基于支持向量机的不平衡数据分类的改进欠采样方法*

赵自翔，王广亮，李晓东
（中山大学信息科学与技术学院//智能传感器网络教育部重点实验室，广东 广州 510006）

摘 要：支持向量机作为一种有监督分类算法，具有小样本，非线性等独特优势，但其在处理不平衡数据分类时效果不够理想。欠采样是一类常用的数据重构方法，它被广泛用于解决不平衡数据的分类问题，然而，传统的随机欠采样方法受随机性影响，稳定性较差。提出一种改进的欠采样方法，并应用在支持向量机上进行分类对比实验。实验结果表明，相比传统随机欠采样方法，该方法的稳定性更好，且在许多情况下可以提高支持向量机对不平衡数据的分类性能。

关键词：支持向量机；不平衡数据；欠采样；稳定性
中图分类号：TP391.4 **文献标志码**：A **文章编号**：0529－6579（2012）06－0010－07

An Improved SVM Based Under-Sampling Method for Classifying Imbalanced Data

ZHAO Zixiang, *WANG Guangliang*, *LI Xiaodong*
（School of Information Science and Technology//Key Lab of Machine Intelligence and Sensor Network, Ministry of Education, Sun Yat-sen University, Guangzhou 510006, China）

Abstract: As a supervised classifier, Support Vector Machine (SVM) has prominent advantages in solving some problems on petty and nonlinear datasets, but it is unsatisfying in tackling with imbalanced datasets. Random under-sampling has been a widely used method to improve SVM's performance on imbalanced data, but its stability is easily influenced by the nature of randomness. A modified SVM based on under-sampling method is presented to classify imbalanced data. Compared with the random under-sampling technique, it is shown through experiments on natural datasets that the new proposed under-sampling method is more stable in classifying imbalanced data, and exhibits improved SVM performance in classifying imbalanced data for many cases.

Key words: support vector machine; imbalanced data; under-sampling; stability

在数据分类问题中，人们往往会遇到这种情况：在一个有两个类别数据的数据集中，一类数据较多，而另一类数据较少，我们把这种数据集称为不平衡数据。传统机器学习方法用于不平衡数据分类时，往往会产生很大的偏向性，即对较多的一类数据（以下简称多数类）有很高的识别率，对较少的一类数据（以下简称少数类）识别率却很低。遗憾的是，在日常生活中，对人们有用的往往是比较少的那类数据[1-2]。以检测信用卡非法交易[3]为例，在信用卡的成千上万次使用记录中，绝大部分是正常而合法的，只有极少数属于非法交易。如果对合法记录识别较好而对极少数的非法交易识别率低下，检测过程将变得毫无意义。简单来说，当把少数类数据错分为多数类的代价远远高于把多数类

* **收稿日期**：2012－05－19
基金项目：国家自然科学基金资助项目（U1135005）
作者简介：赵自翔（1987 年生），男，硕士生；**通讯作者**：李晓东；E-mail：lixd@mail.sysu.edu.cn

错分为少数类的代价时，一些“偏爱”多数类的传统分类方法就不再适用。

多年以来，人们对不平衡数据分类中“偏爱”多数类问题的产生原因进行过一系列研究。Japkowicz 等[4]认为造成不平衡数据分类问题的因素有类间不平衡程度（即类间训练样本数量的比值）、训练样本规模和概念复杂度等，其中以类间不平衡程度最为明显。Prati 等[5]则通过实验证明：即使在一些数量严重不平衡的数据中，只要类重叠不严重，分类器仍然具有良好效果，从而把类重叠也作为引起不平衡数据分类问题的重要因素。在此基础上，一系列针对不平衡数据分类问题的解决方案被人们提出[2]，其中，欠采样就是一种解决不平衡数据分类问题的有效方法。然而，常用的随机欠采样方法因其自身的特点导致分类稳定性较差。针对于此，本文将在讨论欠采样原理及不平衡数据分类问题评价标准的基础上，提出一种改进的欠采样方法（FN 欠采样），并结合支持向量机，通过实验仿真进行对比研究。

