基于类别不平衡数据的层次分类模型[J]. 中国科学技术大学学报,2015,45(1):61-68.

A hierarchical classification model for class-imbalanced data

SHI Peibei[1], LIU Guiquan[2], WANG Zhong[3], WEI Bing[1]

(1. *Department of Public Computer Teaching, Hefei Normal University, Hefei* 230601, *China*;
2. *School of Computer Science and Technology, University of Science and Technology of China, Hefei* 230027, *China*;
3. *Department of Digital Technology, No.* 38 *Research Institute of* CETC, *Hefei* 230088, *China*)

Abstract: Traditional machine learning methods have lower classification performance when dealing with class imbalanced data. A hierarchical classification model for class imbalanced data was thus proposed. With an AdaBoost classifier as its basis classifier, the model builds mathematical models by the features and false positive rates of the classifier, and demonstrates that parameters of the hierarchical classification model could be calculated. First, the hierarchical classification tree was as the structure, and then the classification cost of the hierarchical classification tree mode was obtained as well as a quantitative and mathematical description of the features of each layer. Finally, the classification cost could be converted to a optimization problem, and the solving process of the optimization problem was given. Meanwhile,

收稿日期:2014-06-09;**修回日期**:2014-07-29

基金项目:国家科技支撑计划(2012BAH17B03),安徽省自然科学基金(1408085MF131),安徽省高等学校自然科学项目(KJ2013B212),合肥师范学院魂芯 DSP 产业化研究院开放课题资助.

作者简介:施培蓓(通讯作者),女,1984 年生,硕士/讲师. 研究方向:数据挖掘、机器学习. E-mail: pb_shi@163.com

results of the hierarchical classification are presented. Experiments have been conducted on UCI dataset, and the results show that the proposed method has higher AUC and F-measure compared to many existing class-imbalanced learning methods.

Key words: machine learning; class-imbalanced; hierarchical classification; feature; evaluation criteria

0 引言

不平衡分类是指各类训练样本的数量存在不平衡的情况,如二分类问题中正负样本数量相差较大等.在日常生活中,不平衡分类问题广泛存在,如信用卡欺诈检测[1]、文本分类[2]、信息检索与过滤[3]、市场行为分析[4]、石油勘测[5]、医学诊断[6]、网络入侵检测[7]等.例如,在信用卡欺诈交易检测问题中,绝大部分交易都是正常的交易,只有很少一部分是非法交易,而该部分也是研究者关注的重点对象.如果使用传统的机器学习方法,分类性能将大大降低,因此利用数据挖掘技术提高不平衡分类的精度是现今的难点.近几年,数据挖掘会议也开始关注不平衡分类问题,如AAAI、ICML、ICDM、KDD等.

在计算机视觉中也存在很多不平衡分类问题,如人脸识别、行人检测等,待检测图像中包含人脸或行人的样本很少,绝大部分都是负样本.Sahbi等[8]提出层级树状分类器并成功应用于人脸识别,其单分类器使用的是SVM分类器,且该方法还不能直接用于不平衡分类问题.受该思想的启发,论文提出一种基于类别不平衡数据的层次分类模型HAdaBoost.HAdaBoost采用AdaBoost作为单分类器,采用层次分类树架构建立模型.首先对单分类器的性能进行描述,接着对层次分类模型的分类代价展开分析,得到层次分类模型分类代价的定量描述;然后将模型转换为优化问题,并给出优化问题的具体求解过程;最终给出层次分类模型的计算参数.在UCI数据库[9]上实验,结果表明HAdaBoost方法可以有效地提高不平衡数据的分类性能.

1 相关工作

解决不平衡分类器问题主要包括数据层面和算法层面.数据层面从数据集出发,通过对训练数据集重新采样来解决数据的不平衡程度,主要包括过采样技术和欠采样技术.典型的过采样方法是Chawla等[10]提出的SMOTE过采样方法,该方法通过在训练集中添加有用信息达到增加更多少数类样本的目的,实验结果证明,该方法远优于随机过采样技术.Han等[11]对SMOTE方法进行改进,获得比SMOTE更好的分类效果.Liu等[12]将过采样技术应用于文本分类并取得了良好的分类性能.过采样技术是增加少数类的数据,而欠采样技术主要是去掉噪声和冗余的数据.常用的欠采样技术包括单边选择、编辑技术、一致子集等[13],这些方法主要采用启发式方法,利用KNN规则识别可以去除的样本.由于过采样和欠采样技术都存在一定的缺陷,因此混合采样技术受到越来越多的关注[14].

