文章编号: 1001-0920 (2012) 12-1833-06

一种适用于多类不平衡数据集的模糊关联分类方法

霍纬纲, 高小霞

(中国民航大学 计算机科学与技术学院, 天津 300300)

摘 要: 提出一种适用于多类不平衡分布情形下的模糊关联分类方法, 该方法以最小化 AdaBoost.M1W 集成学习迭代过程中训练样本的加权分类错误率和子分类器中模糊关联分类规则数目及规则中所含模糊项的数目为遗传优化目标, 实现了 AdaBoost.M1W 和模糊关联分类建模过程的较好融合. 通过 5 个多类不平衡 UCI 标准数据集和现有的针对不平衡分类问题的数据预处理方法实验对比结果, 表明了所提出的方法能显著提高多类不平衡情形下的模糊关联分类模型的分类性能.

关键词: 模糊关联分类; 多类不平衡分类; 遗传算法; 集成学习; 数据挖掘

中图分类号: TP18 **文献标志码:** A

A fuzzy associative classification method for multi-class imbalanced datasets

HUO Wei-gang, GAO Xiao-xia

(College of Computer Science and Technology, Civil Aviation University of China, Tianjin 300300, China. Correspondent: HUO Wei-gang, E-mail: wghuo@cauc.edu.cn)

Abstract: A fuzzy associative classification method for multi-class imbalanced datasets is presented. The method implements a better combination of AdaBoost.M1W and the process of building fuzzy associative classification by the genetic optimization objective, which is minimization weighted error rate in the process of ensemble iterative learning and the number of fuzzy association rule and total fuzzy items in the weak fuzzy associative classifier. The experiments of comparing with existing data preprocessing approaches aiming at the imbalanced classification problem show that the proposed method can dramatically improve the classification performance of the fuzzy associative classifier for multi-class imbalanced datasets by five UCI multi-class imbalanced benchmark datasets.

Key words: fuzzy associative classification; multi-class imbalanced classification; genetic algorithm; ensemble learning; data mining

1 引 言

模糊关联分类是数据挖掘研究领域中重要的分类方法之一. 该方法所得的分类模型贴近人类的思维方式, 容易被人理解, 但其分类准确率容易受到模糊关联规则挖掘过程中的模糊支持度阈值的影响, 尤其在数据集类别分布不平衡情形下, 若模糊支持度阈值过高, 则针对出现频率较少类别的样本产生的规则较少, 进而降低了分类性能; 若模糊支持度阈值过低, 则会产生大量无用的规则使得分类模型对训练样本过拟合, 而且影响其可理解性. 对此, 文献 [1] 根据不同类别的样本在训练集中出现的频率采用不同的支持度阈值生成相应的分类规则, 但该方法中的基准支持度仍需主观指定. 文献 [2] 结合数据预处理的方法解决稀有样本产生的模糊规则少的问题, 并分析了数量属性划分模糊区间个数, 以及不同的模糊规则权重、模糊推理策略、t 模算子对基于模糊规则的分类系统在两类不平衡分布情形下的分类性能的影响. 文献 [3] 在此基础上应用参数化的 t 模算子进行模糊分类推理, 并通过遗传进化搜索 t 模算子中参数的较优值来提高不平衡分布下的模糊规则分类器的性能. 文

收稿日期: 2011-07-26; **修回日期:** 2011-09-28.

基金项目: 国家自然科学基金委员会与中国民用航空局联合基金项目(61079007, U1233113); 中国民航局科技计划项目(MHRD201005); 国家自然科学基金青年科学基金项目(61201414); 中央高校基本科研业务费专项资金项目(ZXH2012N001).

作者简介: 霍纬纲(1978−), 男, 讲师, 博士, 从事模糊关联规则挖掘、模糊分类等研究; 高小霞(1980−), 女, 硕士, 从事模糊分类的研究.