1　支持向量机原理及不平衡数据分类优劣的评价标准

1.1　支持向量机原理

支持向量机（Support Vector Machine）由 Cortes 和 Vapnik 于 1995 年提出[6]，它建立在统计学习理论的 VC 维理论和结构风险最小原理基础上，在小样本，非线性，高维识别等方面表现出独特的优势，并在模式识别领域得到了广泛应用。

根据文［6］中所述，设样本集为 (x_i,y_i) $(i=1,\cdots,l)$，$x_i \in \mathbf{R}^d$，$y_i \in \{+1,-1\}$，同时引入松弛变量 ξ_i 和惩罚因子 C 以允许部分错分样本的存在。当分类面为 $<w,x>+b=0$ 时，为了使两类样本间的几何间隔 $2/w$ 最大，可将支持向量机的原型写为：

$$\text{Min}\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i$$

$$\text{s.t.}\ y_i(w\cdot x_i + b) \geqslant 1-\xi_i,$$

$$\xi_i \geqslant 0, i=1,\cdots,l$$

与其他分类算法相比，支持向量机具有解决非线性问题，高维度的独特优势，并且 Japkowicz 等[4]曾通过实验证明：在 BP 神经网络、C4.5 决策树算法[7]以及支持向量机中，不平衡数据对支持向量机带来的影响是相对较小的，因此基于支持向量机解决不平衡数据分类问题，是一个切实可行的方向[8]。

1.2　不平衡数据分类优劣的评价标准

传统模式分类问题中，使用总准确率作为衡量分类器性能的主要指标，但在不平衡数据分类问题中，这不再适用。举例而言，一组少数类和多数类数目分别为 5 和 95 的测试数据，即使少数类数据完全被分错，也很容易达到 90% 以上的总准确率，这显然不够合理。

在机器学习的二分类问题中，对于一个测试数据，如果它被分类器分为正类，而它实际上也是正类，我们称之为正确正类，写为 TP；如果它被分为正类，而实际上却是负类，称之为错误正类，写为 FP。与此类似，可以定义正确负类 TN 和错误负类 FN，如表 1 所示：

表 1　二分类问题中一次预测可能出现的四种结果

Table 1　Four different possible outcomes of a single prediction for a two-class case

	实际为正类	实际为负类
预测为正类	TP	FP
预测为负类	FN	TN

根据这些术语，可以定义如下指标：

查准率（击中率）Precision = TP/（TP + FP）

查全率 Recall = TP/（TP + FN）

虚警率 FA = FP/（FP + TN）一种被广泛使用的评价不平衡问题的方法是 ROC 曲线[9]。ROC 曲线以虚警率为 X 轴，以击中率为 Y 轴，通过调整分类器的决策阈值得到一条曲线。ROC 曲线直观明了，曲线整体越凸，越靠近左上方，表明分类器性能越好。

另一种被广泛使用的评价标准是几何平均数[10-11]，它使用查准率和查全率的几何平均数作为衡量不平衡数据分类优劣的标准，即 Gmean = $\sqrt{\text{Precision}\times\text{Recall}}$。Gmean 同时兼顾了分类器对多数类和少数类数据的分类性能，简单来说，Gmean 越大，表示分类器在解决不平衡问题时性能越好。本文采用 Gmean 作为衡量不平衡数据分类优劣的评价标准。

2　数据重构方法简介

数据重构是解决数据不平衡问题的一类常用方法，通常分为过采样和欠采样两类。

过采样的基本思想是针对二分类问题中的少数类，设法增加其样本数。最基本的过采样是随机过

采样，即随机选取原少数类中的样本，复制并加入到训练集中。

欠采样的基本思想跟过采样相反，它针对二分类中的多数类，人为减少其样本数，从而使训练集数量平衡。随机欠采样是最简单的欠采样方法，通过随机舍弃多数类中的一些样本来达到数量平衡。