算法层面主要包括代价敏感学习、一类学习和集成分类器.代价敏感学习充分考虑不同类别的误分代价.Domingos等[15]提出的MetaCost方法通过估计训练样本的后验概率密度,结合代价矩阵计算每个样本的类别.Chen等[16]提出加权随机森林算法,训练样本最小的类赋予最大的权值.Chew等[17]通过训练集先验信息的分析,利用SVM为不同类别的样本设置惩罚系数.代价敏感学习能有效提高少数类的分类性能,但多数情况下,真实的错分代价难以准确估计.一类学习也被用于不平衡分类,如Raskutti等[18]证明一类学习在特征空间中混杂有大量噪音特征时具有重要作用.Jusczak等[19]将一类学习和重采样技术结合,将有用信息加入训练集.集成分类器是目前解决不平衡分类较为成熟的技术.Zhou等[20]提出代价敏感神经网络与分类器集成相结合的方法,在UCI数据集上的实验结果表明,该方法对二类和多类不平衡问题均有效.Liu等[21]综合重采样技术提出EasyEnsemble和BalanceCascade两种集成分类器.EasyEnsemble分类器的核心思想是每次从负样本集合中独立抽取与正样本相同数量的训练集,然后将其与正样本集合训练AdaBoost分类器,最终结果是所有的AdaBoost算法的输出的总和.而BalanceCascade分类器抽取样本的方法与EasyEnsemble相同,区别是在每层分类器的训练过程中控制AdaBoost分类器的阈值,使得分类器的误报率等于设定值,同时该层被错误分类的负样本重新划分到下一层成为新的数据集合.相比于AdaBoost、Bagging、SMOTEBoost、AsymBoost、随机森林等传统集成分

类器，EasyEnsemble 和 BalanceCascade 方法具有更优的分类性能. Galar 等[22]对基于集成学习的不平衡分类方法进行综述，包括 Bagging、Boosting 及其组合方法，并在 UCI 数据集上进行大量的测试.

2　层次分类模型

2.1　AdaBoost 方法

AdaBoost 算法是 Freund 和 Schapire 根据在线分配算法提出的，他们详细分析了其错误率的上界. AdaBoost 是一种迭代算法，不容易出现过拟合现象，其核心思想是针对同一个训练集训练不同的弱分类器，然后将这些弱分类器级联构成最终的强分类器. 论文采用 Volia[23]提出的 AdaBoost 算法框架，弱分类器 $h(x,f,p,\theta)$是一个关于特征的阈值函数，表示如下：

$$h(x,f,p,\theta)=\begin{cases}1, & \text{if } p\times f(x)<p\times\theta\\ 0, & \text{otherwise}\end{cases}.$$

其中，$f(x)$是样本的特征值，θ 为阈值函数，p 为不等式方向，不等式方向 p 和阈值 θ 可以通过对样本加权平均比较得到.

AdaBoost 算法的流程如下：

Step 1　给定一组训练样本$(x_i,y_i)\mid i=1,\cdots,n$，$T$ 为最大迭代次数，初始化正负样本的权重 $w_i=1/n$；

Step 2　for $t=1\cdots T$

① 归一化样本权重 $w_i^t=w_i^t/(\sum_{j=1}^{n}w_j^t)$；

② 对于每个特征训练一个弱分类器并计算当前误差，选择错误率最低的一个弱分类器，计算公式为：$\varepsilon_t=\min\limits_{f,p,\theta}\sum\limits_i w_i^t\mid h(x_i,f,p,\theta)-y_i\mid$；

③ 对于弱分类器正确分类的样本更新样本权重 $w_i^{t+1}=w_i^t\varepsilon_t/(1-\varepsilon_t)$. 对于错误分类的样本，权重保持不变；

Step 3　输出最终的强分类器

$$h(x)=\begin{cases}\text{true}, & \sum\limits_{t=1}^{T}\alpha_t h_t(x)\geqslant\dfrac{1}{2}\sum\limits_{t=1}^{T}\alpha_t\\ 0, & \text{otherwise}\end{cases}\qquad(1)$$

式中，$\alpha_t=\log\dfrac{1-\varepsilon_t}{\varepsilon_t}$.

对于一个具有 n 个特征的 AdaBoost 算法，我们假定其分类代价 $\cos t=a\times n+b$，a,b 为参数.

2.2　数学模型

HAdaBoost 属于层次分类器，采用二叉层次分类树来描述. 对于不平衡分类问题，正负样本数量相差很大，HAdaBoost 方法采用早筛选的策略，整体分类代价较小. 对于一个判定为正的样本，需要有从根节点到叶子节点的完整路径，而对于一个判定为负的样本，访问分类器较少，所需的代价也较小. 图 1 给出了二叉层次分类树的示意图，其中黑色圆圈表示正样本，白色圆圈表示负样本. 二叉层次分类树为完全二叉树结构，分类搜索采用深度优先搜索方式进行.