献 [4] 提出了一种基于集成学习算法 AdaBoost.M2 的模糊规则分类器, 但该方法要求分类器输出每一训练样本相对每一类别的概率值, 而模糊关联分类器有可能对某个训练样本无法识别分类, 即无法输出样本属于每一类别的概率值.

目前对于不平衡数据集的分类问题的研究主要有以下 3 个方面:

1) 数据预处理. 该方面的工作主要是通过采样的方式使不平衡的样本分布变得比较平衡, 从而提高分类器对稀有样本的识别率. 典型的过采样方法有: 随机过采样、SMOTE (synthetic minority oversaming technique); 下采样方法有: 随机下采样、Tomek-links、ENN (Wilson's Edited Nearest Neighbour Rule) 等; 这些采样方法的具体描述参见文献 [5].

2) 代价敏感学习. 该类方法赋予各个类别不同的错分代价, 在多数不平衡分类问题中, 稀有类是分类的重点, 因此错分稀有类的代价大于错分大类样本, 代价敏感学习的目标为最小化高代价的错分次数和总的错分代价, 以提高分类器在不平衡数据情形下对各个类别的分类准确率. 通常对于给定的数据集, 其真实的各个类别的错分代价依赖于特定的应用领域, 而且难以被估计.

3) 基于 AdaBoost(adaptive boosting)[6]的集成学习. AdaBoost 是一种基于迭代的集成学习方法, 在迭代过程中通过赋予训练样本的权重反映其对分类的重要性, 并由样本上的权值对训练集进行抽样. 在每次迭代学习过程中被错分的样本将赋予更大的权值, 使得下次迭代学习更加关注这些样本; 而对于不平衡数据集的分类问题, 被错分的样本往往是稀有类样本, 因此基于 AdaBoost 的集成学习可提高对稀有类别的识别率.

文献 [7] 对不平衡数据集学习的其他方面的研究工作作了很好的综述. 研究者们还将上述方法相互融合用于不平衡数据集的分类. 文献 [8] 为防止下采样技术对大类样本的信息丢失, 提出了 EasyEnsemble 和 BalanceCascade 算法. EasyEnsemble 对大类样本进行有放回的随机采样, 形成若干与稀有类样本数目相等的大类样本的子集; 然后每个大类样本子集与稀有类样本组合, 并应用集成学习方法 AdaBoost 训练分类器; 最后将所有大类子集形成的分类器进行组合. BalanceCascade 算法与 EasyEnsemble 原理基本相同, 区别点在于: 在每次形成大类样本的子集时, 已经被正确分类的大类样本将被从总的大类样本集中去除. 文献 [9] 将 SMOTE 与 AdaBoost 相结合, 在每次迭代学习过程中由 SMOTE 方法产生新的稀有类样本, 使得下次迭代学习稀有类样本受到更多的关注. 文献 [10] 提出的 DataBoost-IM 算法结合了样本生成技术和 AdaBoost 集成学习方法, 该方法根据大类和稀有类样本数目的比率和迭代学习过程中难以被识别的样本生成新的大类和稀有类样本, 并将它们加入下次迭代过程, 使得分类器对所有类别的样本都具有较高的识别率. 文献 [11] 将训练样本的错分代价引入 AdaBoost 的权值更新规则中, 使得稀有类样本获得更高的权重, 通过大类和稀有类样本错分代价的不同比值下分类器的分类性能确定不同类别的最佳错分代价. 上述文献 [8,10-11] 的方法都是用于两类不平衡数据的分类问题, 文献 [9] 的方法不能用于模糊关联分类模型, 因为在集成迭代学习过程中产生新的稀有类样本会改变数量属性模糊区间的划分范围, 使得分类器中模糊关联分类规则的语义很难确定.

