基于马氏抽样的 SVM 非平衡数据分类算法的泛化性能研究

徐 婕,贺美美
(湖北大学计算机与信息工程学院,湖北武汉 430062)

摘 要: 本文将样本为独立同分布的情形减弱为一致遍历马氏链的情形去研究了非平衡数据分类算法的泛化性能,提出了基于马氏抽样的 SVM 非平衡数据分类算法、基于马氏抽样的 EDSVM 非平衡数据分类算法和基于马氏抽样的 SVM-WKNN 非平衡数据分类算法. 并用 UCI 数据库中的 10 个实际不平衡数据集进行数值实验,实验结果表明基于马氏抽样的上述三种算法的错分率均比基于随机抽样的对应算法的错分率要低,且上述三种算法中,基于马氏抽样的 SVM-WKNN 非平衡数据分类算法的泛化性能最好.

关键词: 马氏抽样; 支持向量机; k 近邻算法; 非平衡数据
中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112 (2018)11-2660-11
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2018.11.013

Research on the Generalization Performance of SVM Imbalanced Data Classification Algorithm Based on Markov Sampling

XU Jie, HE Mei-mei
(*School of Computer Science and Information Engineering, Hubei University, Wuhan, Hubei, 430062, China*)

Abstract: This paper changes the assumption that samples are independent and identically distributed to that samples are uniformly ergodic Markov chains, which make it convenient for us to study the generalization performance of the imbalanced data classification algorithm, and SVM imbalanced data classification algorithm based on Markov sampling, EDSVM imbalanced data classification algorithm based on Markov sampling and SVM-WKNN imbalanced data classification algorithm based on Markov sampling are proposed. The numerical experiments of ten actual imbalanced datasets in the UCI database show that the misclassification rate of the above algorithm based on Markov sampling is lower than that of the corresponding algorithm based on random sampling, and the above three algorithms, SVM-WKNN imbalanced data classification algorithm based on Markov sampling has the best generalization performance.

Key words: Markov sampling; support vector machine; k-nearest neighbor; imbalanced data

1 引言

不管是在学术界还是工业界,对于不平衡数据的研究已经吸引了越来越多学者的关注[1]. 不平衡数据[2]指的是数据集中样本类别是不均衡的,对于标准的二分类问题来说,样本数量多的类称为多数类,样本数量少的类称为少数类. 在实际应用中,不平衡数据出现在许多领域,如医疗诊断[3]、文本分类[4]、信用卡欺诈检测[5]、网络攻击识别[6]、机械故障诊断[7]等. 目前,解决非平衡数据分类问题常用的方法有两种[8]:第一种方法是从数据层面的角度出发,主要方法为重抽样[9],其基本思想是既然样本是不平衡的,那么可以通过某种抽样策略,让样本相对均衡一些;第二种方法是从算法的角度出发,考虑不同错分类情况下的代价差异[10],从而对算法进行优化,使算法在对不平衡数据分类时也能有较好的效果. 近年来,Japkowicz 等人[11]证明了在后向传播神经网络、决策树 C4.5、SVM 几类算法中,SVM 对非平衡数据的影响较为不敏感. 李蓉等人[12]提出了一种 KSVM 算法,通过结合 k 近邻算法和 SVM 各自的优点,在分类阶段计算待测样本与两类支持向

收稿日期:2017-09-04;修回日期:2018-02-05;责任编辑:蓝红杰
基金项目:国家自然科学基金(No.61370002, No.61403132, No.61772011)

量代表点的距离,如果距离大于给定阈值,则用 SVM 来分类,否则以每类支持向量代表点作为训练样本的 k 近邻算法来分类. 王超学等人[13]提出了一种改进型算法—EDSVM(Euclidean Distance NN & Support Vector Machine),该算法对距离分类超平面较远的点使用 SVM 进行分类,对于距离分类超平面较近的点则以全部的支持向量作为训练样本的 k 近邻算法进行分类,提高了样本分类的正确率.

以上算法都是建立在样本是独立同分布的假设下,然而样本是独立同分布的假设无论是在理论上还是在实际应用中都是非常强的,并且很多机器学习应用中产生的数据并不服从独立同分布,于是将独立同分布的情形减弱为非独立的情形,如一致遍历马氏链等,已有一些学者做了相关研究. 如:Xu 等人[14]证明了基于马氏抽样的 SVM 算法的泛化性能比基于随机抽样的 SVM 算法的泛化性能要好. Zou 等人[15]提出了基于马氏抽样的正则化回归算法. Smale 等人[16]将马氏抽样运用到了在线学习中. 本文考虑到样本是独立同分布的假设在很多现实问题中不能满足,且随机抽样的效率比较低,所以引入了新的马氏抽样方法来研究基于一致遍历马氏链样本的 SVM 非平衡数据分类算法的泛化性能,给出了三种基于马氏抽样的非平衡数据分类算法,分别是:基于马氏抽样的 SVM 非平衡数据类算法、基于马氏抽样的 EDSVM 非平衡数据分类算法和基于马氏抽样的 SVM-WKNN 非平衡数据分类算法,三种算法都是先通过马氏抽样获取算法所需的训练样本,再选用不同的算法对数据进行分类. 基于基准库数据集实验结果表明,上述基于马氏抽样分类算法的泛化性能总是优于基于随机抽样分类算法的泛化性能,并且将上述三种算法进行泛化性能的对比后发现,基于马氏抽样的 SVM-WKNN 非平衡数据分类算法的泛化性能最好.

2　相关概念介绍

2.1　支持向量机

支持向量机[17](Support Vector Machine,简称 SVM)是一种应用广泛的分类算法,通过使结构风险最小化来提高泛化能力,实现经验风险和置信范围的最小化,其基本模型定义为特征空间上间隔最大的线性分类器. 给定特征空间上的训练集 $Z=\{z_1=(x_1,y_1),z_2=(x_2,y_2),\cdots z_N=(x_N,y_N)\}$,其中 $x_i\in R^n$,$y_i\in\{+1,-1\}$,$i=1,2,\cdots N$. 相应的分类决策函数为:$f(x)=\mathrm{sign}(wx+b)$,其中:

$$\mathrm{sign}(f(x))=\begin{cases}-1, & f(x)<0\\ +1, & f(x)\geq 0\end{cases}.$$

构造拉格朗日函数可以求得:$w=\sum_{i=1}^{N}\alpha_i^* y_i x_i$,$b=y_i-\sum_{i=1}^{N}\alpha_i^* y_i(x_i\cdot x_j)$,其中 a_i^* 是拉格朗日系数. 若样本线性不可分,可以使用核函数,将输入空间映射到高维特征空间,从而使得原本线性不可分的样本可以在高维特征空间可分,此时所求目标函数为:$\min\frac{1}{2}\|w\|^2+C\sum_{i=1}^{N}\xi_i$,使得 $y_i(w\cdot x_i+b)\geq 1-\xi_i,\xi_i\geq 0,i=1,2,\cdots N$,其中 ξ_i 为松弛变量,C 为惩罚因子.

通过 SVM 训练模型获得的支持向量决定了最终的分类超平面,这不但可以抓住关键样本,“剔除”大量冗余样本,而且表明了 SVM 具有较好的“鲁棒”性. 但 SVM 对位于两个类别的边界区域或重叠区域的样本点进行分类时,会存在一定的分类错误,即 SVM 算法的分类错误一般都发生在最优分类超平面附近.

