结合 Rotation Forest 和 MultiBoost 的 SVM 集成方法

姚 旭 王晓丹 张玉玺 毕 凯

(空军工程大学防空反导学院 西安 710051)

摘 要 针对如何提高集成学习的性能,提出一种结合 Rotation Forest 和 MultiBoost 的集成学习方法——利用 Rotation Forest 中旋转变换的思想对原始数据集进行变换,旨在增加分类器间的差异度;利用 MultiBoost 在变换后的数据集上训练基分类器,旨在提高基分类器的准确度。最后用简单的多数投票法融合各基分类器的决策结果,将其作为集成分类器的输出。为了验证该方法的有效性,在公共数据集 UCI 上进行了实验,结果显示,该方法可获得较高的分类精度。

关键词 集成学习,支持向量机,随机投影,旋转森林,MultiBoost

中图法分类号 TP391 **文献标识码** A

Algorithm for SVM Ensemble Based on Rotation Forest and MultiBoost

YAO Xu WANG Xiao-dan ZHANG Yu-xi BI Kai

(Air Defense Anti-missile College, Air Force Engineering University, Xi'an 710051, China)

Abstract To improve the performance of ensemble learning, an ensemble algorithm which is a combination of Rotation Forest and MultiBoost was proposed as follows: To improve the diversity between classifiers, rotation transformation in rotation forest is introduced to model the new data set, and for higher accuracy, each classifier is trained by MultiBoost on the transformed data set. Finally, majority voting method is utilized to fusion the base classifiers' recognition results. To attest the validity, we made experiments on UCI data sets. The experimental results suggest that our algorithm can get higher classification accuracy.

Keywords Ensemble learning, Support vector machine, Random projection, Rotation forest, MultiBoost

集成学习由于能够显著提高学习系统的泛化性能[1],因此受到了越来越多的关注,已成为模式识别和机器学习领域研究的热点问题。目前,集成学习已经被成功应用于基因数据分析[2,3]、遥感数据分析[4]、图像处理[5]等很多实际应用问题。常用的集成方法有 Bagging[6]、AdaBoost[7]、Random Subspace[8]、Rotation Forest[9]等。在这些方法中,AdaBoost 以其简单、适应性强成为比较流行的一种,并且出现了很多变种,如 MultiBoost[10]等。MultiBoost 是 Bagging 和 AdaBoost 的融合,通过对 AdaBoost 所产生的分类器采用 Wagging(其中 Wagging 是 Bagging 的一个变种)形式的加权机制来实现。Webb[10]用实验证明 MultiBoost 的平均分类误差比 Bagging、Wagging、AdaBoost 的分类误差要低。基于主成分分析(Principal Component Analysis, PCA), Rodríguez[9]等提出了一种新的集成方法——旋转森林(Rotation Forest),并且在 UCI 数据集上证明了该方法的性能优于其它几种方法。因为它在增加分类器间差异度的同时也提高了基分类器的精确度。

通过以上分析,鉴于 MultiBoost 和 Rotation Forest 这两种集成方法的优点,本文试图将二者结合,提出一种基于 MultiBoost 和 Rotation Forest 的 SVM 集成方法 RotMBoost,以期获得更小的预测误差。SVM 集成是集成学习的一种具体实现,Kim 等人[11]首先将 Bagging 集成技术引入到 SVM 分类中,发现 SVM 集成可获得比单 SVM 更好的性能。SVM 集成虽然得到了广泛的应用[12-15],但由于 SVM"稳定"和"高精度"的特点,增加了 SVM 集成的难度。文献[16]指出构造差异性大的集成成员分类器是提高 SVM 集成泛化性能的有效途径。因此本文从如何提高基分类器间的差异性出发,提出一种新的 SVM 集成方法。我们知道,Rotation Forest 是一种基于决策树的集成方法,并且在决策树多分类器集成的应用中效果显著[3,17]。由于 SVM 是一种稳定的分类器,在应用 MultiBoost 和 Rotation Forest 这种通过扰动样本训练基分类器的集成方法上效果可能不是很明显。因此本文将用随机投影(Random Projection, RP)代替 PCA,试图利用 RP 固有的随机性来提高 SVM 集成的性能。

