文章编号:1007-5321(2019)02-0101-07 **DOI**:10.13190/j.jbupt.2018-178

基于模糊粗糙集实例选择的混合算法在信用评分中的应用

刘占峰, 潘 甦
(南京邮电大学 江苏省通信与网络技术工程研究中心,南京 210003)

摘要: 基于聚类算法的混合分类器构建的信息评分系统中,不合理的聚类值或者初始类簇中心点会严重影响分类精度的问题,对此,提出了 2 种基于模糊粗糙集实例选择的新型混合算法. 这 2 种算法仅与数据集的数据结构有关,不受其他外部参数影响. 实验结果表明,基于模糊粗糙集实例选择的 2 种混合算法针对不同结构的数据集表现出了各自的特性,深化了对数据集的理解,提高了准确率.

关 键 词: 模糊粗糙集实例选择; 混合算法; 信用评分
中图分类号: TP181 **文献标志码**: A

Hybrid Algorithm Base on Fuzzy-Rough Instance Selection for Credit Scoring

LIU Zhan-feng, PAN Su
(Jiangsu Engineering Research Center of Communication and Network Technology,
Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: For the credit scoring system built on cluster algorithm based hybrid classifier, the unreasonable clusters number or starting center points of each cluster have severely negative influence on the classification accuracy. In order to solve the problem, two new hybrid algorithms based on fuzzy-rough instance selection were proposed respectively, which are only related to intrinsic data structure of datasets and are not affected by other external parameters. The experimental results show that the proposed hybrid algorithms exhibit their own characteristics for datasets with different structures, which deepens the understanding of data sets and improves the accuracy.

Key words: fuzzy-rough instance selection; hybrid algorithm; credit scoring

信用评分技术用于甄别好坏申请者和评估潜在风险,在金融风险控制领域扮演着重要的角色[1]. 在构建信用评分系统过程中,大数据技术挖掘历史信贷数据中的高价值特征来鉴别高风险申请者[2].

由于冗余和无关特征的影响,基于单个分类器的信用评分系统往往精度不高,研究表明特征选择能够有效提升单分类器性能[3-5]. 此外,孤立和不一致的实例也会严重影响分类器的性能,因此研究人员设计出基于聚类算法的混合分类器[6],用自组织映射(SOM, self-organizing maps)算法确定聚类值和初始类簇中心点. 然后用 k-means 算法进行聚类,将实例分配到相应的类并剔除每类的非典型实例. 由于没有既定的规则用于选择最佳参数,因此 SOM 和 k-means 算法的参数都需要通过反复试验确定. 另

收稿日期: 2018-09-13
基金项目: 江苏省研究生科研与实践创新计划项目(KYCX18_0882); 南京邮电大学江苏省通信与网络技术工程研究中心开放课题
作者简介: 刘占峰(1980—), 男, 博士生.
通信作者: 潘 甦(1969—), 男, 教授, 博士生导师, E-mail: supan@njupt.edu.cn.

外,不合理的聚类值或者初始类簇中心点会严重影响聚类结果. 实例选择技术也可以有效地甄别非典型数据,国内许多研究人员对该技术进行了深入的研究[7-10]. 基于模糊粗糙集理论(FRST, fuzzy-rough set theory)[11]的实例选择(FRIS, fuzzy-rough instance selection)技术[12],仅利用FRST的正域剔除噪声实例,有效提升数据质量和训练时间,并在一些领域取得了成功的应用[13-14],但还尚未应用在信用评分领域.

提出了一种应用在信用评分领域的基于FRIS的新型混合算法,并探讨了基于文献[12]中两种FRIS方法的混合分类器. 实验结果表明,新型混合算法获得了较好的准确率,并且在不同结构的数据集上展现了各自的优势,深化了对数据集的认识.