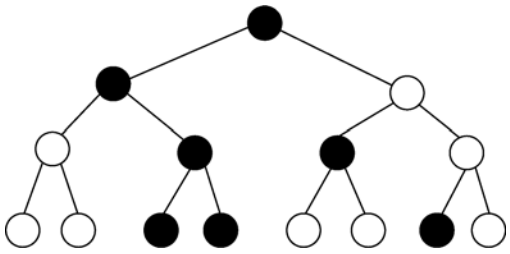

图 1　二叉层次分类树

Fig. 1　Binary hierarchical classification tree

假定二叉层次分类树共 L 层，$C_{l,k}$表示第 l 层的第 k 个分类器，$n_{l,k}$表示第 l 层的第 k 个分类器使用的特征数量，v_l 表示层号为 l 的节点数目，则 $v_l=2^{l-1}$. 考虑到单分类器 AdaBoost 在每层使用的特征数目相同，故 $n_{l,k}=n_l$.

假定 $\delta(l-1;n)$ 表示一个样本被 $1,\cdots,l-1$ 层分类错误的概率且 $\delta(0;n)=1$，则一个样本的分类代价为 $\cos t=\sum\limits_{l=1}^{L}\sum\limits_{k=1}^{v_l}1_{\{C_{l,k}\text{被执行}\}}n_{l,k}$. 由于本文考虑的是不平衡分类问题，正负样本数量差别较大. 相对于负样本对象，我们认为正样本对象的分类代价可忽略不计，则 HAdaBoost 方法的整体代价主要考虑的负样本的分类代价. 针对负样本对象，如果一个负样本对象在第 l 层被访问，那么其被执行的概率等价于该样本被 $1,\cdots,l-1$ 层分类器错误的判断为“正”(即误报率)，因此层次分类器的分类代价如下：

$$E(\cos t)=\sum_{l=1}^{L}\sum_{k=1}^{v_l}1_{\{C_{l,k}\text{被执行}\}}n_{l,k}=\sum_{l=1}^{L}\sum_{k=1}^{v_l}\delta(l-1;n)n_{l,k}=\sum_{l=1}^{L}v_l\delta(l-1;n)n_l=n_1+\sum_{l=2}^{L}v_l n_l\delta(l-1;n)\qquad(2)$$

整个层次分类树的误报率为第 L 层所有单分类器的误报率之和为 $v_L\delta(L;n)$. 为了保证不平衡分类问题的性能最优，就必须分类代价最小，则上述代

价最终可转换为最小化问题. 即

$$\left.\begin{aligned}&\min_{n_1,\cdots,n_L} n_1+\sum_{l=2}^{L} v_l n_l \delta(l-1;n)\\ &\text{s.t.}\quad \begin{cases} v_L\delta(L;n)\leqslant \mu\\ 0<n_l\leqslant N_l\end{cases}\end{aligned}\right\} \tag{3}$$

式中，μ 表示误报率的上限，N_l 表示第 l 层单分类器可以使用的最大特征数量.

根据边际和条件概率可以进一步推导出 $\delta(l-1;n)$ 与 l、n 之间的函数关系. 假定 $\delta(l-1;n)=(\sum_{j=1}^{l}\beta_j n_j)^{-1}$. 其中，$\beta_1,\cdots,\beta_l$ 为常数. 则式(3) 最小化问题可转换为：

$$\left.\begin{aligned}&\min_{n_1,\cdots,n_L} n_1+\sum_{l=2}^{L} v_l n_l(\sum_{j=1}^{l}\beta_j n_j)^{-1}\\ &\text{s.t.}\quad \begin{cases} v_L(\sum_{j=1}^{L}\beta_j n_j)^{-1}\leqslant \mu\\ 0<n_l\leqslant N_l\end{cases}\end{aligned}\right\} \tag{4}$$

2.3 最小化求解

为了使得 HAdaBoost 方法寻优时间最小，本节给出式(4) 的具体求解过程. 我们的目的是求得一组特征值 $n_1,\cdots,n_L$，使得分类模型的分类代价最小. 令 $C=\min_{n_1,\cdots,n_L} n_1+\sum_{l=2}^{L} v_l n_l(\sum_{j=1}^{l}\beta_j n_j)^{-1}$，当 n_L 确定时，具体求解过程如下.

首先对 C 求 n_1、n_j 偏导，可以得到

$$\frac{\partial C}{\partial n_1}=1-\sum_{l=2}^{L}\left[\frac{\beta_1 2^{l-1}n_l}{(\sum_{i=1}^{l-1}\beta_i n_i)^2}\right],$$

$$\frac{\partial C}{\partial n_j}=\frac{2^{j-1}}{\sum_{i=1}^{j-1}\beta_i n_i}-\sum_{l=j+1}^{L}\left[\frac{\beta_j 2^{l-1}n_l}{(\sum_{i=1}^{l-1}\beta_i n_i)^2}\right],$$

$$j\in\{2,L-1\} \tag{5}$$

这样我们得到

$$\frac{\partial C}{\partial n_{j+1}}=0\Rightarrow\sum_{l=j+2}^{L}\left[\frac{2^{l-1}n_l}{(\sum_{i=1}^{l-1}\beta_i n_i)^2}\right]=\frac{2^j}{\sum_{i=1}^{j-1}\beta_i n_i}\frac{1}{\beta_{j+1}},$$

$$\frac{\partial C}{\partial n_j}=0\Rightarrow\frac{2^{j-1}}{\sum_{i=1}^{j-1}\beta_i n_i}-\beta_j\frac{2^j n_{j+1}}{(\sum_{i=1}^{j}\beta_i n_i)^2}=$$

$$\beta_j\sum_{l=j+2}^{L}\left[\frac{2^{l-1}n_l}{(\sum_{i=1}^{l-1}\beta_i n_i)^2}\right] \tag{6}$$

