两层聚类的类别不平衡数据挖掘算法

胡小生[1] 张润晶[2] 钟 勇[1]
(佛山科学技术学院电子与信息工程学院 佛山 528000)[1]
(佛山科学技术学院信息与教育技术中心 佛山 528000)[2]

摘 要 类别不平衡数据分类是机器学习和数据挖掘研究的热点问题。传统分类算法有很大的偏向性,少数类分类效果不够理想。提出一种两层聚类的类别不平衡数据级联挖掘算法。算法首先进行基于聚类的欠采样,在多数类样本上进行聚类,之后提取聚类质心,获得与少数类样本数目相一致的聚类质心,再与所有少数类样例一起组成新的平衡训练集,为了避免少数类样本数量过少而使训练集过小导致分类精度下降的问题,使用SMOTE过采样结合聚类欠采样;然后在平衡的训练集上使用K均值聚类与C4.5决策树算法相级联的分类方法,通过K均值聚类将训练样例划分为K个簇,在每个聚类簇内使用C4.5算法构建决策树,通过K个聚簇上的决策树来改进优化分类决策边界。实验结果表明,该算法具有处理类别不平衡数据分类问题的优势。

关键词 数据挖掘,分类,不平衡数据,K均值聚类
中图法分类号 TP391 **文献标识码** A

Two-tier Clustering for Mining Imbalanced Datasets

HU Xiao-sheng[1] ZHANG Run-jing[2] ZHONG Yong[1]
(College of Electronic and Information Engineering, Foshan University, Foshan 528000, China)[1]
(Information and Education Technology Center, Foshan University, Foshan 528000, China)[2]

Abstract Classification of class-imbalanced data becomes a research hot topic in machine learning and data mining. Most classification algorithms tend to predict that most of the incoming data belongs to the majority class, resulting in the pool classification performance in minority class instances, which are usually much more of interest. In this paper, a two-tier clustering cascading mining algorithm was proposed. The algorithm first constructs balanced training set by clusterd-based under-sampling, using K-means clustering to cluster majority class and extract cluster centroids then merge with all minority class instances to generate a balanced training set for training. To avoid the number of the minority is too small, leading the shortage of training instance, combination of SMOTE over-sampling and cluster-based under-sampling is used; next, using "K-means+C4.5", a method to cascade K-means clustering and C4.5 decision tree algorithm for classifying on the balanced training set, the K-means clustering method is first used to parition the training instances into k clusters, and on each cluster, C4.5 algorithm is used to build decision tree, the decision tree on each cluster refines the decision boundaries by learning the subgroups within the cluster. Experimental results show that the proposed method provides better classification performance than other approaches on both minority and majority classes, and is effective and feasible to deal with the imbalanced datasets.

Keywords Data mining, Classification, Imbalanced data, K-means clustering

1 引言

数据挖掘是从大量的实际应用数据中发现有效的、新颖的、潜在有用的、最终能够被用户理解的知识的过程。数据挖掘主要任务包括:描述与可视化、关联分析、分类与回归、聚类分析。分类在数据挖掘中作为一种预测分析方法,通过对已知样本集的学习构造分类器或者分类模型,然后再用该分类器对大量实际数据分类,达到预测和归类的目的。

如果数据集中某些类的样本数量远远少于其他类别样本数量,则称为类别不平衡数据。现实世界中广泛存在着类别不平衡数据,例如医疗诊断、信用卡欺诈检测、网络入侵检测、文本分类等领域,在这些情况的处理过程中,少数类的识别准确率更为重要,其误分会带来更大的损失,而传统的分类算法假定数据集具有平衡类分布或者相等的错分代价,为保证分类总体精度,通常将少数类划分到多数类来保证整体的分类准确率,这将导致给出的实际分类结果不能令人满意。因此,