1 旋转森林算法

旋转森林(Rotation Forest)的基本设计思想是利用 PCA 对原来的特征轴进行旋转,使得每个基分类器得到不同的训练集。这种做法一方面可以保证基分类器之间的多样性,另一方面,可通过保留所有的主成分并利用整个数据集训练每个基分类器来保证每个基分类器的准确性。假定数据集 $D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_N,y_N)\}$,其中 $\boldsymbol{x}_i\in\boldsymbol{X}_{N\times n}$, $y_i\in\boldsymbol{Y}_{N\times 1}$。

到稿日期:2012-05-31 返修日期:2012-09-09 本文受国家自然科学基金项目(60975026,61273275,61102109)资助。

姚 旭(1982-),女,博士生,主要研究方向为智能信息处理、机器学习等;王晓丹(1966-),女,教授,博士生导师,主要研究方向为智能信息处理、机器学习等。

设 N 为样本个数，n 为特征维数，特征集为 F，基分类器个数为 L，特征集随机分割的块数为 K，旋转森林的基分类器生成步骤如下：

第 1 步　将特征集 F 随机分割成 K 个特征子集。假设每个子集均包含 M 个特征，则 $M=n/K$。

第 2 步　令 F_{ij} 表示用于训练分类器 C_i 的第 j 个特征子集，X_{ij} 为 $\boldsymbol{X}$ 中只包含特征子集 F_{ij} 的样本子集。从 X_{ij} 中的所有样本中随机地抽取 75% 的样本，以构建一个新的样本集 X'_{ij}。之后，对样本集 $X_{ij}{}'$ 采用某种线性变换，以生成一个系数矩阵 C_{ij}。

第 3 步　使用矩阵 C_{ij} 的系数构造一个系数旋转矩阵 R_i，矩阵 R_i 中的各列需要按原始特征集的顺序重新排序，将重新排序后得到的矩阵记为 R_i^a。则对分类器 C_i，用于训练的样本需使用旋转矩阵进行旋转变换，即变换后新的训练集为 $\boldsymbol{X}'=\boldsymbol{X}R_i^a$。

第 4 步　对测试样本分类，融合基分类器的分类结果，输出最终决策。

需要说明的是，在上述算法将特征集 F 随机分为 K 个子集的过程中，这 K 个子集之间可以相交也可以不相交。为了最大化基分类器之间的差异性，文献[9]中采用的是不相交的子集。同时，为了简单，假定 K 是输入特征总数 n 的一个因子，即使得每个特征子集中包含 $M=n/K$ 个特征。此外，算法中选择 X'_{ij} 的样本容量为 X_{ij} 的样本容量的 75%，这是为了避免当不同的基分类器选择了相同的特征子集时，最终不至于得到完全相同的训练集。同时这也增加了基分类器之间的差异性。

2　融合 Rotation Forest 和 MultiBoost 的 SVM 集成方法

2.1　线性变换

在分类问题中，投影技术被广泛地应用于约减特征维数。很多投影方法能够使数据在投影空间中保持原始结构，因此利用投影后的数据训练分类器，不仅能够加速训练，有时也能避免噪声的干扰或者“过拟合”现象。PCA 或许是最流行的投影方法，但其主要缺点是计算复杂度较高。在基于 Rotation Forest 的决策树集成系统中，每一个基分类器都利用 PCA 对数据集进行旋转变换。然而，Rotation Forest 就是利用经过投影的训练集来训练不同的基分类器。这种训练过程中产生的差别并不能够生成彼此之间差异度很大的基分类器，但是个体分类器精度较高。