根据式(6)得到

$$\frac{2^{j-1}}{\sum_{i=1}^{j-1}\beta_i n_i}-\beta_j\frac{2^j n_{j+1}}{(\sum_{i=1}^{j}\beta_i n_i)^2}-\beta_j\frac{2^j}{\beta_{j+1}(\sum_{i=1}^{j}\beta_i n_i)}=0,$$

$$j\neq 1 \tag{7}$$

假定 n_1 已知，根据 $\frac{\partial C}{\partial n_1}=0$ 和 $\frac{\partial C}{\partial n_2}=0$，我们得到

$$1-\frac{\beta_1 2 n_2}{\beta_2^2 n_1^2}-\frac{2\beta_1}{\beta_2\beta_1 n_1}=0\Rightarrow n_2=\frac{1}{2}\frac{\beta_1}{\beta_2}n_1(\beta_2 n_1-2) \tag{8}$$

假定 $2\leqslant j\leqslant l(l\in\{2,L-2\})$，则

$$n_j=2^{-\frac{1}{2}j(j-1)}(\prod_{i=1}^{j-1}\beta_i)\beta_j^{-1}n_1^{j-1}(\beta_j n_1-2^{j-1}) \tag{9}$$

现在我们证明

$$n_{l+1}=2^{-\frac{1}{2}l(l+1)}(\prod_{i=1}^{l}\beta_i)\beta_{l+1}^{-1}n_1^{l}(\beta_{l+1}n_1-2^l) \tag{10}$$

根据式(9)，对于任意 $j\in\{2,\cdots,l\}$，我们有

$$\sum_{i=1}^{j}\beta_i n_i=\beta_1 n_1+\beta_2\frac{1}{2}\beta_1\beta_2^{-1}n_1+$$

$$\beta_3\frac{1}{8}\beta_1\beta_2\beta_3^{-1}n_1^2(\beta_3 n_1-4)+\cdots+$$

$$\beta_{j-1}(2^{-\frac{1}{2}(j-1)(j-2)})(\prod_{i=1}^{j-2}\beta_i)\beta_{j-1}^{-1}n_1^{j-2}(\beta_{j-1}n_1-2^{j-2})+$$

$$\beta_j(2^{-\frac{1}{2}j(j-1)})(\prod_{i=1}^{j-1}\beta_i)\beta_j^{-1}n_1^{j-1}(\beta_j n_1-2^{j-1}) \tag{11}$$

因此，对于 $\forall\, j\in\{2,\cdots,l\}$

$$\sum_{i=1}^{j}\beta_i n_i=2^{-\frac{1}{2}j(j-1)}(\prod_{i=1}^{j-1}\beta_i)n_1^j \tag{12}$$

令 $\pi_j=\prod_{i=1}^{j}\beta_i$，将 $j=l$ 代入式(11)，我们重写式(7) 得到

$$\frac{2^{l-1}}{2^{-\frac{1}{2}(l-1)(l-2)}\pi_{l-1}n_1^{l-1}}-\beta_l\frac{2^l n_{l+1}}{(2^{-\frac{1}{2}l(l-1)}\pi_l n_1^l)^2}-\frac{\beta_l}{\beta_{l+1}}\frac{2^l}{2^{-\frac{1}{2}l(l-1)}\pi_l n_1^l}=0\Rightarrow$$

$$n_{l+1}=2^{-\frac{1}{2}l(l+1)}\pi_l\beta_{l+1}^{-1}n_1^l(\beta_{l+1}n_1-2^l) \tag{13}$$

这样我们证明了式(10). 对于 n_1 将 $j=L-1$ 代入式(5)得到

$$\frac{\partial C}{\partial n_{L-1}}=0\Rightarrow n_1=(\frac{1}{\pi_{L-1}}2^{L(L-1)/2}n_L)^{1/L} \tag{14}$$

我们重写式(4)得到

$$\left.\begin{aligned}&\min_{n_L} L\left\{\frac{1}{\pi_{L-1}}2^{L(L-1)/2}n_L\right\}^{1/L}-\sum_{l=2}^{L}\left\{\frac{2^{l-1}}{\beta_l}\right\}\\ &\text{s.t.}\quad \begin{cases} v_L(\sum_{j=1}^{L}\beta_j n_j)^{-1}\leqslant \mu\\ 0<n_l\leqslant N_l\end{cases}\end{aligned}\right\} \tag{15}$$