到稿日期:2013-01-27 返修日期:2013-04-03 本文受佛山市科技发展专项资金项目(2011AA100061),佛山市产学研专项资金项目(2012HC100272),佛山市教育局智能教育评价指标体系研究项目(DX20120220)资助。
胡小生(1978—),男,硕士,讲师,高级工程师,主要研究方向为机器学习、数据挖掘,E-mail:feihu@fosu.edu.cn;张润晶(1974—),女,高级工程师,主要研究方向为信息检索、信息安全;钟 勇(1970—),男,博士,教授,主要研究方向为信息安全、信息检索、云计算。

如何有效提高少数类的分类性能是数据挖掘领域的重要研究方向。

2 相关工作

鉴于类别不平衡数据分类的重要性，国内外学者进行了大量的研究，主要有以下3个方面：

1)基于数据层面，重构训练样本集。利用数据重抽样技术，改变数据分布，最简单的两种方法是随机过采样(over-sampling)和随机欠采样(under-sampling)，前者对少数类的样本进行复制，使数据集的类别样本达到平衡，后者则随机抽取多数类样本中的一部分达到同样的目的。两者各有优缺点：过采样通过不断复制少数类而使数据规模变大，使分类器学习到的决策域变小，从而容易导致过拟合的问题；欠采样方法通过抽取部分多数类样本使信息丢失严重。目前，很多学者提出改进的数据抽样方法[1-7]。为了避免随机过采样的不足，Chawla等人提出一种SMOTE(Synthetic Minority Over-sampling Technique)算法[1]，该算法通过采用少数类样本合成技术产生新的样本，可使少数类具有更大的泛化空间，但也可能导致分类器过于拟合。Kermanidis等人提出了SMOTE与Tomek links[2]相结合移出多数类冗余数据以平衡数据的方法[3]。Yen等人提出了一种基于聚类的抽样方法SBC(Under-sampling Based on Clustering)[4]，SBC通过聚类后簇内的多数类与少数类的比例确定抽样参数。

2)基于算法层面，修改已有的分类算法或者提出新算法。例如刘胥影等人[8]使用基于级联模型的分类器，该方法通过逐步筛掉多数类样本使数据集趋于平衡，在此过程中训练得到的一系列分类器通过集成方式对预测样本进行分类；针对不平衡数据导致SVM分类超平面偏移问题，Tang等人[9]通过包括代价敏感学习[10,11]、过采样、欠采样等方法解决，Ertekin等人[12]提出一种基于支持向量机(Support Vector Machine，SVM)的主动学习方法，该方法选择最富信息的“未见过”的训练样例，重新训练SVM。

3)将1)和2)有机结合。从数据集和算法两方面同时着手进行调整，这种综合方案可有效发挥前两种方案的优点，无论从理论分析还是从实践上看这种方案都表现出较强的优越性。例如井小沛等人[13]针对入侵检测过程中的不平衡类样本造成的分类超平面偏移现象，在对不同类别样本聚类之后，采用基于修正核函数的支持向量机方法，有效提高了入侵检测的准确率。李雄飞等人[14]提出了不平衡数据学习算法PCBoost，其利用数据合成方法添加少数类样例，平衡训练信息，在弱学习算法的下一次迭代时，通过权值更新过程更多关注错分的困难样例，同时修正扰动数据，消除错误添加的合成样例，该算法有效扩展了少数类边界，而且能够更快地收敛。

本文提出一种两层聚类的不平衡数据挖掘算法，旨在提高少数类的分类精度。首先对初始训练数据集进行基于聚类的欠采样，以在构造出新的平衡训练集的同时，又能防止随机欠采样所导致的重要判别性样本丢失问题；随后，对参与训练的平衡数据集，采用K均值聚类与C4.5决策树算法相结合的级联方法，K均值聚类算法将数据集划分成K个子集，用C4.5决策树算法分别在K个子集上学习得到决策边界；最后，在分类阶段，对预测样本，先将其划分到某个子集中，之后用此子集中的分类模型对样本进行预测分类。