过采样使样本规模变大，增加了训练时间，且容易导致过拟合。欠采样虽然在训练时间上有明显缩短，但是在去掉多数类样本的过程中，可能会去除某些对分类过程“有价值”的点。尽管如此，Drummond 等[12]认为，综合而言，欠采样在性能上优于过采样。

3 改进的欠采样方法

在模式分类的有监督学习中，训练分类器的典型做法是，给定一个输入，计算其输出类别，把它与实际类别标记做比较，并根据差异来改善分类器性能。因此训练集样本是影响分类器性能的重要因素之一。一个良好的训练集应该具备以下特点：（i）不同类别的样本分布均衡；（ii）各样本在特征空间中分布比较集中；（iii）每个类别中的样本，都能够很好地代表该类别的特点。针对特点（i），可以采用分层抽样的方法，使不同类别在训练集中保持原来的比例。特点（ii）涉及的是单个样本的特征维数，我们希望在容易提取、对噪声不敏感，并对区分不同类别很有效的基础上，选取尽量少的特征来构成样本[13-14]。通常情况下，我们需要关注的是怎样使训练集中的每一类样本都能够尽量好地代表这一类数据的特点。

如前所述，欠采样与过采样相比，有着数据量少，训练时间短的优点。但欠采样的缺点也同样明显，容易去掉训练集的多数类样本中有价值的样本点。由于受偶然性的影响，在一次对训练集的多数类随机欠采样的过程中，可能得到很好的效果，也可能得到非常不理想的“失真”训练集。以图 1 的二维数据为例，圆点代表多数类，方点代表少数类。图 1（A）为对原始不平衡训练集的分类情况，分类结果表现为“偏爱”多数类。图 1（B）对多数类数据进行了一次随机欠采样，黑色点为参与训练的样本点。虽然得到的两类数据数量相同，但过程中许多有价值的样本被舍弃，留下来的样本点体现的仅仅是原数据集极其有限的一部分特点。使用该训练集训练的分类器，对训练集的分类效果或许很好，对于其余数据的分类效果却会很差，如图 1（B）所示。一个优秀的训练集，应该是能最大程度地表现原数据集特点的样本集合。在使用随机欠采样方法选取训练集的过程中，这种“失真”情况的发生将会导致分类器稳定性降低。

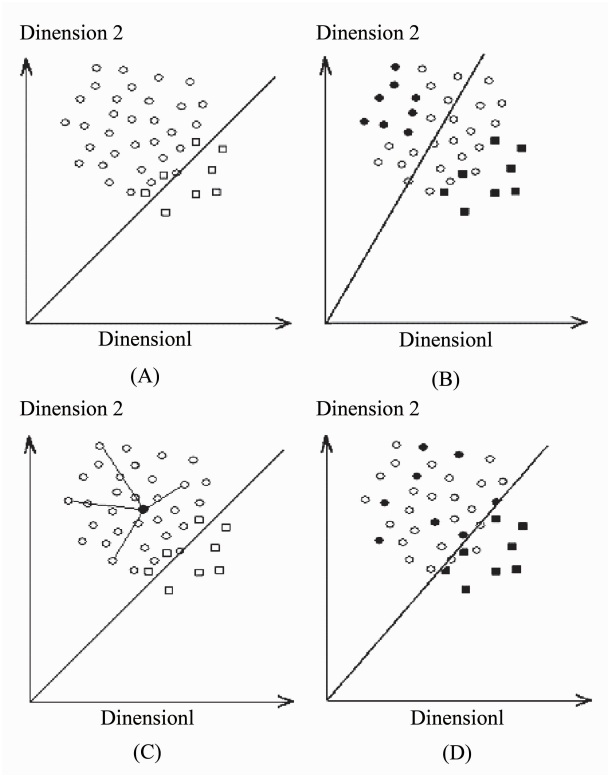

图 1 一次随机欠采样和一次 FN 欠采样结果比较

Fig. 1 Comparison of results between one random under-sampling and one FN under-sampling