针对多类不平衡分类问题, 文献 [12] 提出了基于代价敏感的集成学习方法 AdaC2.M1, 该方法采用遗传算法搜索各个类别的错分代价, 其训练过程较复杂而且求得的错分代价只是代价矩阵每一列之和. 文献 [13] 基于 3 类特殊的代价矩阵通过训练代价敏感的神经网络实验表明, 代价敏感学习方法很难较好地适用于多类不平衡分类问题. AdaBoost.M1 是用于多类分类问题的经典集成学习方法, 但其要求学习过程中每个基本分类器的分类错误率均小于 0.5, 而模糊关联分类器在多类不平衡数据情形下很难满足该条件. 为此, 本文将 AdaBoost.M1W[14]集成学习方法用于模糊关联分类器在多类不平衡情形下的提升, 并且通过遗传算法较好地实现了模糊关联分类模型的建模过程与 AdaBoost.M1W 的融合. 实验结果表明了本文方法的有效性.

2 基于遗传算法的子模糊关联分类器设计

2.1 编码方式与进化算子

本文采用文献 [15] 的算法生成模糊频繁项及模糊关联分类规则 (FACR). 遗传进化种群的个体采用二进制编码方式, 设参与进化的有 N 个 FACR 规则, 则群体中的每个个体可表示为长度为 N 的二进制串 $s_1, s_2, \cdots, s_N$, 每个 $s_i (1 \leqslant i \leqslant N)$ 取值 0 或 1. 0 表示该基因位对应的 FACR 不在构成模糊关联分类模型的规则集中, 1 表示对应的 FACR 包含在模糊关联分类模型中.

种群中个体间采用 HUX (half uniform crossover) 交叉算子, 具体实现步骤如下: 计算参与交叉运算的两个个体间的海明距离, 然后将两个个体不同取值的基因位上的值相互交换生成两个新的个体, 交换的位数为个体间海明距离的一半, 交换的位置在两个个体不同取值的若干个位置上随机选取.

为尽量减少分类模型中FACR规则的数量, 本文采用了有偏向的变异操作[16], 即个体上的每个基因位0至1和1至0的变异按不同的概率进行, 例如1至0的变异概率取0.1, 0至1变异概率取0.001. 当某个个体的基因位上的值全部变为0时, 该个体需要执行重新初始化操作, 即进化个体的每个基因位被重新按概率0.5随机赋值为0或1.

2.2 个体适应度函数设计

因为AdaBoost类的集成学习算法的核心在于每次迭代过程中应尽可能地减小子分类器的加权错误率, 同时为了保证子模糊关联分类器的解释性, 需要减少分类模型中包含的分类规则和模糊项数目, 所以本文进化群体中个体S的适应度函数设计如下:

$$f(S)=\frac{1}{\sum\limits_{h_S(x_i)\neq y_i} D_t(i)}\times\frac{1}{\mathrm{Num_{Rule}}+\mathrm{Num_{FItem}}}. \quad (1)$$

其中: $\mathrm{Num_{Rule}}$和$\mathrm{Num_{FItem}}$分别为个体S表示的分类模型中规则数目和所有规则前件包含的模糊项的数目和, $\sum\limits_{h_S(x_i)\neq y_i} D_t(i)$为个体$S$表示的分类模型的加权错误率. 通过式(1)表示的个体适应度函数来实现子模糊关联分类器的构建过程与下文中所使用的AdaBoost.M1W算法的融合.

2.3 构建子模糊关联分类器的算法描述

为使模糊关联分类模型中能有较多的模糊关联分类规则描述稀有类样本, 本文以最少类别样本在训练集中出现频率的0.5倍为最小模糊支持度挖掘模糊频繁项, 但对于多类不平衡数据集, 这将同时生成大量描述大类样本的模糊关联分类规则, 使得初始规则集过于庞大. 为缩小遗传算法的搜索空间, 本文通过两遍扫描规则集选择模糊相关度大且前件中包含模糊项较少的分类规则, 该方法的具体步骤见文献[17]. 用简单遗传算法构建子模糊关联分类模型, 采用文献[18]中的模糊分类推理模型计算子模糊关联分类模型的加权错误率, 算法的具体流程如下:

1) 对精简后模糊关联分类规则集二进制编码, 初始进化种群;

2) 由式(1)计算种群中每个个体适应度值, 并对种群实施比例选择算子, 对得到的新种群实施HUX交叉算子和有偏向的变异算子;

3) 若满足迭代终止条件, 则输出最优个体表示的子模糊关联规则分类器, 否则转2).