2.2　k 近邻算法

k 近邻算法(k-Nearest Neighbor,简称 KNN)是一种基于距离的算法,它没有明显的前期训练过程. 具体过程为:每来一个未知的样本点,就在此样本点附近找 k 个与之最近的点进行投票,统计投票结果,某类数量最多,就把这个测试样本判定为该类.

KNN 属于懒惰学习方法[13],若测试样本比较大时,因为要扫描全部训练样本并计算距离所以计算量会较大. 由于 KNN 方法主要靠周围有限的邻近样本,而不是靠判别类域的方法来确定所属类别,因此对于类域的交叉或重叠较多的待分样本集来说,KNN 方法较其他方法更为适合.

2.3　EDSVM 算法

EDSVM[13](Euclidean Distance NN & Support Vector Machine)是一种改进的 SVM 和 KNN 组合的不平衡数据分类算法,主要思想是:当待测样本与 SVM 最优分类超平面的距离大于一定值时,用 SVM 进行分类;当待测样本离 SVM 最优分类超平面的距离小于一定值时,以所有的支持向量作为测试样本的近邻样本,用 KNN 进行分类.

EDSVM 算法利用了支持向量能够代表训练样本集的含义,提高了分类器的分类精度. 但 EDSVM 没有考虑到用 KNN 对非平衡数据带来的影响,KNN 对每个样本都赋予了相同的权重,当样本分布不均衡时,可能会导致测试数据被分到样本容量较大的那一类,造成错误分类的情况.

2.4　马尔可夫链

经典的机器学习建立在独立同分布的假设下,然而实际应用中很多模型产生的样本是自然涌现的而非独立同分布,而马尔可夫链更符合机器学习中数据的分布情况,本文的马尔可夫链为一致遍历马氏链,下面是一致遍历马氏链的概念.

记(Z,S)是一个可测空间,马尔可夫链包括一个随

机变量$\{Z_t\}_{t\geq 1}$与一系列转移概率测度$\Pr^n(A|z_i)$,$A\in S$,$z_i\in Z$,则$\Pr^n(A|z_i):=\Pr\{Z_{n+i}\in A|Z_j,j<i,Z_i=z_i\}$.其中$S=\{z_i\}_{i=1}^m$,$\Pr^n(A|z_i)$记为转移概率,从$i$时刻的初始状态$z_i$开始,将在$n$步迭代后的状态$Z_{n+i}$的概率记为集合$A$.马尔可夫链的特性是:"给出当前状态,那么马尔可夫链的将来和过去状态是独立的".即:$\Pr^n(A|z_i)=\Pr\{Z_{n+i}\in A|Z_i=z_i\}$.给定概率空间$(Z,S)$上的两个概率$v_1$和$v_2$,这两个测度$v_1$和$v_2$的全变量差定义为:$\|v_1-v_2\|_{TV}=sup_{A\in S}|v_1(A)-v_2(A)|$.

因此,得到一致遍历马氏链的定义[18].

定义1:如果$\{Z_t\}_{t\geq 1}$满足条件:$\exists\gamma,0<\gamma<\infty,0<\rho<1$,且$\|\Pr^k(\cdot|z)-\pi(\cdot)\|_{TV}\leq\gamma\rho^k$,$\forall k\geq 1,k\in\mathbb{N}$,其中$\pi(\cdot)$是$\{Z_t\}_{t\geq 1}$的平稳分布,则$\{Z_t\}_{t\geq 1}$为一致遍历马氏链.

3 三种基于马氏抽样的非平衡数据分类算法

基于马氏抽样的非平衡数据分类算法通过构建一致遍历马氏链,使得多次抽样后马氏链达到稳定分布,此后的抽样都可以认为是根据样本的分布进行的抽样.三种基于马氏抽样的非平衡数据分类算法的主要思想是:利用马氏抽样获取算法所需的训练样本,再使用不同的算法对样本进行分类.

算法1是基于马氏抽样的SVM算法对非平衡数据进行分类的过程.文献[14]对样本是平衡情况下SVM的泛化性能进行了研究,算法1则将基于马氏抽样的SVM算法推广运用到了非平衡数据的分类上,记$\{Z_j\}_{j=1}^N$为具有一致遍历马氏链的样本序列,损失函数$l(f,z)=(f(x)-y)^2$,α为连续拒绝数,q为常数,则基于马氏抽样的SVM非平衡数据分类算法的具体步骤如下:

算法1 基于马氏抽样的SVM非平衡数据分类算法

输入:训练集D和测试集T

1. 从训练集D中随机取出N个训练样本$\{Z_i\}_{i=1}^N$,使用SVM训练这些样本得到一个初始模型f,令$N_+=0$,$N_-=0$,N是训练样本的个数,N_+、N_-分别代表$\{Z_i\}_{i=1}^N$中正负类样本的个数.
2. 从训练集D中随机抽取一个样本为当前样本Z_t,若Z_t属于正类,则$N_+=N_++1$,否则$N_-=N_-+1$.
3. 从训练集D中随机抽取另外一个样本为候选样本Z_*.
4. 计算候选样本与当前样本在$e^{-l(f,z)}$的比率P,$P=e^{-l(f,z_*)}/e^{-l(f,z_t)}$.
5. 若$P=1$、$y_t=-1$且$y_*=-1$(或$P=1$、$y_t=1$且$y_*=1$)则接受Z_*,转移概率$P'=\min\{1,e^{-y_*f}/e^{-y_tf}\}$;若$P=1$且$y_ty_*=-1$,或$P<1$则接受$Z_*$,转移概率为$P$;如果连续$\alpha$个候选样本$Z_*$不能被接受,则转移概率$P''=\min\{1,qP\}$,此时用$P''$接受$Z_*$,令$Z_{t+1}=Z_*$,若$Z_t$属于正类,则$N_+=N_++1$,否则$N_-=N_-+1$.
6. 若$N_++N_-<N$,则返回第3步,否则停止.于是得到具有马尔可夫链性质的训练样本集$\{Z_j\}_{j=1}^N$.
7. 对训练样本集$\{Z_j\}_{j=1}^N$利用SVM算法训练得到决策函数模型$f(x)$,使$f(x)$对测试集T中的样本进行分类,统计测试集T中错分样本的个数.

输出:算法对测试集的错分率

评注1:马氏抽样属于关联采样,关联采样的效率会更高一些.因为在足够多次马氏抽样之后,马氏样本序列会达到平稳分布,这之后的抽样会体现出样本分布的信息,所以马氏抽样利用样本间的相关性对数据不完全抽样所导致的模型估计置信区间的波动进行了修正.根据文献[14]的研究,为减少抽样时间,在马氏抽样阶段引入了连续拒绝数α和常数q.