1 背景知识

1.1 粗糙集分析

在粗糙集理论中[15],数据被视为信息系统(U, S),其中U和S分别为实例和特征的有限非空集合. S中每一项a都对应$U\to V_a$的映射, 其中V_a为a在实例集合U上的值域. 对于特征集合S的每一个子集B,衡量实例间可区分程度的指标——不可区分关系R_B定义如下:

$$R_B=\{(x,y)\in U^2 \text{ 且 } (\forall a\in B)(a(x)=a(y))\} \tag{1}$$

显而易见,R_B是等价关系. 对于B的不可区分关系的等价类记为$[x]_{R_B}$. 对$A\subseteq U$,R_B的上下近似由下式定义:

$$R_B\downarrow A=\{x\in U\mid [x]_{R_B}\subseteq A\} \tag{2}$$

$$R_B\uparrow A=\{x\in U\mid [x]_{R_B}\cap A\neq\varnothing\} \tag{3}$$

通过包含决策特征可以把信息系统扩展为决策系统$(U,S\cup\{d\})$[16],其中$d(d\notin S)$为决策特征,在信用评分中作为目标变量表示是否违约,其等价类$[x]_{R_B}$称为决策类. 假设$B\subseteq S$,B的正域POS_B包含U中所有可根据B的值预测决策类的实例:

$$\mathrm{POS}_B=\cup_{x\in X}R_B\downarrow [x]_{R_d} \tag{4}$$

若$x\in\mathrm{POS}_B$,对于任何实例,只要其在B中的特征与x取值相等,那么该实例就与x具有相同的决策类.

1.2 模糊粗糙集

将粗糙集上下近似公式(式(2)和式(3))扩展到模糊情况,可以把模糊集和粗糙集整合到一起[11].

对于S的任意子集B,定义模糊集B的不可区分关系如下:

$$(x,y)=\mathscr{T}(\underbrace{R_a(x,y)}_{a\in B}) \tag{5}$$

其中:$\mathscr{T}$为t范数. 对定性属性a,如果$a(x)=a(y)$,则$R_a(x,y)=1$,反之为$R_a(x,y)=0$.

对U中的y,根据上述不可区分关系,定义模糊集B的正域为

$$\mathrm{POS}_B(y)=(\cup_{x\in X}R_B\downarrow R_dx)(y) \tag{6}$$

由于式(6)的计算代价过高,当决策特征是离散值时,用式(7)代替[17]:

$$\mathrm{POS}'_B(y)=(R_B\downarrow R_dy)(y) \tag{7}$$

2 基于FRIS的混合算法

首先介绍几种典型FRIS技术的实现原理及优缺点,并在此基础上提出基于FRIS的混合算法.

2.1 典型FRIS技术

FRIS的核心思想是利用正域信息判断实例的有用程度以及是否应该保留.

对决策系统$(U,S\cup\{d\})$,令a为$(U,S\cup\{d\})$中的数值型特征,取值范围是$l(a)$. 为了表示x和y两个实例对特征a的近似相等,采用下述模糊关系R_a[12]:

$$R_a^{\alpha}(x,y)=\max\left(0,1-\alpha\frac{|a(x)-a(y)|}{l(a)}\right) \tag{8}$$

其中参数$\alpha(\alpha\geqslant 0)$决定$R_a^{\alpha}$的颗粒度. 式(8)并非是定义实例$x$和$y$之间基于特征$a$的相似度的唯一假设,可根据需要采取其他形式的模糊关系公式.

文献[12]中介绍了3种FRIS方法,但FRIS-III方法的计算过于复杂,因此选择FRIS-I和FRIS-II探索其在信用评分系统混合算法中的优势. FRIS-I方法如下文所述,其计算每个实例对正域的隶属度,剔除隶属度小于门限参数τ的实例. FRIS-I方法简单高效,但由于其是根据固定门限参数剔除实例,没有考虑实例剔除后正域的变化,因此通常会剔除比实际更多的实例.

FRIS-I(S,α,τ)

输入:

S,将被约简的实例集合;

输出:

Y,约简后的实例集合;

参数:

α,颗粒度参数;

τ,自定义门限.

