文章编号: 1001-0920 (2012) 01-0104-05

一种新的模糊规则权重方法的非平衡数据分类问题的研究

陈 刚, 冯 丹

(大连海事大学 数学系，辽宁 大连 116026)

摘 要: 针对传统分类算法在处理非平衡数据集所出现的少数类分类准确率较低的问题, 通过引入加权系数和样本分布函数给出了一种新的模糊规则权重的计算方法. 该方法加强了类间的对比度和差异性, 削弱了类内差距. 将该权重方法与 Chi et al 规则生成算法和模糊分类推理模型结合形成新的分类算法, 对具有不同非平衡度的 UCI 数据集进行 Matlab 对比研究, 所得结果验证了该算法的可靠性与有效性.

关键词: 非平衡数据；数据分类；模糊规则权重；数据预处理

中图分类号: TP273 **文献标识码:** A

Research on a new method for fuzzy rule weights in imbalanced data classification problem

CHEN Gang, *FENG Dan*

(Department of Mathematics，Dalian Maritime University，Dalian 116026，China．Correspondent：CHEN Gang, E-mail：chengang@dlmu.edu.cn)

Abstract: For the problem that the traditional classification methods often tend to the majority class and lead a lower classification accuracy to the minority class in imbalanced data, a new calculation method of fuzzy rule weights is proposed. This algorithm not only keeps the pattern matching degree within class in uniform distribution, but also enhances the contrast of inter-class. Then a classification algorithm is designed, which includes the new calculation method of fuzzy rule weights, Chi et al algorithm and fuzzy reasoning model. Finally numerical simulation about the imbalanced data of UCI data sets shows the reliability of the classification algorithm.

Key words: imbalanced data；data classification；fuzzy rules weights；data processing

1 引 言

在现实世界的分类问题中广泛存在着非平衡数据, 即数据中的一类样本在数量上远多于另一类, 而其中少数类样本通常具有重要的作用和价值, 是人们主要关心的对象. 传统的分类算法在处理非平衡数据分类问题时会倾向于多数类, 从而导致少数类的分类精度较低, 因而研究非平衡数据集的分类方法具有重要价值[1-2]. 相关研究[3-5]主要围绕以下 3 个方面展开: 1) 改变数据的分布 (数据层面); 2) 设计新的分类方法 (算法层面); 3) 设计新的分类器性能评价准则 (判别准则).

近年来, 模糊数学由于具有处理人类语言的特殊能力, 其在非平衡数据分类问题中的应用受到越来越多学者的关注. 如何构建初始规则集是设计模糊分类系统的关键. 国内外学者为此进行了深入而广泛的研究: 将神经网络应用于模糊分类规则集的产生, 通过聚类算法生成模糊规则, 群体智能和遗传算法对模糊分类规则集的优化, 将规则编码为粒子; 应用粒子群优化算法进行分类规则的提取, 模糊分类规则的权值启发算法等[6-8]. 其中, 基于模糊规则的分类系统 (FRBCs)[9]处理非平衡数据分类问题具有很好的分类效果[10], 主要有两个方面: 1) 构建具有良好性能的 FRBCs 来处理非平衡数据; 2) 预处理非平衡数据.

目前, 关于模糊规则的分类系统的模糊规则权重的计算方法有 3 种: 第 1 种[11]仅局限于一条规则的各类样本分布的对比, 不足以体现学习空间的全局分

收稿日期: 2010-09-16；**修回日期:** 2010-12-28.

基金项目: 国家自然科学基金项目(60875032/F030504).

作者简介: 陈刚(1964−), 男, 教授, 从事模糊推理、模糊系统优化等研究；冯丹(1984−), 女, 硕士生, 从事数据处理与信息提取的研究.