由于篇幅关系，本文只考虑两类别不平衡的分类问题，两类别情况下，通常将少数类称为正类，多数类为负类，类别标签取值分别为{+1，-1}。当数据集中有两种以上类别时，指定其中一类别的样例为正类，合并其余类别的样例为负类。

3 基于聚类的欠采样方法

对于不平衡数据集，对负类样本进行过采样和在正类进行欠采样均能改变数据分布，使数据达到平衡，但是它们也有缺点，即过采样容易导致过度拟合问题，欠采样会引起信息丢失。为了弥补随机欠采样过程中可能丢失一些有用信息样本的缺点，抽取最富有代表性的训练样本，本文提出了对负类样本进行基于聚类的欠采样方法。

聚类是一种无监督学习方法，按照划分思想以簇内相似、簇间相异原则通过迭代把数据划分到不同的簇中，以求目标函数最小化，从而使生成的簇尽可能地紧凑和独立。通常相似性以欧式空间距离作为测度，样本之间距离近，则相似度高。

数据集$X=\{x_i|i=1,\cdots,p\}$和正整数K，其中x_i是m维向量。寻找映射$f:X\to\{1,\cdots,K\}$使数据集中每个样本x_i映射到某簇$j(1\leqslant j\leqslant K)$中，并使目标准则函数$J$值最小：

$$J=\sum_{j=1}^{K}\sum_{i=1}^{N_j}\|x_i-c_j\|^2 \tag{1}$$

式中，N_j表示簇j的样本数，x_i表示簇j的某个样本，c_j是簇j的中心，对目标准则函数J求偏导数，使其等于0，可求得最优解：

$$c_j=\frac{1}{N_j}\sum_{i=1}^{N_j}x_i \tag{2}$$

由式(2)可知，最优解是各簇中心。各簇中心，也即聚类质心是目标准则函数的最优解，表明聚类质心是各个聚类簇中最富有代表性的样本，最能够表征一个聚类簇内样本的特征。

本方法首先提取正类样本的个数k，并以k为聚类的类数目对负类样本进行K-means聚类，得到k个负类聚类质心，再将k个聚类质心与所有正类样本一起组成一个新的平衡训练集用来训练分类器。整体算法如算法1所示。

算法1 聚类欠采样算法

输入：初始完整数据集D

输出：平衡的训练集S_{train}

1)S=pre_process()；

2)train[S]=随机取数据集S中的80%样本，test[S]=S中剩余的20%样本；

3)提取train[S]中正、负类样本集合S^+、S^-，计算S^+中正类样本数量k，令K=k；

4)对集合S^-的样本进行K均值聚类，得到K个不相交子集及其聚类质心；

5)取出K个聚类质心，记为集合S'；

6)$S_{train}=S^+\cup S'$。

算法1的实质是为了组成一个新的平衡数据集，对负类样本进行聚类压缩，由聚类的质心代表聚类的所有样本，此方法对大规模数据(正类样本数量比较大)取得了很好的效果，但是如果在正类样本数量很小的情况下仍然单纯使用此方法，将会导致输出的平衡训练集过小而难以到达理想的分类精度。针对此问题，提出一个将SMOTE与聚类欠采样相结合的方法，在过采样与欠采样之间寻找一个平衡点，SMOTE

通过生成合成样本来对正类样本进行过采样，对于每个正类样本 x，在其同类中查找 n 个近邻，根据上采样的倍率 N，在样本 x 和被选中的近邻样本之间进行随机插值，构成新的样本，如算法 2 所示。