3 基于AdaBoost.M1W的模糊关联分类器

设有数据集$D=[XY]$, 其中$X=[x_{k,l}]_{M\times n}$为在数量属性上的M个取值, $Y=[y_k]_{M\times 1}$为在类别属性上的取值. 算法具体描述如下.

Step 1: 输入为训练样本集Traindata$=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N),\ x_i\in X, y_i\in Y\}$, $Y=\{1,2,\cdots,|Y|\}$为训练样本类别集合, N为训练样本数目, $|Y|$为类别数目, T为集成学习迭代次数.

Step 2: 用模糊聚类FCM算法对训练样本模糊预处理, 并初始化训练样本权值$D_1(i)=1/N$.

Step 3: For $t=1,2,\cdots,T$.

Step 3.1: 由训练样本权值分布D_t对训练数据集采样, 采用2.3节描述的遗传算法流程构建子模糊关联分类器$h_t: X\to Y$;

Step 3.2: 计算由遗传进化所得子模糊关联分类器$h_t: X\to Y$的加权错误率$\varepsilon_t=\sum\limits_{h_t(x_i)\neq y_i} D_t(i)$;

Step 3.3: If $\varepsilon_t\geqslant 1-1/|Y|$ Then 退出循环, 算法终止;

Step 3.4: 计算分类器$h_t: X\to Y$的权值$\alpha_t=\mathrm{Ln}((|Y|-1)(1-\varepsilon_t)/\varepsilon_t)$;

Step 3.5: 更新训练样本的权值

$$D_{t+1}(i)=\frac{D_t(i)}{Z_t}\times\begin{cases}\mathrm{e}^{-\alpha_t}, & h_t(x_i)=y_i;\\ \mathrm{e}^{\alpha_t}, & h_t(x_i)\neq y_i.\end{cases}$$

其中Z_t为归一化因子.

End For

Step 4: 输出模糊关联集成分类器

$$H(x)=\arg\max_{y\in Y}\left(\sum_{t=1}^{T}\alpha_t\times\begin{cases}1, & h_t(x)=y;\\ 0, & h_t(x)\neq y\end{cases}\right).$$

4 实验结果

实验中采用UCI机器学习数据集(http://www.ics.uci.edu/~mlearn/)中的5个多类不平衡数据集, 各数据集的详细信息如表1所示. 每个数据集的数量属性由FCM聚类算法模糊预处理, 聚类类别数目均为3. 为使训练集和测试集含有相应比例的所有类别的样本, 采用分层抽样的方法将所有数据集划分为3个部分, 随机选择2个部分作为训练集, 剩余部分作为测试集. 构建模糊关联分类模型的遗传算法各参数设置如表2所示, 实验环境为WinXP, CPU 3.0 GHz, 2.0 GB内存, Visual C++6.0.

表 1　实验中所采用的数据集描述

数据集名称	样本数目	数量属性数目	类别数目	类别分布情况
Glass	214	9	6	70/17/76/13/9/29
New_thyroid	215	5	3	150/35/30
Ecoli	332	7	6	143/77/52/35/20/5
Yeast	1 428	6	7	463/429/244/163/51/44/37
Abalone	4 024	7	14	57/115/259/391/568/689/634/487/267/203/126/103/67/58