在统计学习中，正态分布是最普遍的分布情况。当样本的各个特征大多满足这种分布特性时，样本将在特征空间中形成一个中部集中、边缘稀疏的超几何体形状。基于该理论，Tax 等[15]提出了 SVDD 方法，利用紧致超球体对样本分布进行描述。以这种描述为基础，我们进一步提出以下设想：在多数类的特征空间中，假定各个样本的分布是一个接近球体的凸几何体。以图 1（A）的二维空间为例，可以把多数类的分布看作一个接近椭圆的集群。在欠采样过程中，当保留的仅仅是集群中有限的某一个区域中的样本点时，将有大量的有价值点被舍弃；而如果能在集群的每一区域均保留一定量的样本，则能够防止“失真”的最坏情况发生。对于集群上某一区域样本点来说，它们到一个定点的距离应该是相差不大的。于是，为了在一次欠采样过程中尽可能保持训练集中多数类样本原本的类别特点，我们采用如下方法：（规定实验中的少数类为正类，多数类为负类）先找到所有负类

样本点的均值点，如图1（C）中的黑色圆点所示；计算所有负类样本到该均值点的距离，在距离相近的每个小区域中，保留一个点而去掉剩下的点，并将保留下的所有负类样本点作为新的负类样本集和原有的正类样本集一起组成新训练集，如图 1（D）所示。由于在算法过程中我们采用从离均值点最远的点开始舍弃样本，我们把这种欠采样方法称为 FN 欠采样方法（Furthest Neighbor based under-sampling）。下面给出算法的基本流程：

FN Under-sampling（TrainData，T）

1）读入 TrainData //读入训练集

设置参数 T　　//决定将训练集中的多数类缩小的倍数

2）找出 TrainData 中的正类子集 - > P_ TrainData

找出 TrainData 中的负类子集 - > N_ TrainData

//实验中我们将少数类标记为正类，多数类为负类，多数类集合为 N_ TrainData

3）得出 N_ TrainData 中样本的均值 - > MeanSample

4）计算 N_ TrainData 中每个样本与 MeanSample 的距离，并将 N_ TrainData 中各样本按所得距离从大到小排序 - > N_ TrainData_ Sorted

5）设置计数器 timer = 0，遍历 N_ TrainData_ Sorted 中的样本，每经过一个样本，timer 加一，当且仅当 timer 为 T 的整数倍时保留当下样本，其余情况将当下样本从 N_ TrainData_ Sorted 中去掉。

//实质上就是每轮去掉（T - 1）个样本之后，保留一个，如此反复形成新的负类子集

6）将上一步完成后的 N_ TrainData_ Sorted 作为新的负类子集，和正类子集 P_ TrainData 一起形成新的训练集，带入支持向量机进行训练。

4　实验结果和分析

我们采用以下 7 组 UCI 数据对相关算法进行对比实验。

（i）Abalone 数据：它是 UCI 数据库中一组关于鲍鱼年龄预测的数据集，每个样本有 8 个特征，共有 29 个类别，实验中选择类别“18”作为正类，其余统一作为负类。

（ii）Ecoli 数据：它是关于蛋白质研究的数据，共有 8 个类别，每个样本 8 维，其中类别标号“im”的样本为正类，其余为负类。

（iii）Glass 数据：它有 6 个类别，类别“3”被选为正类，其余为负类，每个样本 10 个特征。

（iv）Haberman 数据：它是关于乳腺癌病人生存状况的一组统计数据，分两类，其中取类别“2”为正类，类别“1”为负类.