算法 2 正类样本较小的不平衡数据聚类欠采样算法

输入：初始完整数据集 D，过采样倍率 N

输出：平衡的训练集 S_{train}

1)S=pre_process()；

2)train[S]=随机取数据集 S 中的 80%样本，test[S]=S 中剩余的 20%样本；

3)根据过采样倍率 N，使用 SMOTE 算法对 train[S]进行正类样本扩充得到集合 B，使正类样本规模达到预设值；

4)提取 B 中正、负类样本集合 S^+、S^-，计算 S^+ 中正类样本数量 k，令 K=k；

5)对集合 S^- 的样本 K 均值聚类，得到 K 个不相交子集及其聚类质心；

6)取出 K 个聚类质心，记为集合 S′；

7)$S_{train}=S^+ \cup S'$。

由于数据集中的属性一般有连续型和分类型两种，因此，在算法 1 与算法 2 中的第 1)步均须进行数据预处理 pre_process()过程，具体方法是：对于数据集中的分类型属性，采用二进制编码方式进行转换，将分类型的属性转换为若干个取值 0 和 1 的属性；对于连续型属性，为了消除由于量纲的不同所造成的影响，对输入数据进行最大最小归一化处理，将所有数据归一化到[0,1]之间，其公式如下所示：

$$a_j=\frac{a_j-\mathrm{Min}_A}{\mathrm{Max}_A-\mathrm{Min}_A} \tag{3}$$

式中，Max_A，Min_A 分别表示特征 A 上数据的最大值和最小值。

4 K 均值聚类与 C4.5 决策树算法级联方法

通过聚类的欠采样获得新的平衡训练集之后，使用一种 K 均值聚类与 C4.5 决策树算法相结合的级联分类方法，该方法分为两个阶段：训练阶段和分类阶段。

1. 训练阶段

此阶段，首先通过 K 均值聚类方法将平衡训练集划分为 K 个不相交的子集 C_1，C_2，C_3，…，C_k，然后，在每个子集 C_i（$1 \leqslant i \leqslant K$）上应用 C4.5 决策树算法，得到 K 个分类模型。K 值的确定与上节方法不同，即将初始数据集的类别总数赋给 K，K 均值聚类的目的是尽最大可能将每个类别的样本实例聚集到一个簇内，然而，如果在一个聚类簇内含有不同类别的样本，C4.5 决策树算法在每个聚集内分别进行训练，通过一序列的基于特征空间上的规则来提炼决策边界。

2. 分类阶段

此阶段又可再分为两个子阶段：选择阶段和分类阶段。在选择阶段，对每个分类预测样本计算其与已知的 K 个聚类质心的欧氏距离，找出离其最近距离的聚类，计算此最近距离的聚类的 C4.5 决策树规则；在分类阶段，将测试样例应用到离其最近距离的聚类上学习得到 C4.5 决策规则，最终得到分类结果，整个算法如下所示。

算法 3 K 均值聚类与 C4.5 级联分类算法

(1)训练阶段

输入：平衡的训练集 S_{train}，聚类的数量 K

输出：K 个子集及其聚类质心 r_j（$1 \leqslant j \leqslant K$）

1)在训练集合 S_{train} 中随机选择 k 个样本作为初始聚类中心；

2)对每个训练样例

①计算样本到每个聚类中心的聚类；

②选择最近邻的聚类中心，并将该样本加入到此聚类中心所在的聚类；

③根据新产生的聚类，重新计算该聚类的聚类中心；

3)重复步骤 2)，直到每个聚类不再发生变化；

4)提取最终的 K 个聚类子集及其聚类质心 r_i（$1 \leqslant i \leqslant K$）。

(2)选择阶段

输入：测试样例 Z_i，i=1,2,3,…,N

输出：离测试样例 Z_i 距离最近的聚类 C_i

1)对每个训练样例 Z_i

①计算欧氏距离 $D(Z_i,r_j)$，j=1,2,…,k，找出离 Z_i 距离最近的聚类 C_i；

② 在聚类 C_i 上应用 C4.5 决策树算法，得到决策模型。

(3)分类阶段

输入：测试样例 Z_i，i=1,2,3,…,N

输出：分类结果

1)将测试样例 Z_i 应用到聚类 C_i 上学习得到基于 C4.5 的分类决策模型；

2)由分类决策模型对 Z_i 进行分类；

3)将样例 Z_i 加入到 C_i 中，并且重新计算聚类质心 r_i。

5 实验结果与分析

为了检验本算法的有效性，利用具有不同实际应用背景的 UCI 数据集进行验证。实验设计思路如下：首先选择 UCI 数据集中的类别不平衡数据，与目前处理类别不平衡的主要方法进行对比，包括基于数据抽样的集成方法：SMOTEBoost 和 RUSBoost，并与本文所涉及到的 C4.5 决策树算法和基于聚类的分类方法进行对比，以验证本算法的有效性。对于分类器的分类评价标准，选择更适合类别不平衡数据的分类标准。