表 2 遗传算法参数设置

参数	值
进化种群大小	100
进化迭代次数	50
交叉概率	0.9
变异概率(1 → 0)	0.1
变异概率(0 → 1)	0.001

本文采用分类错误率和查全率的几何平均值两项指标评估分类器的性能, 具体定义如下: 设样本集的总类别数目为 k, $n_{ij}(1 \leqslant i, j \leqslant k)$ 表示分类器将属于 i 类的样本识别为 j 类的样本数目, 则分类器的分类错误率定义为

$$\text{Errorrate} = 1 - \sum_{i=1}^{k} n_{ii} \Big/ \sum_{i,j=1}^{k} n_{ij},$$

类别 i 的查全率定义为

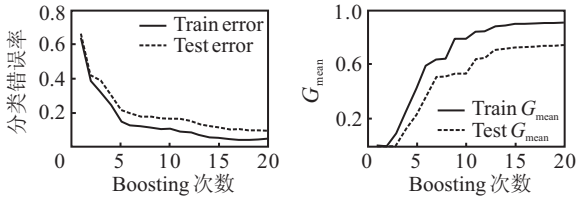
$$\text{Recall}_i = n_{ii} \Big/ \sum_{j=1}^{k} n_{ij},$$

分类器查全率几何均值定义为

$$G_{\text{mean}} = \Big(\prod_{i=1}^{k} \text{Recall}_i \Big)^{1/k}.$$

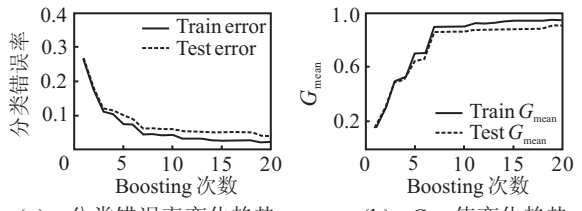
该指标可以反映分类器对所有类别样本识别能力的平均性能.

图 1~图 5 为本文方法在表 1 中各个数据集上的实验结果, 在每个图中 (a) 为训练集和测试集上分类错误率随 Boosting 次数的变化, (b) 为训练集和测试集的 G_{mean} 值随 Boosting 次数的变化, 图中的每个值均为 5 次分层抽样实验结果的均值. 从这些图中不难看出: 1) 随着集成迭代次数的增加, 训练集和测试集上的分类错误率逐步降低, 而 G_{mean} 值逐步升高; 2) 类别和样本数目较多的数据集 Yeast, Abalone 需要较多的迭代次数才能得到较好的分类效果; 3) 除 New_thyroid 数据集外, 其余数据集在集成迭代学习的初期



(a) 分类错误率变化趋势 (b) G_{mean}值变化趋势

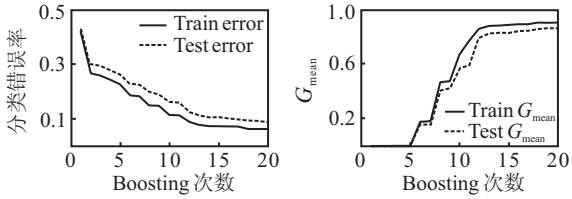
图 1 Glass 训练集和测试集上分类错误率和 G_{mean} 值随 Boosting 次数的变化图



(a) 分类错误率变化趋势 (b) G_{mean}值变化趋势

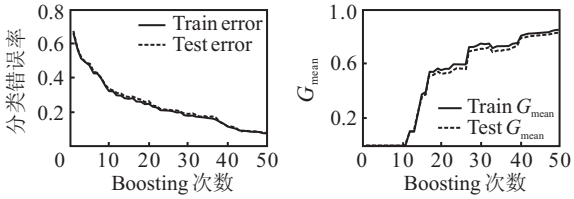
图 2 New_thyroid 训练集和测试集上分类错误率和 G_{mean} 值随 Boosting 次数的变化图

期 G_{mean} 值均出现 0, 这是因为分类器对这些数据集中的某些类别无法识别, 使得这些类别的查全率变为 0, 但随着迭代次数的增加, G_{mean} 值能逐步升高而且达到较高的值. 因此, 本文设计的方法在提高分类器准确率的同时还能较好地识别数据集中的所有类别的样本, 适用于多类不平衡样本的分类.