1　$Y \leftarrow S$

2　foreach $x \in S$

3　　if($\mathrm{POS}_A^{\alpha,S}(x) < \tau$)

4　　$Y \leftarrow Y - \{x\}$

5　return Y

针对 FRIS-I 方法的弊端, FRIS-II 方法每次剔除隶属度最小的实例,然后重新计算每个实例对新数据集正域的隶属度,反复迭代直到所有实例都完全属于正域. 与 FRIS-I 方法相比,其不需要预先设定门限参数,通过动态的正域来剔除噪声实例.

FRIS-II(S,α)

输入:

S,将被约简的实例集合;

输出:

S,约简后的实例集合;

参数:

α,颗粒度参数.

1　while(true)

2　$z \leftarrow 0$,$\rho_z \leftarrow 1$

3　foreach $x \in S$

4　　if($\mathrm{POS}_A^{\alpha,S}(x) < \rho_z$)

5　　$z \leftarrow x$

6　　$\rho_z \leftarrow \mathrm{POS}_A^{\alpha,S}(x)$

7　if ($z \neq 0$)

8　　$S \leftarrow S - \{z\}$

9　else return S

2.2　混合算法

笔者提出的信用评分模型的统一算法框架如图 1 所示,将 FRIS-I 和 FRIS-II 分别应用在该框架里构建混合分类器,详细流程如下.

步骤 1　初始化 FRIS 参数:(α,τ)或者 α. 计算数据集中每个实例的隶属度并根据相应的规则剔除实例,被剔除的实例归到待定数据集.

步骤 2　保留的实例纳入核心数据集. 用 KNN 算法检查待定数据集,凡是 KNN 算法预测类别与实际类别相同的实例一律重新纳入约简后的数据集,形成约简数据集.

步骤 3　在约简数据集上应用 SVM 分类器,并引入交叉检验提升模型的泛化能力.

步骤 4　验证不同的参数组合,得到产生最小交叉检验分类误差的最佳参数组合.

步骤 5　对 2 个分类器使用最佳参数组合构建高性能的混合分类器.

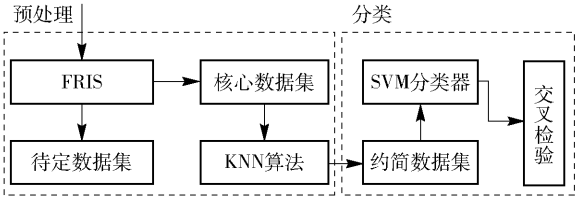

图 1　混合算法架构

3　实验结果

UCI 机器学习库[18]中的 German 和 Australian 信用数据集是 2 个广泛用于信用评分算法评估的数据集,使用这 2 个数据集验证算法便于与其他的研究成果比较. 数据集的基本情况如表 1 所示.

表 1　数据集概况

数据集名称	类别	实例数	名义型特征	数值型特征
German	2	1 000	0	24
Australian	2	690	6	8

在 Ubuntu 系统用 R 语言仿真实验结果,首先在 KNN 算法中令参数 $k=5$,并分别采用线性、多项式和 RBF 3 种不同核函数的 SVM 算法(分别简记为 SVML、SVMP 和 SVMR)对约简后的数据集建模. 为了便于重复实验结果,随机种子设为 123. 为了得到可靠稳定的模型,在数据集上应用 10 折交叉检验(10-CV)[19]和弃一法交叉检验(LOO-CV)[20],并比较分类精度和约简后的实例数. 2 种交叉检验法都是利用规则将原始数据进行分组,一部分作为训练集;另一部分作为测试集,循环遍历计算每次分组结果并平均得到最终模型. 10-CV 与 LOO-CV 不同之处在于分组规则和计算量不同,前者将数据均分为 10 折,计算 10 次;后者每次取一个数据作为分组,计算次数等于数据集大小.