针对SVM对最优分类超平面附近的样本容易错分的问题,文献[13]给出了EDSVM非平衡数据分类算法,它结合SVM与KNN两种算法的优点,提升了分类效果.算法2在EDSVM中加入马氏抽样,进而获取具有一致遍历马氏链性质的训练样本,然后给定经验阈值ε,计算样本与最优分类超平面的距离d,比较d与ε的大小,选取不同算法对样本进行分类,具体步骤如下:

算法2 基于马氏抽样的EDSVM非平衡数据分类算法

输入:训练集D和测试集T

步骤1~步骤6与算法1相同

7. 对训练样本集$\{Z_j\}_{j=1}^N$利用SVM训练得到决策函数模型$f(x)$及支持向量集合SV.
8. 计算测试集T中的每一个待分类样本x与SVM决策函数$f(x)$的距离,即获得x与最优分类超平面的距离$d=f(x)/\|w\|$.
9. 给定阈值ε
   (1)若$|d|\geq\varepsilon$,用SVM的决策函数$f(x)$进行分类;
   (2)若$|d|<\varepsilon$,用KNN进行分类.
10. 统计测试集T中错分样本的个数,计算出算法的错分率.

输出:算法对测试集的错分率

由于EDSVM中的KNN对非平衡数据分类易造成错分,进而提出加入权重因子的SVM-WKNN算法对其进行改进,尽可能的避免了非均衡数据对分类带来的影响.基于马氏抽样的SVM-WKNN非平衡数据分类算法的主要思想是通过马氏抽样获取训练样本,然后根据样本所处的位置,选择将使用SVM还是使用加入权重因子的k近邻算法(WKNN)进行分类,具体步骤如下:

算法3 基于马氏抽样的SVM-WKNN非平衡数据分类算法

输入:训练集D和测试集T

步骤1~步骤8与算法2相同

9. 给定阈值ε

(1)若$|d|\geq\varepsilon$,用SVM的决策函数$f(x)$进行分类;

(2)若$|d|<\varepsilon$,用WKNN算法进行分类.即计算待分类样本x与支持

向量集合 SV 中每一个支持向量的距离,选出距离最近的 k 个支持向量,统计 k 个支持向量的类别,记其中正类支持向量个数和负类支持向量个数为:p_num 和 n_num,统计训练样本集 $\{Z_j\}_{j=1}^{N}$ 中正负类的数目 N_+ 和 N_-。若 $p\_num \times N_+ / N > n\_num \times N_- / N$,则待分类样本 x 为正类,否则为负类.

10. 统计测试集 T 中错分样本的个数,计算出算法的错分率.

输出:算法对测试集的错分率

评注 2:通过对 KNN 算法增加权重因子来处理样本不均衡对算法带来的影响,权重因子由某类样本的训练样本数占总训练样本数的比重求得.

4　数值实验

4.1　实验数据

实验选取的数据集是来自 UCI 数据库[19]的 10 个非平衡数据集,这些数据集都属于二分类,具体信息如表 1 所示. 在实验进行之前对于数据集 Diabetes、Australian、German、Census_income、Covtype 和 Bin_connect4 进行了数据归一化处理.

表 1　10 个实际数据集

数据集	样本维度	训练集		测试集		均衡比例
		正类	负类	正类	负类	
Skin_nonskin	3	33812	129559	17047	64639	1:3.82
Cod-rnd	8	108641	217069	54214	108641	1:2
Diabetes	8	145	239	123	261	1:2.10
Shuttle	9	34108	9392	11478	3022	3.67:1
Pageblocks	10	3266	383	1647	177	8.77:1
Australian	14	151	189	156	194	1:1.25
German	24	364	136	336	164	2.33:1
Census_income	41	12382	187141	6186	93576	1:15.12
Covtype	54	249822	185937	83596	61657	1.35:1
Bin_connect4	126	33312	17356	11161	5728	1.93:1

实验是在 MATLAB 环境下编程实现,算法采用高斯径向基核函数来处理高维数据,其中惩罚参数 C 和高斯核参数 g 通过交叉验证获取最佳取值,算法运行过程中 $C=10$,$g=0.3$,参数阈值 $\varepsilon \in [0,1]$,当 ε 取 0 时,算法为 SVM. 实验中 EDSVM 和 SVM-WKNN 这两个算法需要用到 ε,ε 将采用实验分析的方法进行设定,具体如表 2 所示(以 SVM-WKNN 算法在非平衡数据集 Cod-rnd 上的分类为例).

表 2　参数 ε 与分类性能的关系

评价标准 / 参数 ε	错分率		p-value
	随机抽样	马氏抽样	
$\varepsilon=1/2$	0.1190 ±0.0265	0.0812 ±0.0051	2.6304e-13
$\varepsilon=1/4$	0.0851 ±0.0111	0.0703 ±0.0045	5.7340e-11
$\varepsilon=1/6$	0.0761 ±0.0057	0.0641 ±0.0035	5.6615e-18
$\varepsilon=1/8$	0.0686 ±0.0019	0.0618 ±0.0018	2.5953e-26
$\varepsilon=1/10$	0.0724 ±0.0044	0.0646 ±0.0033	3.8104e-12
$\varepsilon=1/12$	0.0706 ±0.0038	0.0631 ±0.0029	3.5960e-17

由表 2 的实验结果可知,随着参数 ε 的变化,错分率和 p-value 也会随之发生变化,在 $\varepsilon=1/8$ 时,基于两种不同抽样方法的错分率表现出较好的效果,且此时 p-value 最小,统计检验也具有显著性差异,所以算法运行过程中 ε 为 1/8.

4.2　SVM 非平衡数据分类算法数值实验

表 3 为基于随机抽样的 SVM 算法和基于马氏抽样的 SVM 算法在 10 个非平衡数据集上运行 50 次的平均错分率、方差和 t-test 对平均错分率差别的显著性检验的结果 p 值,其中表的第一列为“数据集名-数字”,如“Shuttle-1000”表示用马氏抽样或随机抽样从 Shuttle 的训练集中抽取 1000 个训练样本,即算法中 N = 1000,这些训练样本的均衡比例与原数据集的均衡比例相同,其他数据集类似.