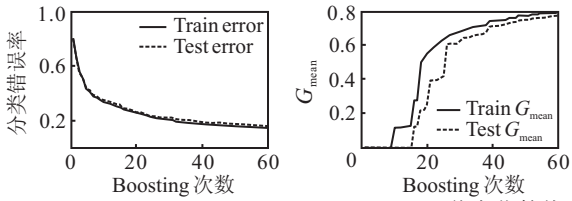
(a) 分类错误率变化趋势 (b) G_{mean}值变化趋势

图 3 Ecoli 训练集和测试集上分类错误率和 G_{mean} 值随 Boosting 次数的变化图



(a) 分类错误率变化趋势 (b) G_{mean}值变化趋势

图 4 Yeast 训练集和测试集上分类错误率和 G_{mean} 值随 Boosting 次数的变化图



(a) 分类错误率变化趋势 (b) G_{mean}值变化趋势

图 5 Abalone 训练集和测试集上分类错误率和 G_{mean} 值随 Boosting 次数的变化图

将处理不平衡数据分类的数据预处理采样的方法与本文方法进行了分类性能比较, 实验中首先对原始数据集进行采样预处理, 然后再进行模糊预处理, 并应用文献 [17] 的方法训练模糊关联分类器. 定义 Classavg = n/k, n 为原始数据集样本数目, k 为类别数目. 对每一数据集中样本数目小于 Classavg 的类别 $i(1 \leqslant i \leqslant k)$ 进行过采样, 由类别 i 的每个样本生成的新样本数目为 $\lfloor \text{Classavg}/n_i \rfloor$, n_i 为原始数据集中类别 i 的样本数目, 大于 Classavg 的类别进行下采样. 实验结果如表 3 所示, 其中 Original 表示直接对原始数据集进行训练, 无采样预处理; Smote+tomelink 表示对数据集先用 Smote 过采样, 然后用 tomelink 下采样; Smote+enn 表示过采样后用 enn 算法下采样; Randomoversample 表示对数据集随机过采样; 本文的方法简记为 GAdaBoost.M1W. 表 3 中的实验结果为 5 次分层抽样的均值, 不难看出, 过、下采样的方法在

表 3　本文方法与数据预处理方法分类性能的比较

数据集	算法	G_{mean}		分类错误率	
		训练样本	测试样本	训练样本	测试样本
Glass	Original	0	0	0.38±0.04	0.42±0.09
	Smote+tomelink	0.62±0.05	0.55±0.04	0.33±0.01	0.40±0.05
	Smote+enn	0.63±0.03	0.62±0.03	0.32±0.01	0.34±0.01
	Randomoversample	0.59±0.05	0.48±0.06	0.37±0.02	0.44±0.03
	GAdaBoost.M1W	**0.90±0.05**	**0.74±0.12**	**0.04±0.02**	**0.09±0.05**
Ecoli	Original	0.68±0.08	0.59±0.11	0.2±0.02	0.25±0.01
	Smote+tomelink	0.76±0.01	0.74±0.02	0.21±0.01	0.22±0.01
	Smote+enn	0.74±0.02	0.74±0.01	0.21±0.01	0.23±0.01
	Randomoversample	0.73±0.02	0.73±0.05	0.24±0.05	0.22±0.02
	GAdaBoost.M1W	**0.90±0.05**	**0.87±0.03**	**0.06±0.03**	**0.09±0.03**
Abalone	Original	0	0	0.72±0.006	0.74±0.01
	Smote+tomelink	0	0	0.73±0.004	0.74±0.006
	Smote+enn	0	0	0.75±0.002	0.76±0.01
	Randomoversample	0	0	0.75±0.003	0.76±0.007
	GAdaBoost.M1W	**0.79±0.03**	**0.77±0.03**	**0.15±0.01**	**0.16±0.02**
New_thyroid	Original	0.79±0.02	0.73±0.05	0.09±0.006	0.11±0.01
	Smote+tomelink	0.90±0.01	0.85±0.02	0.08±0.006	0.1±0.01
	Smote+enn	0.94±0.01	0.89±0.03	0.05±0.01	0.09±0.02
	Randomoversample	0.71±0.02	0.80±0.03	0.22±0.008	0.18±0.02
	GAdaBoost.M1W	**0.96±0.04**	**0.91±0.05**	**0.02±0.01**	**0.04±0.02**
Yeast	Original	0.51±0.05	0.46±0.03	0.43±0.01	0.46±0.02
	Smote+tomelink	0.53±0.03	0.50±0.02	0.47±0.02	0.49±0.01
	Smote+enn	0.53±0.03	0.51±0.02	0.46±0.02	0.48±0.001
	Randomoversample	0.53±0.01	0.50±0.01	0.47±0.03	0.51±0.02
	GAdaBoost.M1W	**0.85±0.08**	**0.82±0.07**	**0.07±0.03**	**0.08±0.03**