布; 第 2 种[12]体现了样本的区域数量分布, 但由于只考虑了同类的样本匹配度, 对于非平衡数据计算的结果可能会偏向于多数类; 第 3 种[13]兼顾了最大值归属类的样本匹配度和其他类别样本匹配度, 更加突出了样本的全局分类, 然而该算法未能体现样本的区域数量分布, 对于分类准确率的计算会产生一定的影响. 对此, 本文提出一种新的模糊规则权重的计算方法, 该方法是通过引入加权系数和样本分布函数得到的. 它既在计算过程中使得样本类内匹配度得到均匀分布, 又在判断类别时加强了类间的对比度, 起到了削弱类内差距和增强类间差异的作用. 然后构建基于模糊规则的分类算法, 该算法由 Chi et al[14]规则生成算法、通用分类推理模型[15]和改进的模糊规则权重的计算方法 3 个部分组成. 最后, 使用非平衡数据集对该分类算法进行验证, 表明了算法的有效性和改进规则权重计算方法的合理性.

2　模糊分类的相关概念

建立分类模型的目的是计算出待分类数据的分类准则, 进行模糊推理. 假设分类输入变量 $M = c_1, c_2, \cdots, c_j$, 待分类数据或称训练集, 是由不同的数据库记录组成的. 每一条记录包含若干个属性, 组成一个特征向量. 训练集的每条记录都与一个特定的类标签相对应. 本文使用的模糊规则形式如下:

$$\begin{aligned}\text{Rule } R_j: \quad & \text{If } x_1 \text{ is } A_{j1} \text{ and} \cdots \text{and } x_n \text{ is } x_{jn},\\ & \text{Then Class is } c_j.\end{aligned}$$

其中: R_j 为第 j 个规则的标号, $x = (x_1, \cdots, x_n)$ 为 n 维样本, $x_i (i = 1, 2, \cdots, n)$ 为样本的第 i 个属性, A_{ji} 为第 j 条规则中第 i 个属性上的前件模糊集, c_j 为第 j 条规则的结论对应的样本类别, r_j 为第 j 条规则的权重. 任意类型的隶属函数(三角、梯形、钟形)都能应用于前件模糊集, 这里使用对称三角模糊集 A_{ij_i} 的隶属函数, 其形式如下:

$$\begin{aligned}u_{ij} &= \max\left(1 - |x - a^{K_{ij_i}}|/b^{k_i}, 0\right),\\ & j_i = 1, 2, \cdots, K_i.\end{aligned} \tag{1}$$

其中

$$a_{j_i}^{K_i} = (j_i - 1)/(K_i - 1),\ j_i = 1, 2, \cdots, K_i;$$
$$b^{K_i} = 1/(K_i - 1).$$

定义 1[13]　训练样本空间中分类 h 的所有样本在第 j 条规则上的匹配度之和称为分类 h 关于规则 j 的匹配度, 记作

$$\begin{aligned}\beta_{\text{Class } h}(R_j) &= \sum_{x_p \in \text{ Class } h} \mu_j(x_p),\\ & h = 1, 2, \cdots, M.\end{aligned} \tag{2}$$

其中 $\beta_{\text{Class } h}(R_j)$ 为属于 h 类的训练样本在模糊规则上的匹配度之和.

定义 2　分类 h 关于规则 j 的匹配度在各分类关于规则 j 的匹配度上进行归一化得到分类匹配度[19], 记作

$$\eta_j^h = \frac{\beta_{\text{Class } h}(R_j)}{\sum_{i=1}^{M} \beta_{\text{Class } h}(R_j)}. \tag{3}$$

定义 3　分类 h 关于规则 j 的匹配度在分类 h 关于各规则的匹配度上进行归一化得 ρ_j^h, 再由各分类样本的数量比例对 ρ_j^h 进行赋权得到规则相对匹配度[12], 记作 $\xi_j^h = m_{c_j}\rho_j^h/m$. 其中: $m_{c_j} (c_j = 1, 2, \cdots, M)$ 为各分类的样本数, 且

$$m = \sum_{c_j=1}^{M} m_{c_j},\ \rho_j^h = \beta_{\text{Class } h}(R_j) \Big/ \sum_{i=1}^{N} \beta_{\text{Class } h}(R_i).$$