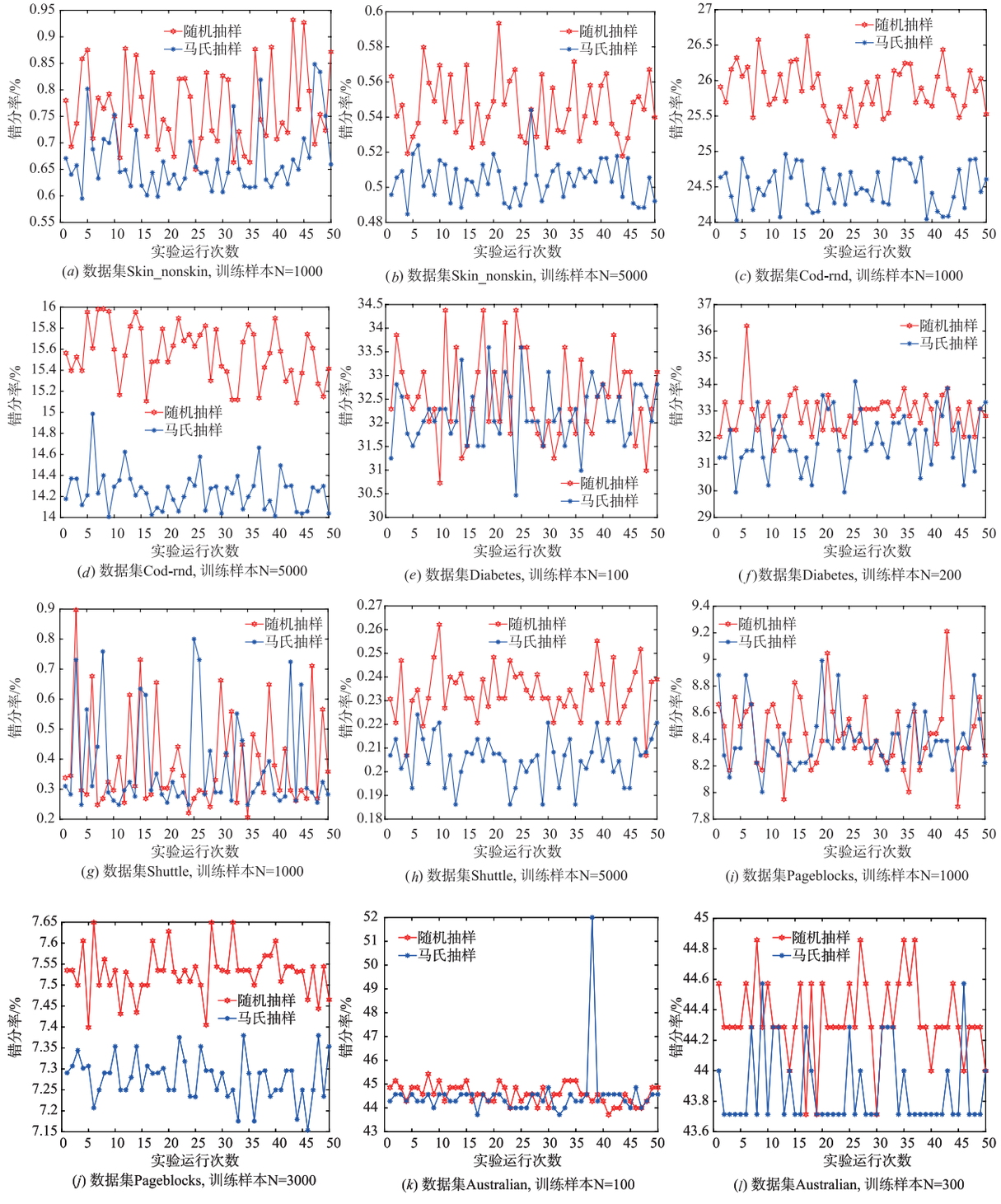
表 3　基于随机抽样的 SVM 和基于马氏抽样的 SVM 的实验对比

数据集	错分率		p-value
	随机抽样	马氏抽样	
Skin_nonskin-1000	0.0077 ±0.0007	0.0067 ±0.0006	1.4006e-05
Cod-rnd-1000	0.2589 ±0.0032	0.2451 ±0.0029	2.4468e-28
Diabetes-100	0.3256 ±0.0091	0.3219 ±0.0064	9.6740e-08
Shuttle-1000	0.0039 ±0.0016	0.0038 ±0.0014	9.2303e-08
Pageblocks-1000	0.0845 ±0.0026	0.0840 ±0.0021	1.4875e-02
Australian-100	0.4458 ±0.0040	0.4451 ±0.0111	2.9240e-02
German-200	0.3117 ±0.0193	0.2962 ±0.0129	7.2051e-07
Census_income-1000	0.0629 ±0.0028	0.0599 ±0.0013	5.9528e-10
Covtype-1000	0.2962 ±0.0080	0.2797 ±0.0053	1.6803e-16
Bin_connect4-1000	0.2971 ±0.0039	0.2756 ±0.0032	3.5243e-32

由表 3 的实验结果可以看出,基于马氏抽样的 SVM 算法的平均错分率比基于随机抽样的 SVM 算法的平均错分率要低,且基于马氏抽样的 SVM 算法比基于随机抽样的 SVM 算法的方差要小,数据集 Australian 在训练样本 N = 100 下的方差除外. 同时由表 3 的统计

显著性值 p-value 可以看出,所有非平衡数据集的 p-value 都小于 0.05,表明从统计检验来看基于两种不同抽样方法的 SVM 非平衡数据分类算法的平均错分率具有显著差异.

为了更好地显示实验结果,图 1 给出了基于随机抽样的 SVM 算法和基于马氏抽样的 SVM 算法对 10 个不平衡数据集重复实验 50 次的错分率对比情况. 由图 1 的 SVM 非平衡数据分类算法的错分率分布情况可以看出,对每个数据集都抽取不同的训练样本进行实验,随着训练样本的增加,SVM 非平衡数据分类算法的泛化性能越好.

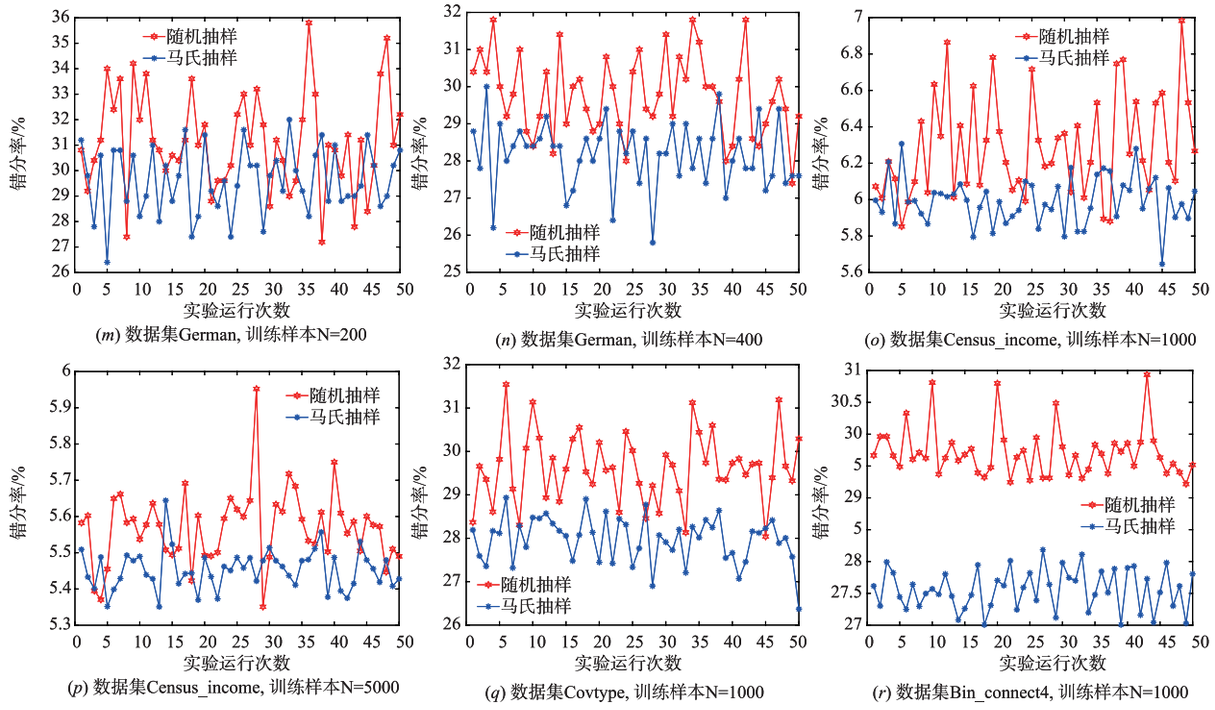


(a) 数据集Skin_nonskin, 训练样本N=1000 (b) 数据集Skin_nonskin, 训练样本N=5000 (c) 数据集Cod-rnd, 训练样本N=1000