（v）Page-blocks 数据：它是典型的文本分类数据，共有 5 个类，类别“5”为正类，其余作为负类，每一个样本 11 个特征。

（vi）Transfusion 数据：它有两类样本，每个样本 5 维，是一个关于血站维护的统计研究数据，我们用类别“1”做正类，类别“2”为负类。

（vii）Yeast 数据：它也是一组关于蛋白质定位研究的数据，共有 10 个类，每个样本 9 个特征，类别“ME2”被选为正类，其余为负类。

在以上七组数据中，统一规定少数类为正类，并选择一定数量比例的正负样本参与实验，如表 2 所示。实验采用分层抽样得到训练集和测试集，以保持固定的不平衡比例。

表 2　数据集描述

Table 2　Datasets description

数据集	特征维数	正类样本数	负类样本数	正负样本比值
Abalone	8	42	4 116	1: 98
Ecoli	8	77	231	1: 3
Glass	10	17	187	1: 11
Haberman	4	81	162	1: 2
Page - blocks	11	115	5 290	1: 46
Transfusion	5	178	534	1: 3
Yeast	9	51	1 428	1: 28

我们先对以上 7 组不平衡数据，分别用普通支持向量机（SVM），经过随机欠采样的支持向量机（Under-SVM）以及采用 FN 欠采样方法的支持向量机（FN U-SVM）做一次对比实验，以证明欠采样技术确实对解决不平衡问题有显著效果。实验过程中，支持向量机的核函数采用线性核函数，惩罚因子 C 采用以下步骤寻优：1）在 $\{2^{-5},2^{-4},\cdots,2^{4},2^{5}\}$ 中找到使 Gmean 最大的 C 取值 $C=2^{n}$；2）在第一步的基础上，在 $\{2^{n-0.9},2^{n-0.8},\cdots,2^{n+0.8},2^{n+0.9}\}$ 中找到使 Gmean 最大的 C 取值 $C=2^{\text{best}}$ 作为实验中使用的惩罚因子。实验采取“10 - Fold”交叉验证进行，每次采用原数据集的约 90% 作为训练集，10% 为测试集，重复 10 次后取平均结果。得到结果如表 3 所示，每一组数据按算法性能优劣由上往下排列。

表 3 算法在测试集上的性能比较

Table 3 Performance on test sets 数据集采用方法查准率

数据集	采用方法	查准率/%	查全率/%	Gmean/%
Abalone	Under-SVM	2. 255 3	75. 833 3	13. 077 7
	FN U-SVM	2. 206 0	71. 666 7	12. 573 7
	SVM	NaN	0	NaN
Ecoli	FN U-SVM	66. 859 4	97. 857 1	80. 886 7
	Under-SVM	66. 274 3	91. 428 6	77. 841 9
	SVM	79. 404 8	75. 000 0	77. 171 0
Glass	Under-SVM	10. 814 9	90. 000 0	31. 198 4
	FN U-SVM	10. 621 5	90. 000 0	30. 918 2
	SVM	NaN	0	NaN
Haberman	FN U-SVM	63. 031 7	63. 055 6	63. 043 6
	Under-SVM	60. 146 5	64. 305 6	62. 191 3
	SVM	83. 214 3	37. 083 3	55. 550 5
Page-blocks	Under-SVM	19. 269 7	85. 738 6	40. 646 7
	FN U-SVM	18. 205 6	89. 375 0	40. 337 6
	SVM	NaN	4. 261 4	NaN
Transfusion	FN U-SVM	39. 303 9	77. 788 2	55. 293 6
	Under-SVM	39. 263 8	77. 388 2	55. 123 1
	SVMNaN	0	NaN	
Yeast	Under-SVM	16. 549 5	81. 333 3	36. 688 2
	FN U-SVM	15. 770 6	85. 333 3	36. 684 6
	SVM	NaN	0	NaN