由上述可以看出，只是一个关于 n_L 的函数. 最终，我们给出式(4)的求解结果：

$$n_l=\begin{cases}(2^{L(L-1)/2}(\prod_{i=1}^{L-1}\beta_i)^{-1}n_L)^{1/L}, & l=1\\ 2^{-l(l-1)/2}(\prod_{i=1}^{l-1}\beta_i)\beta_l^{-1}n_1^{l-1}(\beta_l n_1-2^{l-1}), & \\ \qquad l\in\{2,\cdots,L-1\} & \\ n_L, & (l=L)\end{cases} \tag{16}$$

总结上述的数学模型及最小化求解证明过程，层次分类模型是针对类别不平衡数据问题提出来的，且参数求解可以转换为带约束的优化求解问题. 约束条件的两个参数是误报率和第 l 层单分类器使用的特征上限. 其中，误报率可以人工限定，而特征数量可以根据数据集得知.

层次分类模型是在满足检测率、误报率的前提下使得分类代价达到最优，其完全二叉树结构可以有效地解决训练样本不平衡的问题，且可以将复杂的分类问题划分为难度降低的两个子问题. 针对其分类性能，层次分类模型的误报率可以人工设置，而检测率可以通过设定单分类器的检测率来满足条件，故层次分类模型可以提高分类性能.

3　实验结果及分析

3.1　评价标准

不平衡分类问题的常用指标是 ROC 曲线以及 ROC 曲线下面覆盖的面积 AUC. 首先给出混淆矩阵表，如表 1 所示.

表 1　混淆矩阵

Tab. 1　Confusion matrix

类别	预测正类	预测负类
实际正类	true positives(TP)	false negatives(FN)
实际负类	false positives(FP)	true negatives(TN)

根据混淆矩阵可以得到不平衡分类问题的常用评价指标：检测率 $P=\dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$、召回率 $R=\dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$、F-measure$=\dfrac{2\times P\times R}{P+R}$.

AUC 作为一种可靠的评价标准适用于不平衡分类和代价敏感问题. 对比正负样本得分，若正样本得分高于负样本，则积 1 分；若正样本得分等于负样本得分，则积 0.5 分；若正样本得分小于负样本得分，则积 0 分. 关于 AUC 计算问题，具体见参考文献[24].

3.2　数据集

论文选取 4 个典型的不平衡数据进行测试，数据来源于 UCI 数据集，具体参数如表 2 所示.

表 2　测试数据

Tab. 2　Test data

dataset	特征数目	正类	正样本/负样本	比率
musk2	166	class1	279/5581	20.003
isolet	617	class1	300/7447	24.823
onehr	72	class1	73/2463	33.74
spectf	44	class1	21/212	10.095

对于每个测试数据，给出正类编号，其余均为负类. 每组数据采用 5 折交叉验证，20 次运行统计平均结果. 我们对比了如下方法：

(Ⅰ) AdaBoost(简称 Ada)：采用 2.1 节的算法框架，下面的算法均以 AdaBoost 算法为基准；

(Ⅱ) UnderSampling[25] + AdaBoost(简称 Under)：在负样本集合 N 中随机抽取与正样本相同数量的子集 N′，用正样本集合 P 与负样本子集 N′训练 AdaBoost 分类器；

(Ⅲ) EasyEnsemble(简称 Easy)：在负样本集合 N 中独立抽取与正样本相同数量的子集 N′，用正样本集合 P 与负样本子集 N′训练 AdaBoost 分类器，最终判别函数为各个 AdaBoost 分类器的叠加，实验中设置层数为 4 层；

(Ⅳ) BalanceCascade(简称 Cascade)：在负样本集合 N 中独立抽取与正样本相同数量的子集 N′，用正样本集合 P 与负样本子集 N′训练 AdaBoost 分类器，控制每层 AdaBoost 分类器的阈值，使其分类误报率等于假设值，最终使得最后一层的正负样本规模相当，实验固定层数为 4 层；

(Ⅴ) HAdaBoost(简称 HAda)：实验设置层数为 3 层，表 3 给出 HAdaBoost 方法在 5 个测试数据上的误报率、平均特征、每层特征的数量 n_i 以及对应的参数值 β_i.

表 3　HAdaBoost 方法的参数

Tab. 3　The parameters of HAdaBoost

dataset	误报率	平均特征	n_1	n_2	n_3	β_1	β_2
musk2	0.05	37	16	29	60	0.416 4	0.276 5
isolet	0.005	33	18	32	60	0.393 4	0.221 6
nehr	0.07	14	9	10	20	0.816 6	0.314 7
spectf	0.04	15	7	8	15	0.653 0	0.590 3

考虑到每个数据集的复杂程度不同，其可分度

有差异,根据人工经验设置误报率达到最优分类性能.图 2 给出了 musk2 数据集上不同误报率对应的 F-measure 结果,最终的误报率依据最优分类结果设置选取.