5.1 UCI 数据

本文选取 sonar、breast-w、vehicle、artificial、pendigits、car、letter 和 page-blocks 共 8 组 UCI 数据进行测试，表 1 是用于实验的数据信息，包括数据集大小、少数类和多数类样例数量和比例等。对于含有多个类别的数据，将其中的一类作为少数类，合并剩下的各个类别为一个整体作为多数类。例如，将 page-blocks 的类别 5 作为少数类，合并其他的类作为多数类。

表 1 UCI 数据集信息

数据集	样例数目	少数类	多数类	不平衡度
sonar	208	97	111	1.14
breast-w	699	241	458	1.90
vehicle	846	199	647	3.25
artificial	5109	708	4401	6.21
pendigits	10992	1055	9937	9.42
car	1728	65	1663	25.58
letter	20000	734	19266	26.25
page-blocks	5473	115	5358	46.59

通常情况下，将不平衡度[1.5,3.5)、[3.5,9.5)、[9.5,+∞)分别称为低度不平衡范围、中度不平衡范围、高度不平衡范围。表 1 所列数据集具有不同不平衡比例，其中 sonar 数据集基本是平衡数据集，选取该数据集的目的是验证本文算法对一般数据集的有效性。

5.2 评价标准

在传统的分类学习中，一般采用分类精度(分类正确的样本个数站总样本个数的百分比)作为评价指标，然而对于不平衡数据集，这一指标实际意义不大，因为它反映的是多数类样本的分类测试结果。因此，为了充分反映少数类的分类情况，需要提出更为合理的评价标准，目前常用的类别不平衡评价标准是建立在混淆矩阵(见表 2)基础上的评价指标：查全率(Recall)、查准率(Precision)、F-measure、AUC(Area Under Roc Curve)值等[15]。

表 2 两类混淆矩阵

	预测正类	预测负类
实际正类	TP	FN
实际负类	FP	TN

在某些应用中，人们更加关注少数类样本的分类性能，F-measure 就是用于衡量少数类分类性能的指标。F-measure 是 Recall 和 Precision 的调和均值，其取值接近两者的较小者，因此，较大 F-measure 值表示 Recall 和 Precision 都较大。

$$F\text{-}measure=\frac{2\times Recall\times Precision}{Recall+Precision} \tag{4}$$

式中，$Recall=\frac{TP}{TP+FN}$，$Precision=\frac{TP}{TP+FP}$。

AUC 是 ROC(Receiver Operating Characteristic)曲线下方图面积，研究表明 AUC 能够全面描述分类器在不同判决阈值时的分类性能。AUC 值总是在 0 和 1 之间，其越大说明分类性能越好。对于两类别数据分类，AUC 值计算公式如下：

$$AUC=\frac{TP_{rate}+TN_{rate}}{2} \tag{5}$$

式中，正类分类准确率 $TP_{rate}=\frac{TP}{TP+FN}$，负类分类准确率 $TN_{rate}=\frac{TN}{TN+FP}$。

本文采用 F-measure 和 AUC 值作为评价标准。

5.3 实验结果

为方便比较，实验中对数据采用五折交叉验证方式。为保证数据在进行分组过程中不平衡度保持一致，采用分层采样，将数据集中的少数类和多数类样本分别随机分为 5 等份，两两随机组合得到 5 个不平衡度一致的子集，将其中 4 个子集作为训练集，1 个子集作为测试集，重复 5 次，以平均值作为最终的分类结果。