从表 3 中可以看到，普通支持向量机对测试集的分类效果很差。对于一部分严重不平衡的数据，如 Abalone、Page-blocks 等，更是把少数类当作噪音，将测试样本统统分为多数类，从而使 TP 和 FP 均为 0，无法求出查准率。而两种欠采样方法均得到了较好的 Gmean，使支持向量机性能受不平衡数据影响更小。两种欠采样方法之间，性能表现各有优劣，这是因为随机欠采样方法的随机性，使得结果时好时坏。

为了进一步比较两种欠采样方法的性能，较少偶然性的干扰，我们再进行如下实验：在上一步实验的基础上，对于每一组数据，分别进行五组对比，每组分别使用两种欠采样方法重复上步实验 100 次，200 次，500 次，1 000 次和 2 000 次。对每一组数据的实验结果，比较采用随机欠采样和改进欠采样分类时得到的 Gmean，并统计各自 Gmean 较大的次数，结果如表 4 所示。

表 4 两种欠采样性能比较

Table 4 Performance comparison of two under-sampling methods

数据集	重复次数 T	Under-SVM 胜出次数	FN U-SVM 胜出次数
Abalone	100	42	58
	200	94	106
	500	215	285
	1 000	422	578
	2 000	885	1 115
Ecoli	100	39	60
	200	88	112
	500	210	290
	1 000	442	558
	2 000	931	1 069
Glass	100	33	62
	200	69	126
	500	193	297
	1 000	377	598
	2 000	748	1 198
Haberman	100	41	59
	200	96	104
	500	236	264
	1 000	465	535
	2 000	872	1 128
Page-blocks	100	32	68
	200	81	119
	500	196	304
	1 000	370	630
	2 000	718	1 282
Transfusion	100	38	62
	200	96	104
	500	190	310
	1 000	440	560
	2 000	912	1 088
Yeast	100	51	49
	200	105	95
	500	255	245
	1 000	551	449
	2 000	1 056	944

为了更直观比较两种欠采样方法的性能，我们将上表结果绘制成图像如图 2 所示。

从实验结果可以看到，相比随机欠采样，在除 Yeast 外的六组数据的测试中，FN 欠采样方法取得较好效果的次数都更多。部分实验数据中，两种方法的胜出次数之和并不等于重复实验总数，这是因为随机欠采样方法在随机抽样的过程中，可能会抽

到“失真”数据而造成所有样本被分为多数类，出现无法计算 Gmean 的情况。同时可以发现，随着重复次数的增加，虽然改进后的方法仍然效果更佳，但是优势开始慢慢减弱，以 Transfusion 的实验结果为例，$T=100$ 时 FN 欠采样胜出的次数约是随机欠采样的 1.63 倍，而当 $T=2\ 000$ 时，仅为约 1.18 倍。这说明当实验次数足够多时，随机欠采样方法依然有着很好的综合性能。

FN 欠采样对于 Yeast 的实验效果不好，可以解释为数据维数造成的影响。对于一个数据分类问题，样本的各个特征中一部分特征是对分类结果“强相关”的，一部分是“弱相关”的，也有些是“不相关”的。受到“不相关”特征的影响，即使是“强相关”与“弱相关”特征完全一致的两个样本，它们之间的距离也可能会非常远[16]，这就让以均值样本和样本间距离为基础的 FN 欠采样方法失去了意义。Khoshgoftaar 等[17]也曾通过实验证明，在处理不平衡问题时，不同的特征选择过程会对分类效果产生重要影响。