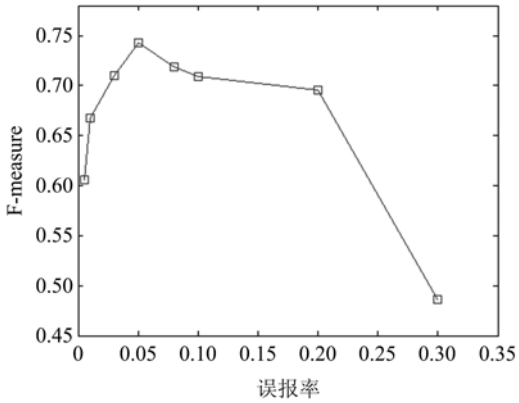

图 2 musk2 数据集上不同误报率的 F-measure 曲线图

Fig. 2 F-measure curve of different false positive rate on musk2 data

3.3 结果分析

表 4 和表 5 分别给出 Ada、Under、Easy、Cascade 和 HAda 方法在 4 个测试数据上的 AUC 和 F-measure 的统计结果,其中每个数据集的最优结果用粗体表示.

从表 4 和表 5 的统计结果可以看出,这 4 个测试数据的分类难度不同. musk2 和 isolet 分类相对容易,原因是其 AUC 统计结果较高(高于 0.9).而另外 2 个测试数据的难度较高,他们的 F-measure 结果均较低(低于 0.5).对于不同的测试数据,每种算法排名不同,如在 spectf 测试数据的统计结果依次为 HAda、Easy、Cascade、Under、Ada.表 4 和表 5 的共同特点是 HAda 方法的 AUC 和 F-measure 的结果均最优,且 F-measure 的优势较为明显,其次是 Cascade 和 Easy 方法,最差是 AdaBoost 和 Under 方法.

图 3 和图 4 分别给出每种方法在 5 个测试数据上的 AUC 曲线图和 F-measure 曲线.其中,横坐标为特征数目,纵坐标为统计结果. HAda 方法使用的是平均特征,对比其他四种方法,在使用相同数目的特征时,论文所提方法可以取得最优的结果,且 HAda