为了得到更好的分类精度,在仿真中为每个核函数构建参数搜索网格,以期获得最佳的参数组合.

3.1　混合分类器 1 的实验结果

表 2 ~ 表 7 显示了混合分类器 1(HC1, hybrid classifier 1)在不同颗粒度系数和门限系数参数组合时经 2 种交叉检验后的分类精度和约简后实例数. 当颗粒度系数取值较大时,实例选择能力不显著,故此处选择实例选择作用明显的区间[0.2,1.0]. 门限系数取值过小,则剔除噪声能力变差,此处选择{0.7,0.8,0.9}3 个值来观察噪声数据剔除情况.

表2 $\tau=0.7$ 时 HC1 的实例数和分类精度(German 信用数据)

α	实例数	10折交叉检验			弃一法交叉检验		
		SVML	SVMP	SVMR	SVML	SVMP	SVMR
0.2	732	0.948 1	0.964 5	0.942 7	0.959 0	0.960 4	0.948 1
0.3	819	0.860 8	0.879 2	0.870 5	0.866 9	0.886 4	0.868 1
0.4	877	0.832 4	0.855 2	0.840 3	0.840 4	0.849 5	0.835 8
0.5	923	0.810 5	0.826 7	0.796 3	0.800 7	0.826 7	0.805 0
0.6	948	0.791 0	0.802 7	0.783 8	0.790 1	0.811 2	0.788 0
0.7	964	0.788 5	0.786 3	0.787 4	0.792 5	0.791 5	0.785 3
0.8	973	0.776 0	0.787 3	0.782 2	0.779 0	0.795 5	0.779 0
0.9	986	0.766 7	0.780 0	0.774 8	0.776 9	0.786 0	0.774 8
1.0	990	0.770 7	0.789 0	0.771 8	0.777 8	0.789 9	0.773 7

表3 $\tau=0.8$ 时 HC1 的实例数和分类精度(German 信用数据)

α	实例数	10折交叉检验			弃一法交叉检验		
		SVML	SVMP	SVMR	SVML	SVMP	SVMR
0.2	714	0.959 3	0.973 3	0.960 8	0.967 8	0.966 4	0.965
0.3	766	0.894 2	0.920 4	0.895 6	0.896 9	0.926 9	0.896 9
0.4	845	0.841 4	0.869 7	0.841 5	0.839 1	0.868 6	0.847 3
0.5	904	0.814 2	0.827 3	0.826 3	0.827 4	0.829 6	0.826 3
0.6	942	0.793	0.801 5	0.793	0.798 3	0.806 8	0.796 2
0.7	952	0.785 6	0.792	0.786 7	0.783 6	0.804 6	0.785 7
0.8	964	0.788 5	0.786 3	0.787 4	0.792 5	0.791 5	0.785 3
0.9	973	0.776	0.787 3	0.782 2	0.779	0.795 5	0.779
1.0	980	0.780 6	0.777 5	0.777 5	0.771 4	0.790 8	0.777 6

表4 $\tau=0.9$ 时 HC1 的实例数和分类精度(German 信用数据)

α	实例数	10折交叉检验			弃一法交叉检验		
		SVML	SVMP	SVMR	SVML	SVMP	SVMR
0.2	713	0.977 6	0.983 2	0.977 5	0.973 4	0.986 0	0.979 0
0.3	742	0.940 7	0.946 1	0.938 0	0.944 7	0.954 2	0.942 0
0.4	822	0.855 2	0.867 5	0.855 3	0.860 1	0.871 0	0.861 3
0.5	870	0.834 4	0.851 7	0.828 7	0.832 2	0.847 1	0.834 5
0.6	913	0.813 7	0.819 2	0.807 3	0.817 1	0.828 0	0.811 6
0.7	940	0.794 7	0.800 0	0.792 6	0.794 7	0.808 5	0.793 6
0.8	952	0.791 1	0.795 1	0.777 3	0.786 8	0.804 6	0.782 6
0.9	964	0.788 5	0.786 3	0.787 4	0.792 5	0.791 5	0.785 3
1.0	972	0.776 8	0.790 3	0.781 0	0.778 8	0.795 3	0.782 9