(d) 数据集Cod-rnd, 训练样本N=5000 (e) 数据集Diabetes, 训练样本N=100 (f) 数据集Diabetes, 训练样本N=200

(g) 数据集Shuttle, 训练样本N=1000 (h) 数据集Shuttle, 训练样本N=5000 (i) 数据集Pageblocks, 训练样本N=1000

(j) 数据集Pageblocks, 训练样本N=3000 (k) 数据集Australian, 训练样本N=100 (l) 数据集Australian, 训练样本N=300



(*m*) 数据集German, 训练样本N=200　(*n*) 数据集German, 训练样本N=400　(*o*) 数据集Census_income, 训练样本N=1000

(*p*) 数据集Census_income, 训练样本N=5000　(*q*) 数据集Covtype, 训练样本N=1000　(*r*) 数据集Bin_connect4, 训练样本N=1000

图1　基于随机抽样的SVM算法和基于马氏抽样的SVM算法对10个不平衡数据集重复实验50次的错分率对比情况

4.3　EDSVM 非平衡数据分类算法数值实验

表 4 给出了基于随机抽样的 EDSVM 算法和基于马氏抽样的 EDSVM 算法分别在 10 个非平衡数据集上运行 50 次的平均错分率、方差和 t-test 对平均错分率差别的显著性进行检验结果 p 值,表的第一列为“数据集名-数字”,如“Diabetes-100”表示用马氏抽样或随机抽样从 Diabetes 训练集抽取 100 个训练样本,即 EDSVM 中的 $N=100$,抽取训练样本的均衡比例与原数据集的均衡比例相同,其他数据集类似.

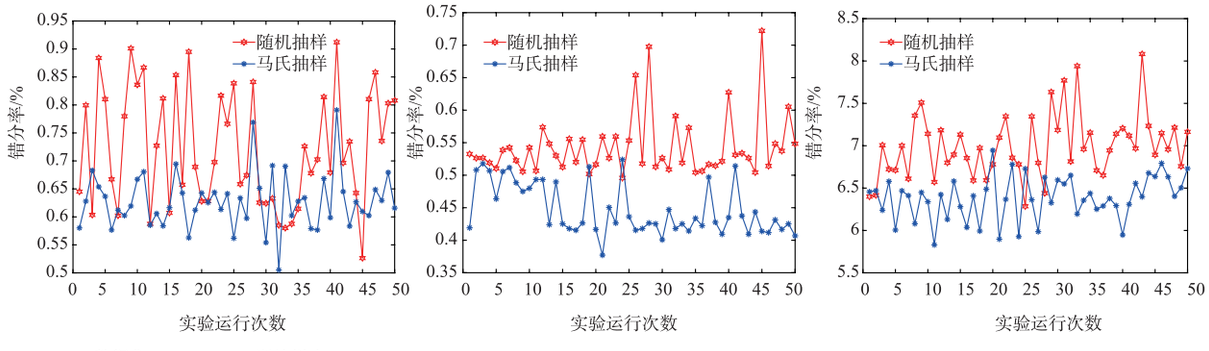
由表 4 显示的实验对比结果可以看出,基于马氏抽样的 EDSVM 算法比基于随机抽样的 EDSVM 算法分类准确率更高,并且基于马氏抽样的 EDSVM 算法比基于随机抽样的 EDSVM 算法的方差要小. 同时由表 4 中实验结果的最后一列统计显著性值 p-value 可以看出,所有非平衡数据集的统计显著性值 p-value 都小于 0.01,表明从统计检验来看基于两种不同抽样方法(随机抽样或者马氏抽样)的 EDSVM 非平衡数据分类算法的平均错分率差异较为显著.

为了更好地显示实验结果,图 2 给出了基于随机抽样的 EDSVM 算法和基于马氏抽样的 EDSVM 算法对 10 个不平衡数据集重复实验 50 次的错分率对比情况.

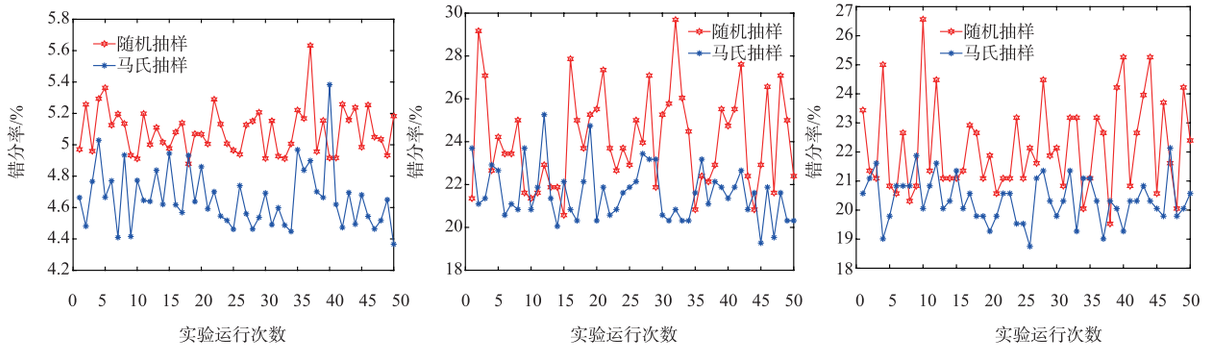
由图 2 的 EDSVM 非平衡数据分类算法错分率的分布情况可以看出,随着训练样本的增加,EDSVM 非平衡数据分类算法的学习性能越好.

表 4　基于随机抽样的 EDSVM 和基于马氏抽样的 EDSVM 实验对比

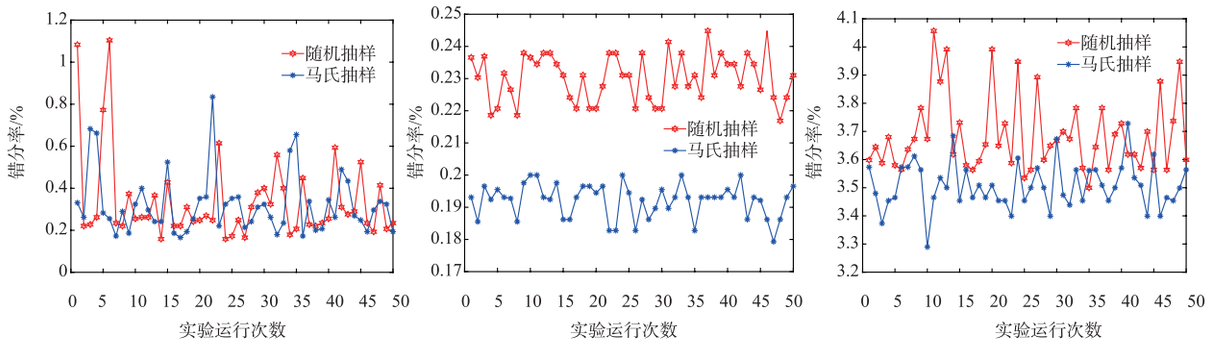
数据集	错分率		p-value
	随机抽样	马氏抽样	
Skin_nonskin-1000	0.0072 ±0.0010	0.0065 ±0.0007	1.8799e −06
Cod-rnd-1000	0.0700 ±0.0038	0.0638 ±0.0026	2.5799e −12
Diabetes-100	0.2411 ±0.0227	0.2159 ±0.0128	4.6873e −08
Shuttle-1000	0.0034 ±0.0020	0.0032 ±0.0015	6.7280e −03
Pageblocks-1000	0.0369 ±0.0032	0.0351 ±0.0029	3.3834e −04
Australian-100	0.1669 ±0.0161	0.1538 ±0.0110	5.1637e −05
German-200	0.3052 ±0.0191	0.2809 ±0.0142	7.4247e −06
Census_income-1000	0.0610 ±0.0016	0.0593 ±0.0010	1.1014e −04
Covtype-1000	0.2870 ±0.0076	0.2740 ±0.0046	6.5122e −14
Bin_connect4-1000	0.2920 ±0.0043	0.2723 ±0.0040	3.4282e −30