3　规则权重的改进算法

基于文献 [13] 给出一种改进的权重计算方法, 既兼顾了最大值归属类和其他类别的样本匹配度, 又体现了样本的全局密度分布, 且突出了样本的全局分布, 对不同数据集都能得到较高的分类准确率. 下面给出改进的规则权重方法的计算步骤:

Step 1: 计算加权规则匹配度. 在规则匹配度中引入加权指数 α[8], 从数学上看, 它控制着各样本在模糊类间的分享程度, 加权规则匹配度可表示为

$$W_j(x_p) = [\mu_j(x_p)]^\alpha = [T(\mu_j(x_{p1}), \cdots, \mu_j(x_{pn})]^\alpha, \tag{4}$$

其中 α 为正常数, 通常在 $\alpha = 0, 0 < \alpha < 1$ 和 $\alpha > 1$ 三种情况下进行讨论. 对于常数 α 的作用, 将在后面的算例中得到体现.

Step 2: 计算平均加权规则匹配度. 为了避免数据集各类样本分布差别较大, 一类比较集中而另一类比较分散的情况对分类准确率造成的影响, 对每一个样本类, 计算训练样本在规则 R_j 上的匹配度之和并平均化, 即

$$\tilde{\beta}_{\text{Class } h}(R_j) = \frac{1}{m_{c_j}} \sum_{x_p \in \text{Class } h} W_j x_p. \tag{5}$$

其中: $h = 1, 2, \cdots, M$; m_{c_j} 为第 M 类对应的训练样本个数.

Step 3: 引入样本分布函数. 引入样本分布函数 $P_{c_j} = m_{c_j}/m$, 突出样本的全局密度分布, 其中 $m = \sum_{c_j=1}^{M} m_{c_j}$ 为训练样本的总个数.

Step 4: 计算最大匹配度. 将平均加权规则匹配度与样本分布函数相结合, 即使类内规则匹配度得以平均, 又增大了类间的差异. 从而, 找到类 $\hat{h}_j$, $\tilde{\beta}_{\text{Class } h}(R_j)$ 的最大值为

$$\tilde{\beta}_{\text{Class}\,\hat{h}_j}(R_j) = \max_{j=1}^{M}(P_1 \cdot \tilde{\beta}_{\text{Class}\,1}(R_j), \cdots, P_M \cdot \tilde{\beta}_{\text{Class}\,M}(R_j)). \quad (6)$$

Step 5: 规则权重的确定. 规则权重 r 确定如下:

$$r_j = \frac{\tilde{\beta}_{\text{Class}\,\tilde{h}_j}(R_j) - \bar{\tilde{\beta}}}{\sum_{h=1}^{M} \tilde{\beta}_{\text{Class}\,h}(R_j)}, \quad (7)$$

其中

$$\bar{\tilde{\beta}} = \frac{1}{c-1} \sum_{h=1(h\neq\hat{h}_j)}^{M} \tilde{\beta}_{\text{Class}\,h}(R_j). \quad (8)$$

通过对规则权重计算方法的改进, 使得在计算过程中样本类内匹配度得到均匀分布, 同时在判断类别时又加强了类间的对比度, 起到了减小类内差距、加强类间差异的作用.

4 基于模糊规则的分类算法

基于模糊规则的分类需要模糊规则提取和模糊推理两个环节, 本文采用的分类算法由 Chi et al 规则生成算法、通用分类推理模型和改进的模糊规则权重的计算方法 3 个部分组成, 下面给出具体的描述.

4.1 Chi et al 算法

为了生成模糊规则, 设计一个确定变量之间关系和建立特征空间与类空间联系的方法如下: 1) 建立语言划分. 一旦每个特征的变量区域被确定, 便可以计算模糊划分. 2) 对每个样本生成模糊规则: ① 使用合取算子(通常使用 T-范数)计算不同模糊区域样本的匹配度; ② 将样本分配到具有最大隶属度的模糊区域; ③ 生成模糊规则, 即规则前件由所选的模糊区域确定, 规则后件为样本的类标号; ④ 计算规则权重. 该算法是生成模糊规则的基本方法, 可以通过分别对特征空间进行不同的模糊划分来生成模糊规则, 具有很强的可操作性.