表5 $\tau=0.7$ 时 HC1 的实例数和分类精度(Australian 信用数据)

α	实例数	10折交叉检验			弃一法交叉检验		
		SVML	SVMP	SVMR	SVML	SVMP	SVMR
0.2	648	0.890 5	0.916 8	0.892 2	0.888 9	0.910 5	0.893 5
0.3	648	0.890 5	0.916 8	0.892 2	0.888 9	0.910 5	0.893 5
0.4	648	0.890 5	0.916 8	0.892 2	0.888 9	0.910 5	0.893 5
0.5	649	0.883 1	0.913 6	0.893 8	0.889 1	0.909 1	0.893 7
0.6	651	0.874 1	0.904 8	0.891 0	0.860 2	0.909 4	0.890 9
0.7	651	0.892 5	0.912 3	0.900 3	0.875 6	0.918 6	0.892 5
0.8	652	0.886 6	0.915 8	0.886 4	0.880 4	0.918 7	0.894 2
0.9	654	0.876 3	0.912 9	0.894 6	0.868 5	0.906 7	0.891 4
1.0	656	0.888 7	0.913 0	0.881 0	0.859 8	0.905 5	0.890 2

表6 $\tau=0.8$ 时 HC1 的实例数和分类精度(Australian 信用数据)

α	实例数	10折交叉检验			弃一法交叉检验		
		SVML	SVMP	SVMR	SVML	SVMP	SVMR
0.2	648	0.890 5	0.916 8	0.892 2	0.888 9	0.910 5	0.893 5
0.3	648	0.890 5	0.916 8	0.892 2	0.888 9	0.910 5	0.893 5
0.4	648	0.890 5	0.916 8	0.892 2	0.888 9	0.910 5	0.893 5
0.5	649	0.890 7	0.915 3	0.892 4	0.889 1	0.909 1	0.895 2
0.6	650	0.882 8	0.906 2	0.895 2	0.867 7	0.909 2	0.892 3
0.7	651	0.874 1	0.904 8	0.891 0	0.860 2	0.909 4	0.890 9
0.8	651	0.892 5	0.912 3	0.900 3	0.875 6	0.918 6	0.892 5
0.9	652	0.889 7	0.918 7	0.889 5	0.880 4	0.918 7	0.894 2
1.0	653	0.876 0	0.911 3	0.889 9	0.876 0	0.915 8	0.892 8

表7 $\tau=0.9$ 时 HC1 的实例数和分类精度(Australian 信用数据)

α	实例数	10折交叉检验			弃一法交叉检验		
		SVML	SVMP	SVMR	SVML	SVMP	SVMR
0.2	648	0.890 5	0.916 8	0.892 2	0.888 9	0.910 5	0.893 5
0.3	648	0.890 5	0.916 8	0.892 2	0.888 9	0.910 5	0.902 8
0.4	648	0.890 5	0.916 8	0.892 2	0.888 9	0.910 5	0.902 8
0.5	648	0.890 5	0.916 8	0.892 2	0.888 9	0.910 5	0.893 5
0.6	649	0.890 7	0.915 3	0.892 4	0.889 1	0.909 1	0.895 2
0.7	650	0.882 8	0.906 2	0.895 2	0.867 7	0.909 2	0.892 3
0.8	651	0.874 1	0.904 8	0.891 0	0.860 2	0.909 4	0.890 9
0.9	651	0.892 5	0.912 3	0.900 3	0.875 6	0.918 6	0.892 5
1.0	651	0.886 5	0.906 2	0.901 9	0.875 6	0.918 6	0.892 5

从分类结果来看，2 个信用数据集在 2 种交叉检验下的最佳性能都是通过多项式核函数的 SVM 算法达到的. German 信用数据集在 2 种交叉检验下的最佳分类精度分别是 0.983 2 和 0.986 0，最优的实例数是 713，是原始数据集大小的 71.3%. Australian 信用数据集在 2 种交叉检验下的最佳分类精度都是 0.918 7，最优的实例数是 652，是原始数据集大小的 94.5%. 上述分析表明，German 数据集中孤立实例和不一致实例的占比更大.