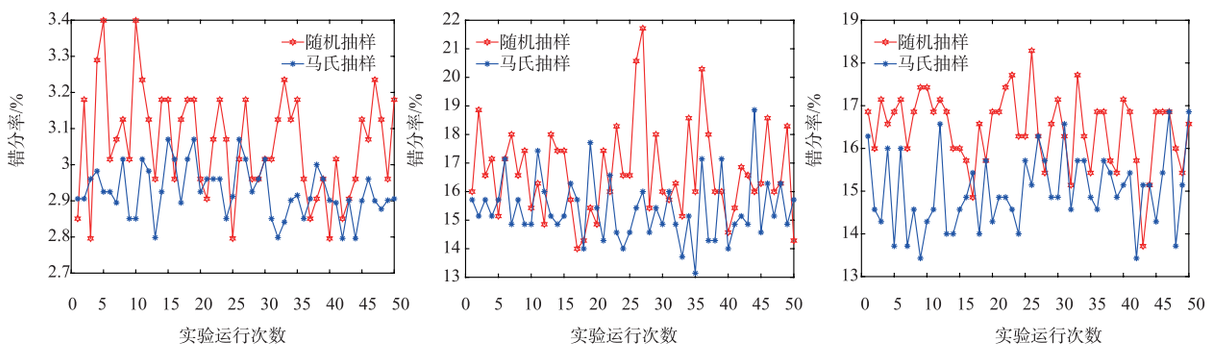
(*a*) 数据集Skin_nonskin, 训练样本N=1000 (*b*) 数据集Skin_nonskin, 训练样本N=5000 (*c*) 数据集Cod-rnd, 训练样本N=1000



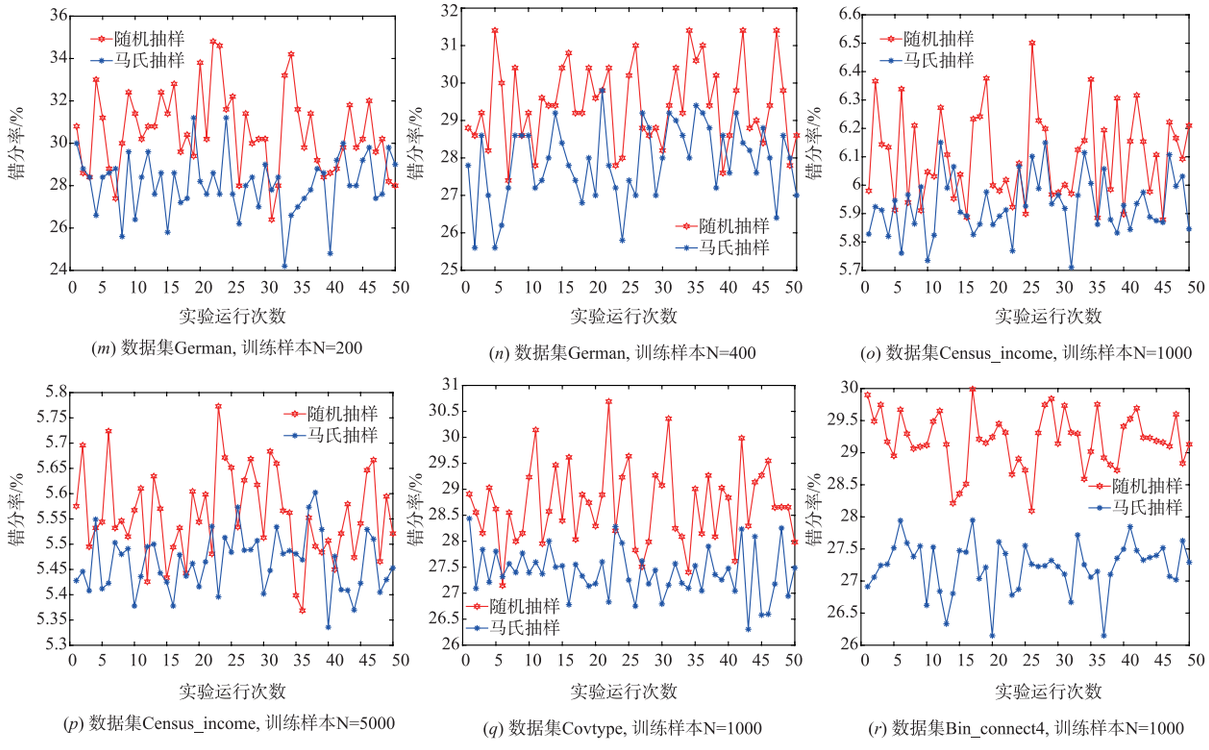
(*d*) 数据集Cod-rnd, 训练样本N=5000 (*e*) 数据集Diabetes, 训练样本N=100 (*f*) 数据集Diabetes, 训练样本N=200



(*g*) 数据集Shuttle, 训练样本N=1000 (*h*) 数据集Shuttle, 训练样本N=5000 (*i*) 数据集Pageblocks, 训练样本N=1000



(*j*) 数据集Pageblocks, 训练样本N=3000 (*k*) 数据集Australian, 训练样本N=100 (*l*) 数据集Australian, 训练样本N=300



(*m*) 数据集German, 训练样本N=200　(*n*) 数据集German, 训练样本N=400　(*o*) 数据集Census_income, 训练样本N=1000

(*p*) 数据集Census_income, 训练样本N=5000　(*q*) 数据集Covtype, 训练样本N=1000　(*r*) 数据集Bin_connect4, 训练样本N=1000