表 3 和表 4 分别给出在 8 个不同的不平衡数据集上本文两层聚类的算法与 SMOTEBoost、RUSBoost、C4.5 和基于聚类的分类方法(ClassificiationviaClustering，CC)的比较结果。为了直观比较，SMOTBoost、RUSBoost 算法的基分类器均采样 C4.5，CC 算法使用经过基于聚类的欠采样处理后的平衡训练集进行模型训练。

表 3 不同算法的 F-measure 值比较

数据集	SMOTEBoost	RUSBoost	C4.5	CC	本文算法
sonar	0.705	0.785	0.706	0.763	**0.883**
breast-w	0.94	0.941	0.913	0.979	**0.984**
vehicle	0.924	0.901	0.87	0.618	**0.945**
artificial	0.639	0.654	0.569	0.39	**0.672**
pendigits	**0.993**	0.942	0.948	0.71	0.984
car	**0.897**	0.612	0.614	0.714	0.88
letter	0.885	0.718	0.88	0.697	**0.952**
page-blocks	0.65	0.55	0.629	0.36	**0.714**
平均值	0.829	0.763	0.766	0.654	**0.877**

表 4 不同算法的 AUC 值比较

数据集	SMOTEBoost	RUSBoost	C4.5	CC	本文算法
sonar	0.746	0.798	0.712	0.779	**0.894**
breast-w	0.959	0.959	0.94	0.984	**0.989**
vehicle	0.963	0.947	0.936	0.696	**0.974**
artificial	0.894	0.871	0.894	**0.901**	0.888
pendigits	**0.998**	0.98	0.99	0.824	0.997
car	**0.991**	0.954	0.972	0.668	**0.991**
letter	0.969	0.944	0.992	0.705	**0.997**
page-blocks	0.981	0.94	**0.987**	0.783	0.985
平均值	0.938	0.924	0.928	0.793	**0.964**

根据 F-measure 的定义，其取值接近于 Recall 和 Precision 的较小值，因此，如果算法获得了较大的 F-measure 值，则说明算法的 Recall 和 Precision 都是较大值，也即算法的少数类查全率和查准率都较高。从表 3 中可以看出，本文算法在所测试的 8 个 UCI 数据集中，除了在 pendigits 和 car 两个数据集上的 F-measure 度量指标略低于 SMOTEBoost 外，在其他数据集上均优于其他算法。通过计算这 8 组 UCI 数据在不同算法条件下的 F-measure 平均值，也可以看出所提方法要明显优于其他算法，在平均性能上比次优算法有 5.8% 的性能提升，与 C4.5 决策树算法进行比较有 14.5%的提升，与基于聚类的分类算法相比有 34%的提升。

AUC 定义说明，如果算法获得较大的 AUC 值，则说明算法在整体上分类性能好。从表 4 中可以看出，对于 AUC 值的度量，本文算法在 sonar、breast-w、vehicle、car、letter、page-blocks 上均不低于其他算法，在 artificial 和 pendigits 数据集上也仅仅略低于最优算法。

结合表 1、表 3、表 4 可以看出，随着不平衡度的提高，在少数类的分类评价指标上，RUSBoost 算法分类效果要低于 SMOTEBoost 等其他算法，特别是在 car、letter 和 page-blocks 上，RUSBoost 算法的 F-measure 值最低，主要原因是随机欠采样的数据抽样方法导致丢失许多有用的负类样例信息，而且，少数类样本总数量过少，特别是 car、page-blocks 数据集上少数类数量分别只有 65 和 115，这也是一个重要因素。与之对应，本文算法采用基于聚类的欠采样方法，但是在所测试的数据集上少数类分类性能在整体上不低于相比较的 SMOTEBoost 等算法，特别是在不平衡度最高的 letter 和 page-blocks 数据集上，F-measure 度量指标表现优异，这说明基于聚类的数据欠采样方法能够有效避免随机欠采样方法的不足，作为一种欠采样方法，其对少数类样本分类效果明显。将基于聚类的分类算法 CC 与本文算法进行比较，在两者采用相同的训练集情况下，在 F-measure 和 AUC 两个度量指标上评价，CC 算法分类效果不佳，究其原因，基于聚类的分类方法仅仅考虑样本属性之间的关联关系，是基于无监督学习方式，忽略了作为监督判别性的类别标签信息；而本文所提出的基于 K 均值聚类与 C4.5 决策树算法相级联的方法，将无监督学习方式与监督学习方式相结合，充分考虑了属性之间的关联性以及类别标签的监督判别性，最终分类效果提升明显。