3.2　混合分类器 2 的实验结果

表 8 和表 9 显示了混合分类器 2（HC2）在不同颗粒度系数下经过 2 种交叉检验后的分类精度和约简后实例数. 颗粒度系数的取值原则参考 HC1. 从整体上来看，依然是多项式核函数 SVM 算法性能最佳，但 Australian 数据集的最佳分类精度却是线性核函数 SVM 算法获得. German 信用数据集在 2 种交叉检验下的最佳分类精度分别是 0.977 5 和 0.978 8，并且取得最优值时的颗粒度参数相同，对应的最优实例数是 709，是原始数据集大小的 70.9%. Australian 信用数据集在弃一法交叉检验下的最佳分类精度是 0.988 1，相应的实例数是 505，为原始数据集大小的 73.2%. 与 HC1 相比，HC2 不需要调整门限参数 τ，虽然 German 信用数据集的最佳分类精度略有下降，但约简数据集实例数更少. 对于 German 信用数据集来说，HC2 约简后的实例数介于 HC1 当 τ 分别取 0.8 和 0.9 得到的实例数之间. Australian 信用数据集约简实例数在 HC2 中变化比较大，并且用更少的实例得到了比 HC1 中更好的分类精度，这说明 HC2 获得了比 HC1 质量更好的约简数据集.

表 8　HC2 的实例数和分类精度（German 信用数据集）

α	实例数	10 折交叉检验			弃一法交叉检验		
		SVML	SVMP	SVMR	SVML	SVMP	SVMR
0.2	709	0.967 5	0.977 5	0.969 0	0.966 1	0.978 8	0.969 0
0.3	757	0.914 1	0.934 0	0.918 1	0.915 5	0.941 9	0.919 4
0.4	845	0.839 1	0.856 9	0.847 4	0.829 6	0.854 4	0.846 2
0.5	887	0.800 5	0.813 9	0.807 2	0.812 9	0.818 5	0.805 0
0.6	921	0.792 8	0.796 9	0.798 1	0.798 0	0.802 4	0.797 0
0.7	943	0.788 9	0.788 8	0.792 1	0.793 2	0.799 6	0.792 2
0.8	960	0.786 4	0.789 5	0.784 4	0.790 6	0.792 7	0.782 3
0.9	968	0.771 7	0.787 2	0.779 0	0.783 1	0.789 3	0.782 0
1.0	974	0.774 2	0.786 6	0.765 8	0.780 3	0.783 4	0.777 2

表 9　HC2 的实例数和分类精度（Australian 信用数据集）

α	实例数	10 折交叉检验			弃一法交叉检验		
		SVML	SVMP	SVMR	SVML	SVMP	SVMR
0.2*	309	–	–	–	–	–	–
0.3	505	0.986 2	0.982 2	0.976 3	0.988 1	0.982 2	0.974 3
0.4	534	0.941 9	0.951 3	0.951 3	0.940 1	0.956 9	0.947 6
0.5	555	0.953 3	0.953 1	0.947 9	0.953 2	0.956 8	0.951 4
0.6	582	0.957 1	0.955 3	0.951 9	0.957 0	0.955 3	0.948 5
0.7	606	0.942 3	0.950 5	0.945 5	0.945 5	0.957 1	0.947 2
0.8	608	0.947 4	0.952 3	0.947 3	0.949 0	0.952 3	0.945 7
0.9	619	0.945 2	0.948 4	0.940 4	0.940 2	0.946 7	0.941 8
1.0	626	0.931 2	0.940 9	0.932 8	0.932 9	0.940 9	0.934 5

* 当 $\alpha=0.2$，约简数据集过小，对原始数据集不具有代表性

2 种混合分类器与线性判别分析（LDA，linear discriminant analysis）、逻辑回归（LR，logistics regression）和神经网络（NN，neural network）等未做实例选择基准分类器[21]的最佳分类精度比较见表 10. 从比较结果来看，2 种混合分类器取得了明显优于基准分类器的分类精度，尤其是 HC2 对 Australian 数据集展示了更加强大的信用评估能力. 对German数据集，2 种混合分类器没有明显的分类精度差异.