4.2 模糊分类算法的设计

考虑由一个新样本 $X_p = (x_{p1}, x_{p2}, \cdots, x_{pn})$ 和规则库组成的 N 个模糊规则, 推理步骤如下:

Step 1: 数据预处理. 对于非平衡数据集, 先对其进行 SMOTE 预处理, 使得处理后的数据集的各类样本数量大致平衡, 然后再进行模糊化处理. 本文中, 使用最小-最大标准化方法[8], 将样本各属性数据在数值上进行模糊化, 使模糊隶属函数在数值上具有相同的 $[0, 1]$ 闭区间, 模糊化公式如下:

$$X_p = \frac{x_p - \min_{j=1}^{m} x_p^j}{\max_{j=1}^{M} x_p^j - \min_{j=1}^{M} x_p^j}. \quad (9)$$

Step 2: 隶属函数模糊划分[8]. 一个数据库由输入空间的模糊划分和前件模糊集的隶属函数两部分组成, 假设第 i 个输入向量 x_i 的区间被均匀分成 K_i 个模糊子集, 标记为 $A_{i1}, A_{i2}, \cdots, A_{ik_i}$, $i = 1, 2, \cdots, n$. 则 n 维的输入空间被分成 $K_1, K_2, \cdots, K_n$ 个模糊子空间.

Step 3: 分类推理. 考虑由一个新样本 $X_p = (x_{p1}, x_{p2}, \cdots, x_{pn})$ 和规则库组成的 N 个模糊规则, 具体推理步骤如下:

Step 3.1: 计算规则匹配度. 计算样本 $X_p = (x_{p1}, x_{p2}, \cdots, x_{pn})$ 对所有规则前件的匹配度, 可以用合取算子 (通常使用 T-范数) 运算得到, 即

$$\mu_j(x_p) = T(\mu_{j1}(x_{p1}), \cdots, \mu_{jn}(x_{pn})). \quad (10)$$

其中: $j = 1, 2, \cdots, N$; $\mu_{ji}(x_{pi})$ 为 A_{ji} 的隶属函数.

Step 3.2: 计算规则关联度. 计算样本 $X_p = (x_{p1}, x_{p2}, \cdots, x_{pn})$ 对各分类所有规则的关联度

$$b_j^k = T(\mu_j(x_p), r_j^k). \quad (11)$$

Step 3.3: 计算样本分类健全度

$$Y_k = f(b_j^k, j = 1, 2, \cdots, N),\ k = 1, 2, \cdots, M. \quad (12)$$

对于规则集中的新样本分类问题, 本文采用两种模糊推理方法:

1) 单一优胜法: 每一个新样本的分类由单一优胜规则的后件类确定, 定义如下:

$$Y_k = \max(b_j^k). \quad (13)$$

其中: $j = 1, 2, \cdots, N$; $k = 1, 2, \cdots, M$; $k = c_j$.

2) 投票加权法: 对每一个模糊规则的后件类进行投票, 每一个类的投票总强度计算如下:

$$Y_k = \sum_{j=1, c_j=k}^{N} b_j^k. \quad (14)$$

Step 3.4: 对决策函数进行分类. 对所有类的样本分类健全度, 应用决策函数 $F(Y_1, \cdots, Y_M) = l$ 进行分类, 此函数将确定最大值相对应的类标号.

Step 4: 计算规则权重. 使用改进的规则权重方法分别对规则权重进行计算, 然后代入 Step 3 进行数据分类.

5 算例分析

文献 [15] 已经证明采用 SMOTE 预处理方法对降低数据的非平衡性的效果优于其他方法. 因而, 使用 SMOTE 方法对数据集进行预处理, 使各类数据集达到数量上的大致平衡, 然后使用基于模糊规则的分类算法对不同非平衡度的数据集进行分类. 将上述两种方法相结合, 对原始数据进行重采样, 以降低数据的非平衡性, 然后采用非平衡数据的分类算法进行分类. 这里使用 4 组具有从低到高不同非平衡度[2]的数据集. 表 1 分别从数据集、属性数、类别名、类分布和非平衡度对数据集进行描述.