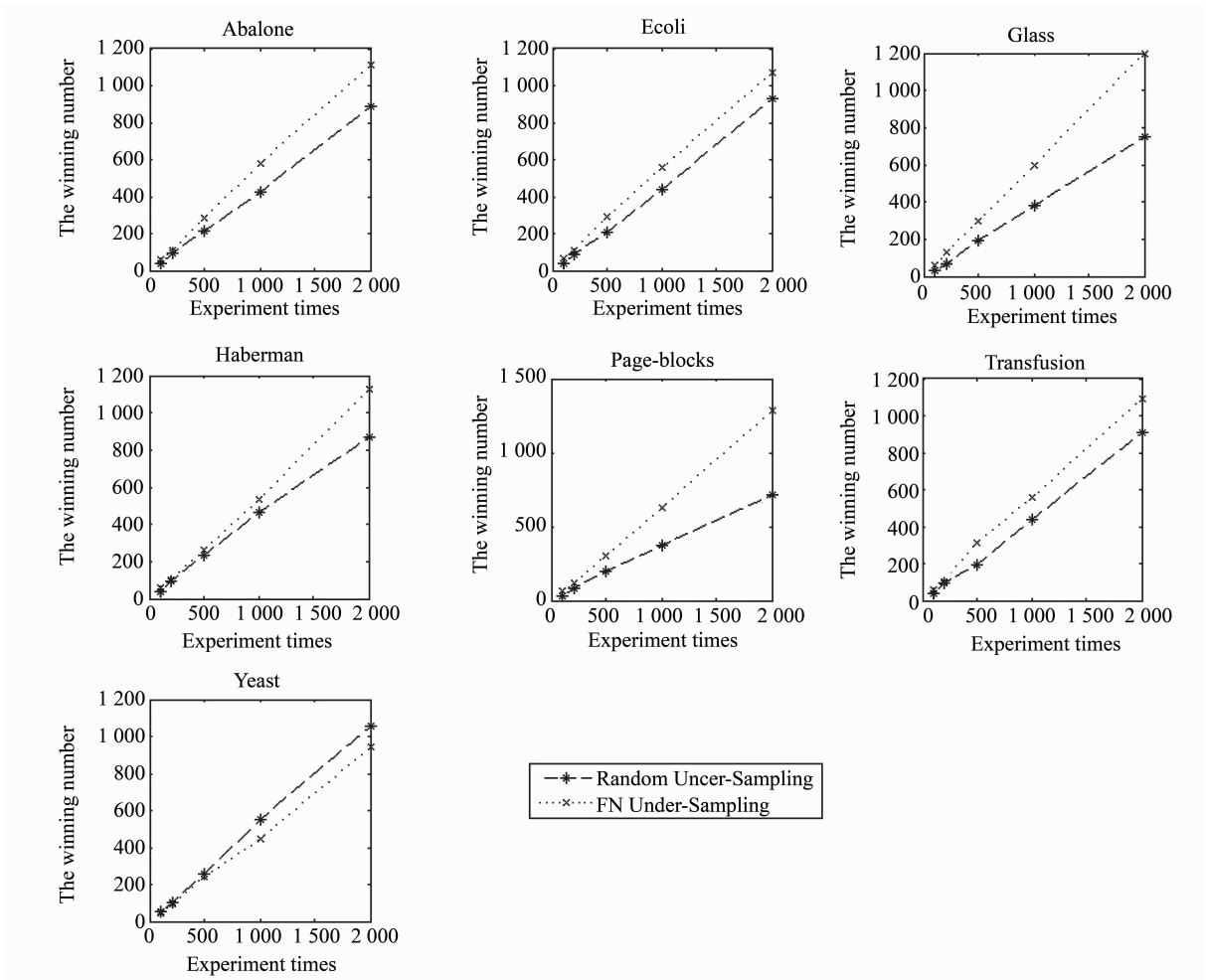

图 2　随机欠采样和改进欠采样方法在五组重复实验中胜出次数比较

Fig. 2　Comparison of winning numbers between Random Under-Sampling and FN Under-Sampling

根据以上分析可以说明，对低维数据使用本文提出的 FN 欠采样方法更有意义，实验中几组低维数据效果较好的现象恰好说明了这点。对于高维数据来说，由于有更多的特征，受“不相关”特征的影响而让该方法失效的可能性越大。

需要注意的是，本文的 FN 欠采样方法对于不平衡比值为整数的数据集，能使训练集中两类样本完全平衡。对于训练集不平衡比值不为整数的情况，可以将改进后的欠采样方法与随机欠采样相结合，先用改进方法使两类样本数尽量接近，再由随机欠采样进一步平衡，在最大程度上提高分类器稳定性。