表 10　不同算法对 Australian 和 German 信用数据集的分类精度

算法	Australian	German
LDA	0.852 0	0.660 0
LR	0.857 0	0.724 0
NN	0.868 3	0.752 0
HC1	0.918 7	0.986 0
HC2	0.988 1	0.978 8

图 2 和图 3 是根据密度聚类算法[22]绘制的 German 和 Australian 数据集的 MDS 图. MDS 图表明，German 的数据分布比较分散，而 Australian 数据相对比较集中. 从表 2 ~ 表 7 可以看出，固定 τ 值，German 信用数据的约简结果随着 α 的改变而剧烈变化，而 Australian 信用数据的约简结果却对比变化不大. 再观察表 8 和表 9，随着 α 值的变化，2 个数据集的约简结果变化趋势相似，约简结果完全由颗粒度系数决定. 上述现象说明，由于 German 数据集分散的特性，HC1 通过固定门限剔除了更多的孤立和不一致实例. 而 HC2 通过动态的正域探索数

据结构,比 HC1 剔除了更多的 Australian 数据集中的噪声实例,但却用更少的实例获得了更好的分类精度. 从实验结果来看,HC1 更适合分散型的数据集,而 HC2 对相对集中的数据集更加有效.

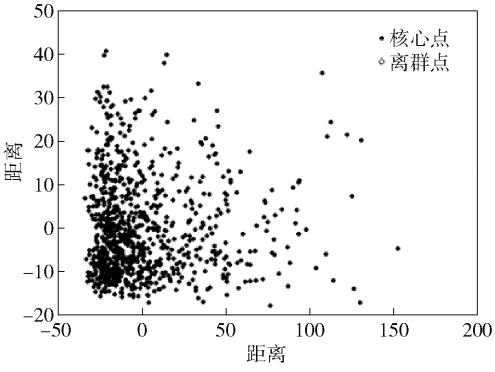

图2 German 数据集的 MDS 图

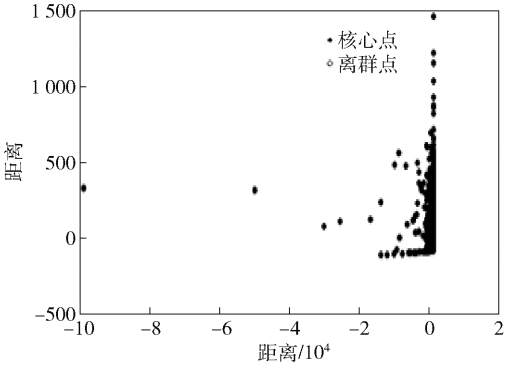

图3 Australian 数据集的 MDS 图

4 结束语

提出的基于 FRIS 的信用评分混合算法不受外部参数影响,通过选取适当的参数,性能要远远优于 LDA、LR 和 NN 等基准分类器. 由于 2 种分类器的原理不同,HC1 更适合分散型的数据集,而 HC2 对相对集中的数据集更加有效. 在保证分类性能的情况下,同时约简实例和特征[23]是值得进一步探索的领域.

参考文献:

[1] Yi Baiheng, Zhu Jianjun. Credit scoring with an improved fuzzy support vector machine based on grey incidence analysis[C]// IEEE International Conference on Grey Systems and Intelligent Services. New York: IEEE Press, 2015: 173-178.