表 1　非平衡数据集

数据集	属性数	类别名	类分布	非平衡度
Iris 1	4	(I-S, re)	(33.33, 66.67)	2.00
Class 1	9	(bu,re)	(32.71, 67.29)	2.06
Class 7	9	(hd, re)	(13.55, 86.45)	6.38
Class 5	9	(con, re)	(6.07, 93.93)	15.47

SMOTE 算法只能整数倍增加少数类样本, 因此在多数情况下, 该方法不能达到少数类和多数类样本数量的绝对均衡. 实验表明, 本文中改进的权重计算方法对样本数量相对平衡的数据集能够得到很好的分类效果. 如表 2 所示, 预处理后的非平衡数据基本上达到了多数类和少数类在样本数量上的平衡.

表 2　预处理后的非平衡数据集

数据集	属性数	类别名	类分布	非平衡度
Iris 1	4	(I-S, re)	(50.00, 50.00)	1.00
Class 1	9	(bu,re)	(49.30, 50.70)	1.03
Class 7	9	(hd, re)	(47.68, 52.31)	1.09
Class 5	9	(con, re)	(49.24, 50.76)	1.03

5.1　数据分类结果

基于模糊规则的分类系统使用如下参数设置:

1) 模糊划分: $k = 3, k = 4$;

2) α 的取值: $\alpha = 0.2, 0.4, 0.6, 0.8, 1$;

3) 规则权重计算方法: 改进的计算方法;

4) 模糊推理方法: 采用单一优胜法和投票加权法.

将表 3∼表 6 中最优准确率进行归纳, 如表 7 所示. 可以看出, 基于模糊规则的分类系统对经过预处理后的非平衡数据集有很好的分类效果, 且分类准确率随着取值的改变而取得最优值.

表 3　Iris 数据集的分类准确率

α	单一优胜法		投票加权法	
	$k=3$	$k=4$	$k=3$	$k=4$
0.2	100.00	100.00	100.00	100.00
0.4	100.00	100.00	100.00	100.00
0.6	100.00	100.00	100.00	100.00
0.8	100.00	100.00	100.00	100.00
1.0	100.00	100.00	100.00	100.00

表 4　Class 1 数据集的分类准确率

α	单一优胜法		投票加权法	
	$k=3$	$k=4$	$k=3$	$k=4$
0.2	72.54	82.04	71.48	83.10
0.4	71.13	75.70	71.48	83.80
0.6	71.13	75.00	71.48	79.23
0.8	70.42	74.30	71.48	76.76
1.0	73.24	73.94	71.83	76.41

表 5　Class 7 数据集的分类准确率

α	单一优胜法		投票加权法	
	$k=3$	$k=4$	$k=3$	$k=4$
0.2	93.56	97.94	95.62	99.23
0.4	94.59	96.91	96.93	98.20
0.6	94.59	97.16	97.16	97.94
0.8	95.36	97.42	97.16	98.20
1.0	95.36	97.68	96.65	97.94

表 6　Class 5 数据集的分类准确率

α	单一优胜法		投票加权法	
	$k=3$	$k=4$	$k=3$	$k=4$
0.2	95.20	97.98	90.91	97.47
0.4	93.94	97.98	91.16	98.23
0.6	94.19	98.48	91.41	98.23
0.8	94.44	98.99	91.41	98.99
1.0	95.20	98.99	91.92	98.74

表 7　最优准确率

数据集	Iris 1	Class 1	Class 7	Class 5
单一优胜法	100.00	82.04	97.94	98.99
投票加权法	100.00	83.80	99.23	98.99

5.2　不同算法分类结果比较

为了评价本文提出的基于模糊规则的分类算法的性能, 将文献 [15] 中经 SMOTE 预处理的 Chi et al, Ishibuchi et al, 决策树分类法 (C4.5) 和未经过预处理的 E-算法, 同本文方法分别对非平衡数据集进行分类, 实验结果比较见表 8.