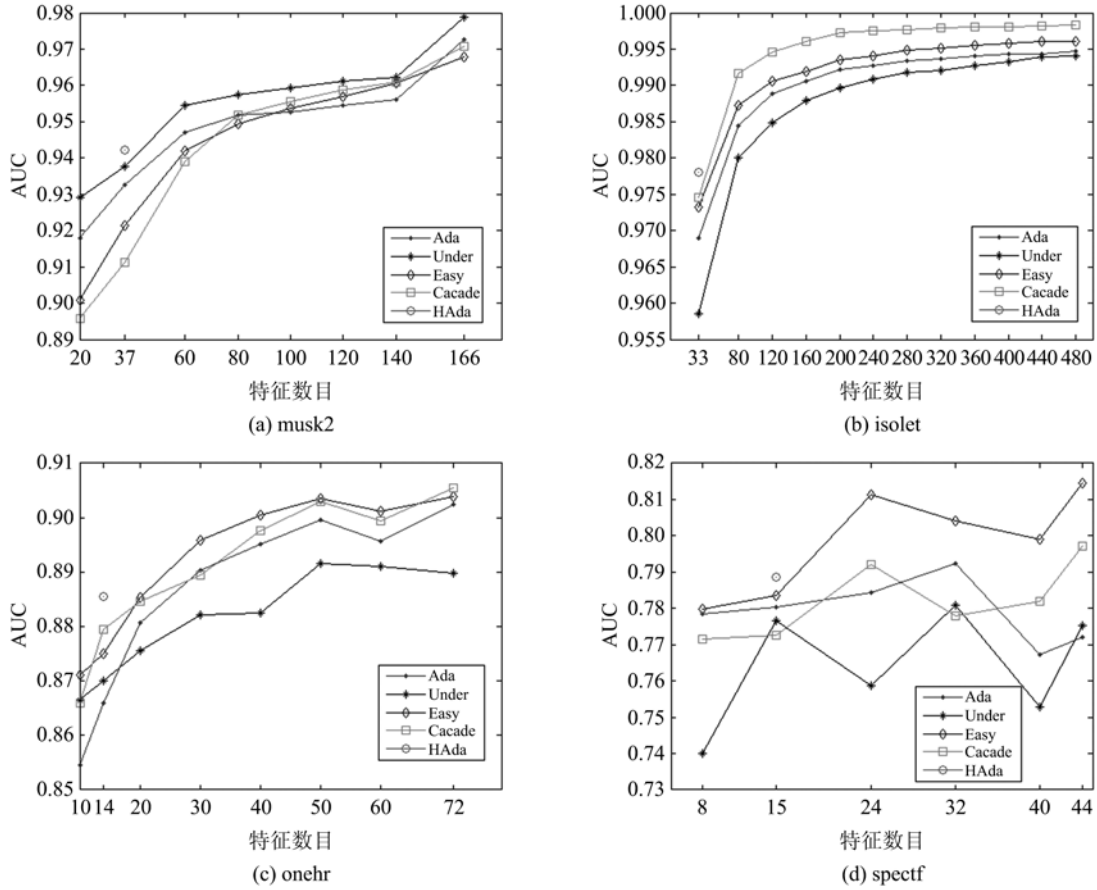

图 3 不同测试数据上 AUC 曲线图

Fig. 3 AUC curves on different algorithms

表 4　不同算法的 AUC 统计结果

Tab. 4　AUC statistical results of different algorithms

AUC	HAda	Ada	Under	Easy	Cascade
musk2	0.941 2±0.019 8	0.911 0±0.014 5	0.913 9±0.014 9	0.917 0±0.018 8	0.915 9±0.023 8
isolet	0.978 0±0.010 4	0.967 6±0.006 4	0.968 9±0.008 5	0.968 6±0.010 4	0.973 2±0.008 3
onehr	0.889 2±0.043 6	0.869 2±0.041 9	0.868 9±0.045 8	0.870 5±0.042 4	0.875 7±0.043 3
spectf	0.783 2±0.120 9	0.752 4±0.104 9	0.745 8±0.115 1	0.781 8±0.091 1	0.752 7±0.102 9

表 5　不同方法的 F-measure 统计结果

Tab. 5　F-measure statistical results of different algorithms

F-measure	HAda	Ada	Under	Easy	Cascade
musk2	0.743 0±0.040 8	0.647 7±0.045 6	0.561 1±0.052 9	0.546 6±0.052 6	0.650 0±0.049 1
isolet	0.820 4±0.036 8	0.608 2±0.045 6	0.637 6±0.046 1	0.574 6±0.047 3	0.676 1±0.047 6
onehr	0.405 0±0.077 5	0.339 5±0.069 4	0.334 5±0.066 8	0.338 0±0.074 3	0.377 3±0.075 8
spectf	0.457 5±0.119 3	0.408 7±0.101 3	0.410 0±0.117 1	0.445 7±0.110 1	0.411 6±0.099 8

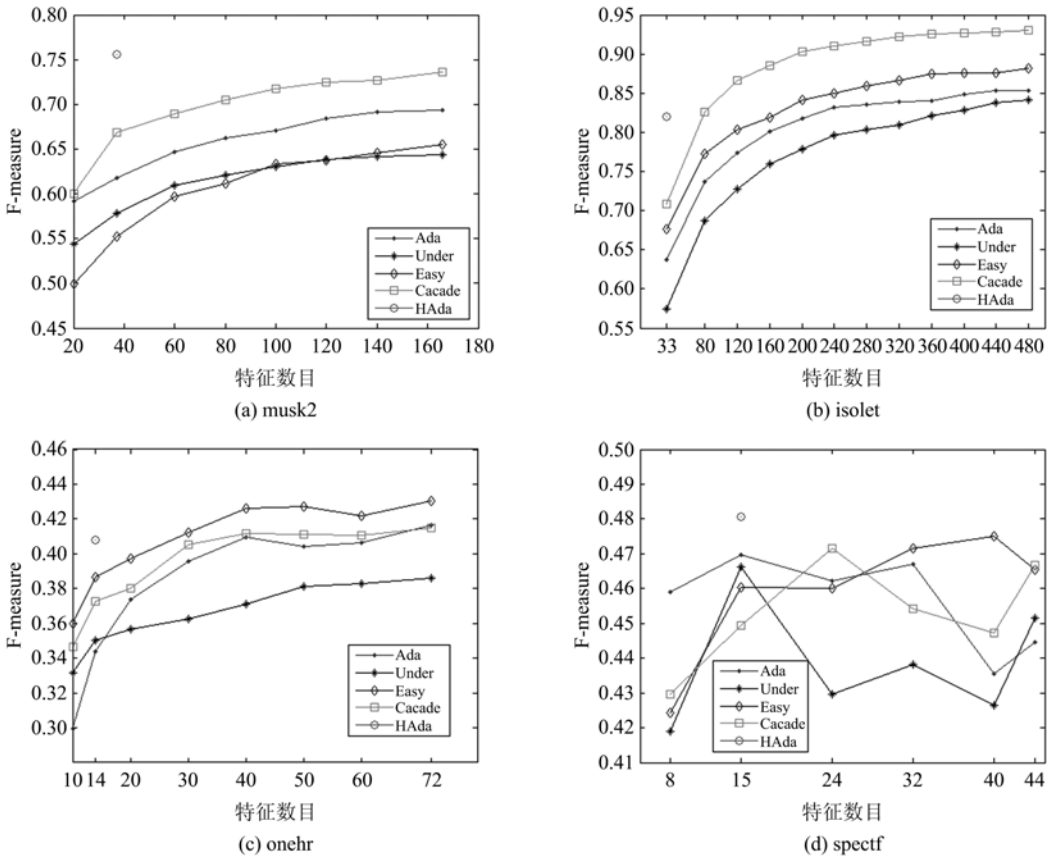

图 4　不同测试数据上 F-measure 曲线图

Fig. 4　F-measure curves on different test data

方法的 F-measure 结果优势较为明显. 另外,从 musk2 和 spectf 数据集的统计曲线可以看出,HAda 方法的 F-measure 值优于其他四种方法的最优结果.