[2] Lahsasna A, Ainon R N, Wan T Y. Credit scoring models using soft xomputing methods: a survey[J]. The International Arab Journal of Information Technology, 2010, 7(2): 115-123.

[3] Ramya R S, Kumaresan S. Analysis of feature selection techniques in credit risk assessment[C]// International Conference on Advanced Computing and Communication Systems. New York: IEEE Press, 2015: 1-6.

[4] Lin C C, Chang C C, Li F C, et al. Features selection approaches combined with effective classifiers in credit scoring[C]// IEEE International Conference on Industrial Engineering and Engineering Management. New York: IEEE Press, 2011: 752-757.

[5] Yao Ping. Fuzzy rough set and information entropy based feature selection for credit scoring[C]// International Conference on Fuzzy Systems and Knowledge Discovery. New York: IEEE Press, 2009: 247-251.

[6] Hsieh N C. Hybrid mining approach in the design of credit scoring models[J]. Expert Systems with Applications, 2005, 28(4): 655-665.

[7] 田春娜, 高新波, 李洁. 基于嵌入式 Bootstrap 的主动学习示例选择方法[J]. 计算机研究与发展, 2006, 43(10): 1706-1712.
Tian Chunna, Gao Xinbo, Li Jie. An example selection method for active learning based on embedded bootstrap algorithm[J]. Journal of Computer Research and Development, 2006, 43(10): 1706-1712.

[8] 刘振兴. 主动示例选择算法及其在人脸检测中的应用[D]. 西安: 西安电子科技大学, 2010.

[9] 韩光辉. 基于欧式距离的实例选择算法研究[D]. 保定: 河北大学, 2011.

[10] 张宁. 基于近邻分类的实例选择算法研究[D]. 保定: 河北大学, 2009.

[11] Dubois D, Prade H. Putting rough sets and fuzzy sets together[M]// SLOWINSKI R. Intelligent decision support. Berlin: Springer, 1992: 203-232.

[12] Jensen R, Cornelis C. Fuzzy-rough instance selection[C]// International Conference on Fuzzy Systems. New York: IEEE Press, 2010: 1-7.

[13] Jhawar A, Chan C S, Monekosso D, et al. Fuzzy-rough based decision system for gait adopting instance selection[C]// IEEE International Conference on Fuzzy Systems. New York: IEEE Press, 2016: 1127-1133.

[14] Kang Xiaomeng, Liu Xiaopeng, Zhai Jjunhai, et al. Instances selection for NN with fuzzy rough technique[C]// International Conference on Machine Learning and Cybernetics. New York: IEEE Press, 2011: 1097-1100.

[15] Pawlak Z. Rough sets: theoretical aspects of reasoning

about data [M]. Norwell: Kluwer Academic Publishing, 1992.

[16] Jensen R, Shen Qiang. Computational intelligence and feature selection: rough and fuzzy approaches [M]. Hoboken: Wiley, 2008.

[17] Cornelis C , Jensen R , Hurtado G, et al. Attribute selection with fuzzy decision reducts[J]. Information Sciences, 2010, 180(2): 209-224.

[18] Frank A, Asuncion A. UCI machine learning repository [EB/OL]. (2010-09-16). http: //archive. ics. uci. edu/ml.

[19] Geisser S. The predictive sample reuse method with applications [J]. Journal of the American Statistical Association, 1975, 70(350): 320-328.

[20] Geisser S. A predictive approach to the random effect model [J]. Biometrika, 1974, 61(1): 101-107.

[21] Yao Ping. Hybrid classifier using neighborhood rough set and SVM for credit scoring[C]//International Conference on Business Intelligence and Financial Engineering. New York: IEEE Press, 2009: 138-142.

[22] Rodriguez, Laio A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492-1496.

[23] Parthalain N M, Jensen R. Simultaneous feature and instance selection using fuzzy-rough bireducts[C]//IEEE International Conference on Fuzzy Systems. New York: IEEE Press, 2013: 1-8.