Abalone 上对训练集和测试集的 G_{mean} 值均为 0, 相对原始数据集无提高, 而在其他数据集上过、下采样方法对 G_{mean} 值有明显提高, 但同时提高了数据集 Ecoli, Yeast, Page-block 的分类错误率, 这可能是因为过采样使得总样本数目增加引起的. 由此可见, 过、下采样方法处理多类不平衡数据集的分类具有一定的局限性, 而本文方法相比过、下采样方法在实验中所用的 5 个数据集的 G_{mean} 和分类错误率上均有较好的实验结果.

5　结　论

本文提出了一种适用于多类不平衡数据的模糊关联分类方法. 该方法结合了 AdaBoost.M1W 集成学习算法和遗传构建模糊关联分类模型的过程, 通过在遗传算法的适应度函数中引入加权分类错误率和 AdaBoost.M1W 的自适应采样过程, 实现了对大类样本和稀有类样本的同时识别. 实验结果表明了本文方法的有效性和可行性. 下一步将研究如何用更少的集成学习迭代次数得到较好的多类不平衡分类性能, 通过将代价敏感学习方法与本文方法相融合, 研究不同类型的代价矩阵对本文方法分类性能的影响.

参考文献(References)

[1] Bing Liu, Yiming Ma, Ching Kian Wong. Improving an association rule based classifier[C]. Proc of the 4th European Conf on Principles of Data Mining and Knowledge Discovery. Lyon, 2000: 504-509.

[2] Alberto Fernández, Salvador García, María José del Jesusb, et al. A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets[J]. Fuzzy Sets and Systems, 2008, 159(18): 2378-2398.

[3] Alberto Fernandez, Maria José del Jesus, Francisco Herrera. On the influence of an adaptive inference system in fuzzy rule based classification systems for imbalanced data-sets[J]. Expert Systems with Applications, 2009, 36(6): 9805-9812.

[4] 方敏, 王宝树. 基于 AdaBoost 的改进模糊分类规则集成学习[J]. 电子与信息学报, 2005, 27(5): 835-837.
(Fang M, Wang B S. Advance ensemble learning of fuzzy classification rules based on AdaBoost[J]. J of Electronics & Information Technology, 2005, 27(5): 835-837.)

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

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

[7] Haibo He, Edwardo. A garcia learning from imbalanced

Data[J]. IEEE Trans on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.

[8] Xu-Ying Liu, Jianxin Wu, Zhi-Hua Zhou. Exploratory underSampling for class-imbalance learning[J]. IEEE Trans on Systems, Man, and Cybernetics, Part B: Cybernetics, 2009, 39(2): 539-549.

[9] Chawla N V, Lazarevic A, Hall L O, et al. SMOTEBoost: Improving prediction of the minority class in boosting[C]. Proc of the 7th European Conf on Principles and Practice of Knowledge Discovery in Databases. Dubrovnik, 2003: 107-119.

[10] Guo H, Viktor H L. Learning from imbalanced data sets with boosting and data generation: The databoost-IM approach[J]. SIGKDD Explorations, 2004, 6(1): 30-39.

[11] Sun Y, Kamel M S, Wong A K C, et al. Cost-sensitive boosting for classification of imbalanced data[J]. Pattern Recognition, 2007, 40(12): 3358-3378.