结束语 本文提出两层聚类的级联分类算法来处理类别不平衡数据问题，首先，对多数类样本进行基于聚类的欠采样，为后续的分类器提供新的平衡训练集；然后，在新训练集上采用 K 均值聚类与 C4.5 决策树算法相结合的级联分类算法，将无监督学习方式的聚类与监督学习方式的决策树算法相融合，通过两层聚类的方式，在数据层面和算法层面分别进行调整，以便更好地处理类别不平衡的分类问题。实验结果

表明，该算法能够有效提高小类样本以及整体分类性能。

由于数据集本身的多样性和复杂性，样本的分布也呈现多样性，如果能够准确估计多数类样本的潜在分布，根据不同的分布采取不同的聚类方式以及智能采样技术，将会显著提高分类性能。此外，考虑对数据集用不同聚类方式进行多次聚类，然后通过某种策略对聚类结果进行融合，进一步提高少数类和数据集的整体分类性能是今后需要进一步研究的内容。

参 考 文 献

[1] Chawla N V, Bowyer K, Hall L, et al. SMOTE: Synthetic Minority Over-sampling Technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357

[2] Tomek I. Two modifications of CNN[J]. IEEE Transaction on Systems, Man and Communications, 1976, 26(1): 769-772

[3] Kermanidis K, Maragoundakis K, Fakotakis N, et al. Learning greek verb complements: addressing the class imbalance[C]// Proceedings of the 20th International Conference on Computational Linguistics. Geneva, Switzerland, 2004: 1065-1071

[4] Yen Show-jane, Lee Yue-shi. Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset[C]// Proceedings of Intelligent Control and Automation, Series: Lecture Notes in Control and Information Sciences. Berlin/Heidelberg: Springer, 2006: 731-740

[5] 蒋盛益，苗邦，余雯. 基于一趟聚类的不平衡数据下抽样算法[J]. 小型微型计算机系统，2012，33(2)：232-236

[6] 李雄飞，李军，屈成伟，等. 数据挖掘中平衡偏斜训练集的方法研究[J]. 计算机研究与发展，2012，49(2)：346-353

[7] 韩敏，朱新荣. 不平衡数据分类的混合方法[J]. 控制理论与应用，2011，28(10)：1485-1489

[8] 刘胥影，吴建鑫，周志华. 一种基于级联模型的类别不平衡数据分类方法[J]. 南京大学学报：自然科学版，2006，42(2)：148-155

[9] Tang Y, Zhang Y Q, Chawla N V, et al. SVMs modeling for highly imbalanced classifications[J]. IEEE Transaction on Systems, Man, and Cybernetics, Part B: Cybernetics, 2009, 39(1): 281-288

[10] 凌晓峰，Sheng V S. 代价敏感分类器的比较研究[J]. 计算机学报，2007，30(8)：1203-1212

[11] 翟云，杨炳儒，曲武. 不平衡类数据挖掘研究综述[J]. 计算机科学，2010，37(10)：27-32

[12] Ertekin S, Huang J, Bottou L, et al. Learning on the border: active learning in imbalanced data classification[C]// Proceedings of the ACM Conference on Information and Knowledge Management. Lisbon, Portugal, 2007: 127-136