表 8　不同算法分类准确率的比较

数据集	Chi-S	Ish-S	E-Alg	C4.5-S	本文方法
Iris 1	100.00	100.00	100.00	100.00	100.00
Class 1	74.44	72.22	71.56	94.23	83.80
Class 7	94.75	85.78	80.21	98.14	99.23
Class 5	98.87	87.03	84.82	98.42	98.99

从表 8 中可以看出, 除了在 Class 1 数据集上分类的准确率不及 C4.5 算法外, 本文方法都能得到较高的准确率, 且明显优于 Chi et al, Ishibuchi et al 和 E-算法的分类情况, 这充分说明该系统具有较好的分类准确率.

6　结　论

针对非平衡数据的分类问题, 首先使用 SMOTE 算法对其进行预处理, 将模糊规则生成算法与模糊分类推理模型相结合, 设计了一种有效调节平衡数据与非平衡数据分类准确率的分类算法; 同时, 对模糊规则权重的计算方法进行了修改, 使其更具合理性. 最后, 采用该方法对具有不同非平衡度的 UCI 数据集进行对比实验, 实验结果表明该分类器具有很高的分类准确率, 且优于其他分类算法的准确率.

参考文献(References)

[1] Haibo He, Edwardo A Garcia. Learning from imbalanced data[J]. IEEE Trans on Knowledge and Data Engineering, 2009, 21: 645-663.

[2] Orriols-Puig A, Bernado-Mansilla E. Evolutionary rule-based systems for imbalanced datasets[J]. Soft Computing, 2009, 13(3): 213-225.

[3] Mazurowski M, Habas P, Zurada J Lo, et al. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance[J]. Neural Networks, 2008, 21(2/3): 427-436.

[4] Huang Y M, Hung C M, Jian H C. Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem[J]. Nonlinear Analysis: Real World Application, 2006, 7(4): 720-747.

[5] Mu Chen Chen, Long Sheng Chen, Chun Chin Hsu, et al. An information granulation based data mining approach for classifying imbalanced data[J]. Information Sciences, 2008, 178(8): 3214-3227.

[6] Xu L, Chow M, Taylor L. Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification E-algorithm[J]. IEEE Trans on Power Systems, 2007, 22(1): 164-171.

[7] Ishibuchi H, Yamamoto T. Comparison of heurisric criteria for fuzzy rule selection in classification problems[J]. Fuzzy Optimization Decision Making, 2004, 3(2):119-139.

[8] Ken Nozaki, Hisao Ishibuchi, Hideo Tanaka. A simple but powerful heuristic method for generating fuzzy rules from numerical data[J]. Fuzzy Sets and Systems, 1997, 86: 251-270.

[9] Zolghadri Jahromi M, Taheri M. A proposed method for learning rule weights in fuzzy rule-based classification systems[J]. Fuzzy Sets and Systems, 2008, 159: 449-459.

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

[11] Ishibuchi H, Yamamoto T. Rule weight specification in fuzzy rule-based classification systems[J]. IEEE Trans on Fuzzy System, 2005, 13(4): 428-435.

[12] Li Jie, Deng Yi-Ming, Shen Shi-Tuan. Classification rule extraction based on fuzzy area distribution and classification reasoning algorithm[J]. J of Omputers, 2008, 31(6): 934-941.

[13] Hisao Ishibuchi, Tomoharu Nakashima, Tadahiko Murata. Performance evaluation of fuzzy classifier systems for multidimensional pattern classification problems[J]. IEEE Trans on Systems, Man, and Cybernetics－Part B: Cybernetics, 1999, 29(5): 601-618.

[14] Chi Z, Yan H, Pham T. Fuzzy algorithms with applications to image processing and pattern recognition[M]. World Scientific. Singapore, 1996.

[15] Fernandez A, Garcia S, Jose del Jesus M. A study of the behavior of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets[J]. Fuzzy Sets and Systems, 2008, 159: 2378-2398.

下 期 要 目