图2　基于随机抽样的EDSVM算法和基于马氏抽样的EDSVM算法对10个不平衡数据集重复实验50次的错分率对比情况

4.4　SVM-WKNN 非平衡数据分类算法数值实验

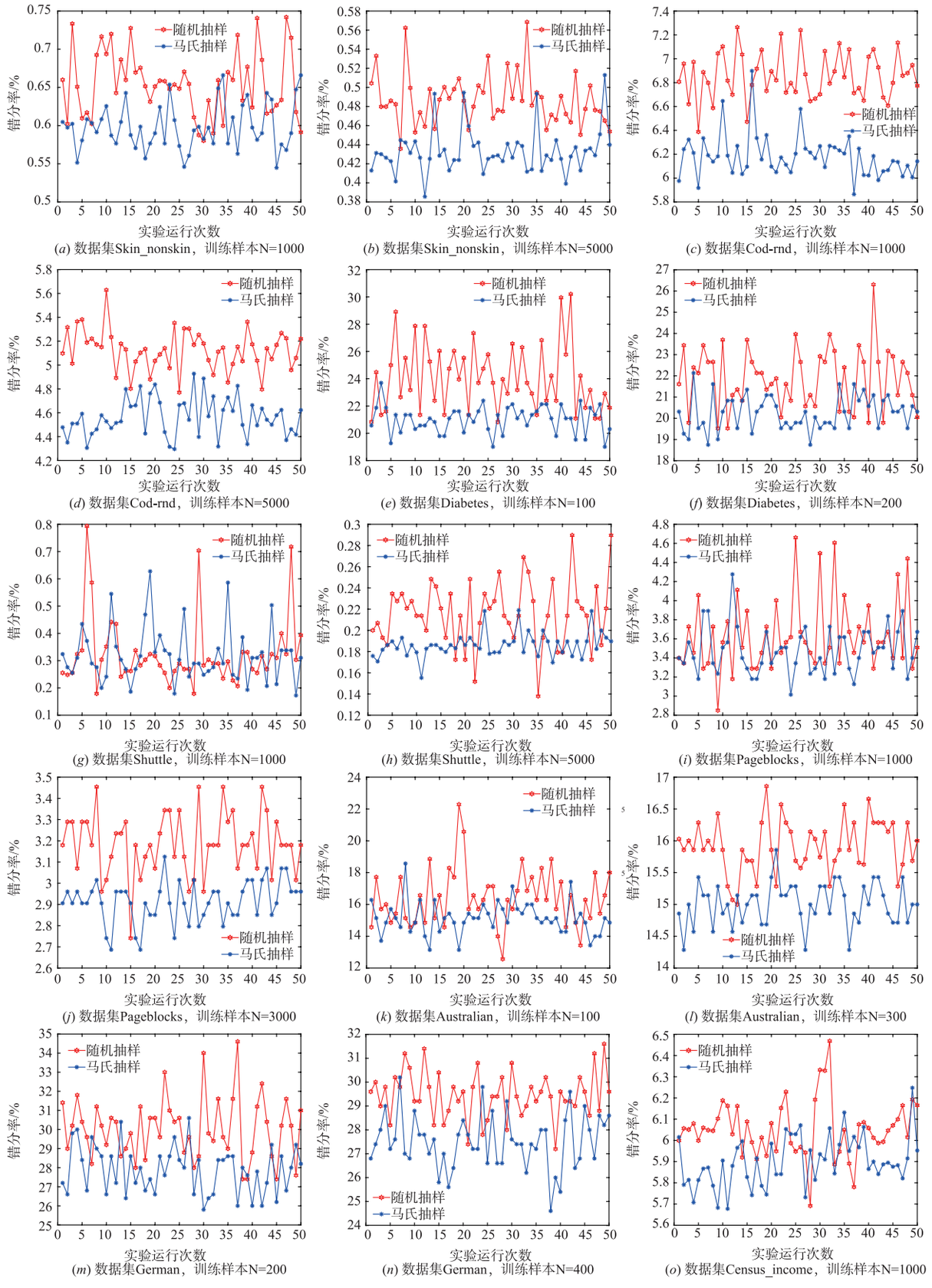
表5 为基于随机抽样的 SVM-WKNN 算法和基于马氏抽样的 SVM-WKNN 算法在 10 个非平衡数据集上运行 50 次的平均错分率、方差和 t-test 对平均错分率差别的显著性检验的结果 p 值,表的第一列为“数据集名-数字”,如“German-200”表示使用马氏抽样或随机抽样从非平衡数据集 German 的训练集中抽取了 200 个训练样本,即 SVM-WKNN 算法中的 N = 200,抽取的训练样本的均衡比例与原数据集的均衡比例相同,其他数据集类似.

由表 5 中的实验结果可以看出,基于马氏抽样的 SVM-WKNN 算法比基于随机抽样的 SVM-WKNN 算法分类准确率高,且基于马氏抽样算法的方差也较小,说明基于马氏抽样的算法更加稳定. 同时由表 5 中的统计显著性值 p-value 可知,所有非平衡数据集的 p-value 均小于 0.01,表明从统计检验来看基于两种不同抽样方法的 SVM-WKNN 非平衡数据分类算法的平均错分率差异较为显著.

为了更好地显示实验结果,图 3 给出了基于随机抽样的 SVM-WKNN 算法和基于马氏抽样的 SVM-WKNN 算法对 10 个不平衡数据集重复实验 50 次的错分率对比情况. 由图 3 的 SVM-WKNN 非平衡数据分类算法错分率的分布情况可以看出,同一数据集抽取的训练样本越多,最终 SVM-WKNN 非平衡数据分类算法的泛化性能越好.

表 5　基于随机抽样的 SVM-WKNN 算法和基于马氏抽样的 SVM-WKNN 算法的实验对比

数据集	错分率		p-value
	随机抽样	马氏抽样	
Skin_nonskin-1000	0.0065 ±0.0004	0.0060 ±0.0003	1.1851e −08
Cod-rnd-1000	0.0686 ±0.0019	0.0618 ±0.0018	2.5953e −26
Diabetes-100	0.2407 ±0.0244	0.2103 ±0.0010	5.9752e −09
Shuttle-1000	0.0032 ±0.0013	0.0031 ±0.0010	2.0706e −03
Pageblocks-1000	0.0363 ±0.0038	0.0348 ±0.0025	4.9230e −06
Australian-100	0.1642 ±0.0177	0.1514 ±0.0102	4.7966e −03
German-200	0.3032 ±0.0144	0.2806 ±0.0133	1.0173e −07
Census_income-1000	0.0605 ±0.0013	0.0590 ±0.0012	1.9220e −05
Covtype-1000	0.2840 ±0.0065	0.2717 ±0.0038	1.0268e −15
Bin_connect4-1000	0.2842 ±0.0029	0.2654 ±0.0028	2.6735e −37



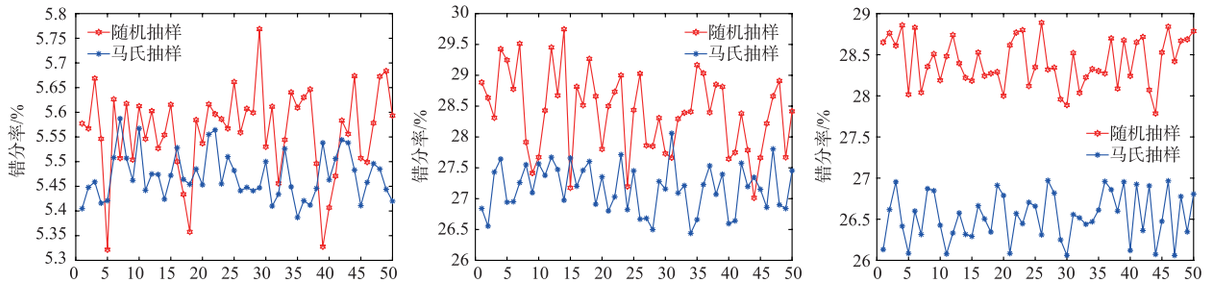
(*a*) 数据集Skin_nonskin，训练样本N=1000 (*b*) 数据集Skin_nonskin，训练样本N=5000 (*c*) 数据集Cod-rnd，训练样本N=1000