5 结 语

本文在分析支持向量机、欠采样方法基本思想的基础上，讨论了不平衡数据分类优劣的评价标准，进而针对随机欠采样方法可能出现的问题，提出了一种适用于支持向量机的FN欠采样方法。该方法基于均值样本以及样本间距离实现。UCI数据实验结果表明，对于低维数据和一部分高维数据来说，FN欠采样方法有较好的效果，避免了随机欠采样方法偶然性带来的“失真”影响，提高了支持向量机分类不平衡数据的稳定性。然而在对一些比较特殊的高维不平衡数据进行分类时，由于受维度影响，FN欠采样方法表现不够理想。在进一步的工作中，将针对这种影响进行研究，并结合特征选择算法提出相应的解决方法。

参考文献：

[1] WEISS G M. Mining with rarity: A unifying framework [J]. ACM SIGKDD Explorations Newsletter-Special issue on learning from imbalanced datasets, 2004, 6(1): 7 – 19.

[2] HE H B, GARCIA. Learning from imbalanced data [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263 – 1284.

[3] CHAN P K, STOLFO S J. Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection [C]//The Fourth International Conference on Knowledge Discovery and Data Mining, 1998: 164 – 168.

[4] JAPKOWICZ N, STEPHEN S. The class imbalance problem: A systematic study [J]. Intelligent Data Analysis, 2002, 6(5): 429 – 449.

[5] PRATI R C, BATISTA G, MONARD M C. Class imbalances versus class overlapping: an analysis of a learning system behavior [C]//MICAI 2004: Advances in Artificial Intelligence, 2004: 312 – 321.

[6] CORTES C, VAPNIK V. Support-vector networks [J]. Machine Learning, 1995, 20(3): 273 – 297.

[7] QUINLAN J R. C4.5 programs for machine learning [M]. San Mateo, Calif: Morgan Kaufmann Publishers, 1993.

[8] TANG Y C, ZHANG Y Q, CHAWLA N V, et al. SVMs modeling for highly imbalanced classification [J]. IEEE Transactions on Systems, Man, and Cybernetics, 2009, 39(1): 281 – 288.

[9] WANG X H, SHU P, CAO L, et al. A ROC curve method for performance evaluation of support vector machine with optimization strategy [C]//IFCSTA 09. 2009, 2: 117 – 120.

[10] PRATI R C, BATISTA G, MONARD M C. A study of the behavior of several methods for balancing machine learning training data [J]. ACM SIGKDD Explorations Newsletter-Special issue on learning from imbalanced datasets, 2004, 6(1): 20 – 29.

[11] BATUWITA R, PALADE V. FSVM-CIL: fuzzy support vector machines for class imbalance learning [J]. IEEE Transactions on Fuzzy Systems, 2010, 18(1): 558 – 571.

[12] DRUMMOND C, HOLTE R C. C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling [C]//The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2001: 198 – 207.

[13] SAEYS Y, INZA I, LARRANAGA P. A review of feature selection techniques in bioinformatics [J]. Oxford Journals: Bioinformatics, 2007, 23(19): 2507 – 2517.

[14] TED W W, SAHINER B, HADJIISKI L M, et al. Effect of finite sample size on feature selection and classification: a simulation study [J]. Medical Physics, 2010, 37(2): 907 – 920.

[15] TAX D M J, DUIN R P W. Support vector data description [J]. Machine Learning, 2004, 54:45 – 66.

[16] 陈振洲,李磊,姚正安. 基于SVM的特征加权KNN算法 [J]. 中山大学学报:自然科学版, 2005,44: 17 – 20.

[17] KHOSHGOFTAAR T M, GAO K H. Feature selection with imbalanced data for software defect prediction [C]//ICMLA ‘09, 2009: 235 – 240.