[12] Sun Y, Kamel M S, Wang Y. Boosting for learning multiple classes with imbalanced class distribution[C]. Proc of the 6th Int Conf on Data Mining. Hong Kong, 2006: 592-602.

[13] Zhi-Hua Zhou, Xu-Ying Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem[J]. IEEE Trans on Knowledge and Data Engineering, 2006, 18(1): 63-77.

[14] Günter Eibl, Karl Peter Pfeiffer. How to make AdaBoost.M1 work for weak base classifiers by changing only one line of the code[C]. Proc of the 13th European Conf on Machine Learning. Helsinlci, 2002: 72-83.

[15] 霍纬纲, 邵秀丽. 基于TD-FP-Growth的模糊关联规则挖掘算法[J]. 控制与决策, 2009, 24(10): 1504-1508.
(Huo W G, Shao X L. Algorithm for mining fuzzy association rule based on TD-FP-growth[J]. Control and Descision, 2009, 24(10): 1504-1508.)

[16] Hisao Ishibuchi, Takashi Yamamoto. Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining[J]. Fuzzy Sets and Systems, 2004, 141(1): 59-88.

[17] Ferenc Peter Pach, Attila Gyenesei, Janos Abonyi. Compact fuzzy association rule-based classifier[J]. Expert Systems with Application, 2008, 34(4): 2406-2416.

[18] 霍纬纲, 邵秀丽. 一种基于多目标进化算法的模糊关联分类方法[J]. 计算机研究与发展, 2011, 48(4): 567-575.
(Huo W G, Shao X L. A fuzzy associative classification method based on multi-objective evolutionary algorithm[J]. J of Computer Research and Development, 2011, 48(4): 567-575.)

(上接第1832页)

[7] Carlsson C, Fullér R, Heikkil M, et al. A fuzzy approach to R and D project portfolio selection[J]. Int J of Approximate Reasoning, 2007, 44(2): 93-105.

[8] Liao S H, Ho S H. Investment project valuation based on a fuzzy binomial approach[J]. Information Sciences, 2010, 180(11): 2124-2133.

[9] Zhu D M, Zhang T, Chen D L. A new fuzzy pricing approach to real option[J]. J of Northeastern University: Natural Science, 2008, 29(11): 1544-1547.

[10] Xia Y Q, Chen J F. Fuzzy optimization of real options valuation for multi-phase R and D project[J]. J of Shanghai Jiaotong University, 2009, 43(4): 583-586.

[11] Mun J. Real options analysis: Tools and techniques for valuing strategic investments and decisions[M]. New York: J Wiley and Sons, 2002: 223-229.

[12] Zadel L A. Fuzzy sets[J]. Information and Control, 1965, 8(3): 338-353.

[13] Chielana F, Herrera F, Herrera-Vied-Ma E. Integrating three representation models in fuzzy multipurpose decision making based in fuzzy preference relations[J]. Fuzzy Sets and Systems, 1998, 97(1): 177-194.

[14] Bordogna G, Fedrizzi M, PASI G. A linguistic modeling of consensus in group decision making based on OWA operations[J]. IEEE Trans on Systems, Man and Cybernetics, Part A: Systems and Humans, 1997, 27(1): 132-142.

[15] 彭祖赠, 孙韫玉. 模糊数学及其应用[M]. 武汉: 武汉大学出版社, 2007: 177-179.
(Peng Z Z, Sun Y Y. Fuzzy mathematics and application[M]. Wuhan: Wuhan University Press, 2007: 177-179.)

[16] Cook W D. Distance-based and ad hoc consensus models in ordinal preference ranking[J]. European J of Operational Research, 2006, 172(2): 369-385.

[17] Ramanathan R, Ganesh L S. Group preference aggregation methods employed in AHP: An evaluation and an intrinsic process for deriving members' weightages[J]. European J of Operational Research, 1994, 79(2): 249-265.