(*d*) 数据集Cod-rnd，训练样本N=5000 (*e*) 数据集Diabetes，训练样本N=100 (*f*) 数据集Diabetes，训练样本N=200

(*g*) 数据集Shuttle，训练样本N=1000 (*h*) 数据集Shuttle，训练样本N=5000 (*i*) 数据集Pageblocks，训练样本N=1000

(*j*) 数据集Pageblocks，训练样本N=3000 (*k*) 数据集Australian，训练样本N=100 (*l*) 数据集Australian，训练样本N=300

(*m*) 数据集German，训练样本N=200 (*n*) 数据集German，训练样本N=400 (*o*) 数据集Census_income，训练样本N=1000



(*p*) 数据集Census_income，训练样本N=5000　(*q*) 数据集Covtype，训练样本N=1000　(*r*) 数据集Bin_connect4，训练样本N=1000

图3　基于随机抽样的SVM-WKNN算法和基于马氏抽样的SVM-WKNN算法对10个不平衡数据集重复实验50次的错分率对比情况

评注 3:由图 1、图 2、图 3 以及表 3、表 4、表 5 的实验结果可以看出,三中算法不管是哪种算法,基于马氏抽样算法的错分率均比基于随机抽样相应算法的错分率低,且由表 3、表 4、表 5 的实验结果表明,对于同一种抽样方法(随机抽样或马氏抽样),SVM 算法的错分率最大,EDSVM 算法次之,SVM-WKNN 算法的错分率最小. 同时由表 3、表 4、表 5 中的统计显著性值 p-value 可知,从统计检验来看基于不同抽样方法的三种非平衡数据分类算法的错分率具有显著差异.

5　结论

在机器学习过程中,经常存在不平衡数据的分类问题,标准的分类学习算法对非平衡数据进行分类时,通常是基于样本是独立同分布的假设,但这并不符合实际应用中样本的分布情况. 本文通过马氏抽样将样本是独立同分布的情形减弱为了一致遍历马氏链,提出了基于马氏抽样的 SVM 非平衡数据分类算法和基于马氏抽样的 EDSVM 非平衡数据分类算法,并对 EDSVM 算法的不足进行了改进,形成了学习性能更好的基于马氏抽样的 SVM-WKNN 非平衡数据分类算法. 实验结果表明,对马氏抽样后的样本进行训练得到的分类器泛化性能更好. 如何将算法更好的应用在超高维数据与多分类情况中是下一步研究的内容.

参考文献

[1] Chawla N, Japkowicz N. Special issue on learning from imbalanced data sets[J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 1 – 6.

[2] Maciej A M, Piotr A H. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance [J]. Neural Networks, 2008, 21(2): 427 – 436.

[3] 翟云,王树鹏,马楠,杨炳儒,张德政. 基于单边选择链和样本分布密度融合机制的非平衡数据挖掘方法[J]. 电子学报,2014,42(7):1311 – 1319.
ZHAI Yun, WANG Shupeng, MA Nan, et al. A data mining method for imbalanced datasets based on one-sided link and distribution density of instances[J]. Acta Electronica Sinica, 2014, 42(7): 1311 – 1319. (in Chinese)

[4] Moreo A, Esuli A, Sebastiani F. Distributional random oversampling for imbalanced text classification[A]. Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. Pisa: ACM, 2016. 805 – 808.

[5] 欧阳震诤,罗建书,胡东敏,吴泉源. 一种不平衡数据流集成分类模型[J]. 电子学报,2010,38(1):184 – 189.
OUYANG Zhenzheng, LUO Jianshu, HU Dongmin, WU Quanyuan. An ensemble classifier framework for mining imbalanced data streams [J]. Acta Electronica Sinica, 2010, 38(1): 184 – 189. (in Chinese)

[6] Li Y L, Zhu Y H, Yang P. Text classification for imbalanced data sets[J]. Information Science and Engineering, 2008, 2(20 22): 778 – 781.

[7] Duan L X, Xie M Y, Bai T B, Wang J J. A new support vector data description method for machinery fault diagnosis with unbalanced datasets[J]. Expert Systems with Applications, 2016, 64: 239 – 246.

[8] Sui H F, Yang B R, Zhai Y. The problem of classification in imbalanced data sets in knowledge discovery[J]. Computer Application and System Modeling, 2010, 9(22 24): 658 – 661.

[9] 郝秀兰,陶晓鹏,徐和祥,胡运发. KNN 文本分类器类偏斜问题的一种处理对策[J]. 计算机研究与发展,2009,46(1):52 – 61.
HAO Xiulan, TAO Xiaopeng, XU Hexiang, HU Yunfa. A strategy to class imbalance problem for KNN text classifier [J]. Computer Research and Development, 2009, 46(1): 52 – 61. (in Chinese)

[10] He H B, Edwardo A G. Learning from imbalanced data [J]. Knowledge and Data Engineering, 2009, 21(9): 1263 – 1282.

[11] Japkowicz N. The class imbalance problem: Significance and strategies [A]. Proceedings of 2000 International

Conference on Artificial Intelligence (ICAI2000) [C]. Las Vegas, Nevada: Computer Science & Information Technology, 2000. 111 – 117.

[12] 李蓉, 叶世伟, 史忠植. SVM-KNN 分类器: 一种提高 SVM 分类精度的新方法[J]. 电子学报, 2002, 30(5): 745 – 748.
LI Rong, YE Shiwei, SHI Zhizhong. SVM-KNN classifier: A new method of improving the accuracy of SVM classifier[J]. Acta Electronica Sinica, 2002, 30(5): 745 – 748. (in Chinese)

[13] 王超学, 张超, 马春森. 改进 SVM-KNN 的不平衡数据分类[J]. 计算机工程与应用, 2016, 52(4): 51 – 55.
WANG Chaoxue, ZHANG Chao, MA Chunsen. Improved SVM-KNN algorithm for imbalanced datasets classification[J]. Computer Engineering and Applications, 2016, 52(4): 51 – 55. (in Chinese)

[14] Xu J, Tang Y Y, Zou B, Xu Z B, Li L Q, Lu Y. The generalization ability of SVM classification based on Markov sampling[J]. IEEE Trans on Cybernetics, 2015, 45(6): 1169 – 1179.

[15] Zou B, Tang Y Y, Xu Z B, Li L Q, Xu J, Lu Y. The generalization performance of regularized regression algorithms based on Markov sampling[J]. IEEE Trans on Cybernetics, 2014, 44(9): 1497 – 1507.

[16] Smale S, Zhou D X. Online learning with Markov sampling[J]. Analysis and Applications, 2009, 7(1): 87 – 113.

[17] Vapnik V. Statistic Learning Theory[M]. New York: John Wiley, 1998.

[18] Vidyasagar M. Learning and Generalization with Applications to Neural Networks [M]. London: Springer-Verlag, 2003.

[19] UCI 机器学习存储库[OL]. http://archive.ics.uci.edu/ml/index.php.

作者简介

徐 婕 女, 1975 年生, 湖北武汉人, 教授, 主要研究方向为机器学习、计算机网络.
E-mail: frangipani@hubu.edu.cn

贺美美 女, 1993 年生, 陕西渭南人, 硕士研究生, 主要研究方向为机器学习.
E-mail: hemei.work@foxmail.com