4　结论

本文提出了一种基于类别不平衡数据的层次分类模型,采用层次分类树结构,实现模型参数可计算. 在 4 个类别不平衡测试数据上进行实验,结果表明,层次分类模型具有较优的分类结果. 论文考虑层次分类架构,并实现二叉层次树的求解,主要针对二分类问题. 对于不平衡分类的多分类问题,如果采用 k 叉层次树求解是未来的一个研究方向.

参考文献(References)

[1] Phua C, Alahakoon D, Lee V. Minority report in fraud detection: classification of skewed data[J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 50-59.

[2] Sun A X, Lim E P, Liu Y. On strategies for imbalanced text classification using SVM: A comparative study[J]. Decision Support Systems, 2009, 48(1): 191-201.

[3] Turney P D. Learning algorithms for key phrase extraction[J]. Information Retrieval, 2000, 2(4): 303-336.

[4] Burez J, van den Poel D. Handling class imbalance in customer churn prediction[J]. Expert Systems with Applications, 2009, 36(3): 4 626-4 636.

[5] Brekke C, Solberg A H S. Oil spill detection by satellite remote sensing[J]. Remote sensing of environment, 2005, 95(1): 1-13.

[6] Plant C, Böhm C, Tilg B, et al. Enhancing instance-based classification with local density: a new algorithm for classifying unbalanced biomedical data[J]. Bioinformatics, 2006, 22(8): 981-988.

[7] Branch J W, Giannella C, Szymanski B, et al. In-network outlier detection in wireless sensor networks[J]. Knowledge and information systems, 2013, 34(1): 23-54.

[8] Sahbi H, Geman D. A hierarchy of support vector machines for pattern detection[J]. Journal of Machine Learning Research, 2006, 7: 2 087-2 123.

[9] Blake C, Keogh E, Merz C J. UCI repository of machine learning databases[EB/OL]. http://www.ics.uci.edu/_mlearn/MLRepository.html.

[10] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[11] Han H, Wang W Y, Mao B H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning[C]//Advances in Intelligent Computing. Berlin Heidelberg, Germany: Springer, 2005: 878-887.

[12] Liu A, Ghosh J, Martin C E. Generative oversampling for mining imbalanced datasets[C]// Proceedings of International Conference on Data Mining. Las Vegas, USA: IEEE Press, 2007: 66-72.

[13] Batista G E, Prati R C, Monard M C. A study of the behavior of several methods for balancing machine learning training data[J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 20-29.

[14] Weiss G M, Provost F J. Learning when training data are costly: The effect of class distribution on tree induction[J]. Journal of Artificial Intelligence Research, 2003, 19: 315-354.

[15] Domingos P. MetaCost: A general method for making classifiers cost-sensitive[C]// Proceedings of the International Conference on Knowledge Discovery and Data Mining. San Diego, USA: ACM Press, 1999: 155-164.

[16] Chen C, Liaw A, Breiman L. Using random forest to learn imbalanced data[R]. TR666, Statistics Department, University of California at Berkeley, 2004.

[17] Chew H G, Bogner R E, Lim C C. Dual ν-support vector machine with error rate and training size biasing[C]// Proceedings of the 26th International Conference on Acoustics, Speech and Signal Processing. Salt Lake City, USA: IEEE Press, 2001, 2: 1 269-1 272.

[18] Raskutti B, Kowalczyk A. Extreme re-balancing for SVMs: A case study[J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 60-69.

[19] Juszczak P, Duin R P W. Uncertainty sampling methods for one-class classifiers[C]// Proceedings of International Conference on Machine Learning. Washington, USA: IEEE Press, 2003: 81-88.

[20] Zhou Z H, Liu X Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(1): 63-77.

[21] Liu X Y, Wu J X, Zhou Z H. Exploratory undersampling for class-imbalance learning[J]. IEEE Transactions on Systems, Man, and Cybernetics, 2009, 39(2): 539-550.

[22] Galar M, Fernandez A, Barrenechea E, et al. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches[J]. IEEE Transactions on Systems, Man, and Cybernetics, 2012, 42(4): 463-484.

[23] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[C]// Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. London, IEEE Press, 2001, 1: I-511-518.

[24] Liu X Y, Li Q Q, Zhou Z H. Learning imbalanced multi-class data with optimal dichotomy weights[C]// IEEE 13th International Conference on Data Mining. Omaha, USA: IEEE Press, 2013: 478-487.

[25] Drummond C, Holte R C. C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling[EB/OL]. http://www.site.uottawa.ca/~nat/Workshop2003/drummondc.pdf.