[13] 井小沛，汪厚祥，聂凯. 一基于修正核函数 SVM 的网络入侵检测[J]. 系统工程与电子技术，2012，34(5)：1036-1039

[14] 李雄飞，李军，董元方，等. 一种新的不平衡数据学习算法 PC-Boost[J]. 计算机学报，2012，35(2)：202-209

[15] 林智勇，郝志峰，杨晓伟. 若干评价准则对不平衡数据学习的影响[J]. 华南理工大学学报：自然科学版，2010，4(38)：126-135

(上接第 260 页)

[2] Huang C L, Wang C J. A GA-based feature selection and parameters optimization for support vector machines[J]. Expert Systems with Applications, 2006, 35(4): 231-240

[3] 王世卿，曹彦. 基于遗传算法和支持向量机的特征选择研究[J]. 计算机工程与设计，2010，31(18)：4088-4092

[4] Tian Jin, Li Min-qiang, Chen Fu-zan. Dual-population based coevolutionary algorithm for designing RBFNN with feature selection[J]. Expert Systems with Applications, 2010, 37(10): 6904-6918

[5] Oja E. Subspace methods of pattern recognition[M]. New York: Research Studies Press, 1983

[6] Qiu Guo-ping, Fang Jian-zhong. Classification in an informative sample subspace[J]. Pattern Recognition, 2008, 41(3): 949-960

[7] Cevikalp H, Neamtu M, Barkana A. Kernel common vector method: A Novel nonlinear subspace classifier for pattern recognition[J]. IEEE Transactions On Systems, Man and Cybernetics, 2007, 37(4): 937-951

[8] Wen Ying. An improved discriminative common vectors and support vector machine based face recognition approach[J]. Expert Systems with Applications, 2012, 39(4): 4628-4632

[9] Sakano H, Mukawa N, Nakamura T. Kernel mutual subspace method and its application for object recognition[J]. Electronics and Communications in Japan (Part II: Electronics), 2005, 88(6): 45-53

[10] Kazuhiro F, Osamu Y. The kernel orthogonal mutual subspace method and its application to 3D object recognition[C]// 8th Asian Conference on Computer Vision. Tokyo, 2007: 467-476

[11] Zhu Mei-hong, Li Ai-hua. Random subspace method for improving performance of credit cardholder classification[C]// Modeling Risk Management for Resources and Environment in China. Berlin, 2011: 257-264

[12] Watanabe S, Lambert P F, Kulikowski C A, et al. Evaluation and selection of variables in pattern recognition[C]// Computer and Information Sciences II. New York, 1967

[13] Zhang Peng, Peng Jing, Domeniconi C. Kernel pooled local subspaces for classification[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2005, 35(3): 489-502

[14] Kitamura T, Takeuchi S, Abe S, et al. Subspace-based support vector machines for pattern classification[J]. Neural Networks, 2009, 22(5/6): 558-567

[15] Wang Li-po, Zhou N, Chu Feng. A General Wrapper Approach to Selection of Class-Dependent Features[J]. IEEE Transactions on Neural Networks, 2008, 19(7): 1267-1278

[16] Vapnik V N. The nature of statistical learning theory [M]. Springer, 1999

[17] Shin H, Cho S. Invariance of neighborhood relation under input space to feature space mapping[J]. Pattern Recognition Letters, 2005, 26(6): 707-718

[18] 翟俊海，李胜杰，王熙照. 基于粗糙集技术的压缩近邻规则[J]. 计算机科学，2012，39(2)：236-239

[19] Wang Ji-gang, Neskovic P, Cooper L N. Training data selection for support vector machines[C]// Advances in Natural Computation. Changsha, 2005: 554-564

[20] 周晓飞，姜文瀚，杨静宇. 基于子空间样本选择的最近凸包分类器[J]. 计算机工程，2008，34(12)：167-168

[21] 姜文瀚，周晓飞，杨静宇. 核子类凸包样本选择方法及其 SVM 应用[J]. 计算机工程，2008，34(16)：212-214

[22] Zhou Xiao-fei, Jiang Wen-han, Tian Ying-jie, et al. Kernel subclass convex hull sample selection method for svm on face recognition[J]. Neurocomputing, 2010, 10-12(73): 2234-2246