与 PCA 相比，随机投影(Random Projection，简称 RP)是一种运算成本相对较低且实现简单的投影方法。它能够降低模型构建和分类的复杂度，同时维数约减去除了噪声的干扰，提高了分类精度。RP 用于集成基于两个方面，一是通过投影变换，为集成模型的构建提供差异度；二是 RP 将原始数据集投影到一个低维空间，并保持了原始数据集的几何结构，从而保证了模型的精确度。在利用 RP 进行降维的过程中，假设原始数据集为 $\boldsymbol{X}_{N\times n}$(其中 N 为样本个数，n 为特征维数)，通过 $n\times d$ 随机矩阵 R 将其投影到 d 维空间上，则有：

$$\boldsymbol{X}'_{N\times n}=\boldsymbol{X}_{N\times n}\times R_{n\times d} \tag{1}$$

式中，$\boldsymbol{X}'_{N\times n}$ 为 $\boldsymbol{X}_{N\times n}$ 在 d 维子空间上的投影，即将数据集从 n 维降到了 d 维。根据 Johnson-Lindenstrauss 引理可知，如果一个向量空间中的数据点被随机投影到一个具有一定维数的子空间上，那么数据向量间的相似性近似保持不变。使用上述随机投影算法的关键在于确定随机矩阵 R 和低维空间的维数 d。矩阵 R 的元素 r_{ij} 的确定方法有 3 种：

(1) 每一个元素都来自高斯随机数发生器，文中记作 Gaussian；

(2) $r_{ij}=\sqrt{3}\times x, x\in\{-1,0,1\}$，其中取 -1 和 1 的概率均为 1/6，取 0 的概率为 2/3，文中记作 Sparse；

(3) $r_{ij}\in\{-1,1\}$，其中取 -1 和 1 的概率均为 1/2，文中记作 Binary。

另一个问题是确定低维空间的维数 d。理论研究说明，k 阶高斯混合函数空间中的向量被投影到 $O(\log k)$ 维子空间中仍能保持近似的分类特性。由于直接利用的经验很少，而且初始数据的分布也是未知的，故常采用实验的方法来确定低维空间的维数 d。

SVM 是一种高精度且稳定的分类器，对训练集上的微小扰动不敏感。利用经典的集成方法很难得到优于单分类器的集成系统。因此，在 SVM 集成中，我们试图利用 RP 固有的随机性作为产生差异性的来源，并旨在提高 SVM 的精度。因此，在这里利用 RP 的目的并不是对数据进行降维，而是为了提高 SVM 集成的性能。在文献[18]的研究中已经证明采用 Binary 的 Rot-RP125%(其中 125% 为变换后特征的维数与原始特征维数的比率)在 SVM 集成中的效果最好。因此本文采用同样的方法生成旋转矩阵。

2.2　算法描述

在集成系统的设计中，差异性和准确性是两个关键因素，即一方面应该保证每个基分类器的准确性；另一方面，应该使得每个基分类器的错误尽可能不相关。Krogh 和 Vedelsby 也指出，如果可以在不影响基分类器误差的情况下增加它们之间的差异性，则可以进一步改善集成分类器的泛化性能。MultiBoost 是 Bagging 和 AdaBoost 的合成，它充分利用 Bagging 和 AdaBoost 各自对方差和偏差的影响来进一步降低分类误差。因此 MultiBoost 用于分类器集成可以保证基分类器的准确性。同时，RP 是一种简单的投影方式，它能够降低模型的计算复杂度，同时它天生的不稳定性和随机性可以作为产生差异性的来源。我们在本节提出一种新的集成方法 RotMBoost。该方法的主要思想是利用 Rotation Forest 中的旋转变换技术对原始训练样本进行变换，然后在变换后的数据集上利用 MultiBoost 来训练基分类器，最后以简单的多数投票法融合各基分类器的预测结果，并将其作为集成分类器的输出。同时，由于 Rotation Forest 的算法复杂度与数据特征维数直接相关，因此如果能事先去除噪声特征的干扰，算法的性能将会有很大程度的提高。特征选择的方法有很多，在这里只是对数据作一个预处理，所以采用经典的特征选择方法 ReliefF 去除无关特征。RotMBoost 方法的具体描述如下：

训练阶段

输入：训练集 $D=\{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),\cdots,(\mathbf{x}_N,y_N)\}=[\mathbf{X}\quad\mathbf{Y}]$，$\mathbf{x}_i\in\mathbf{X}_{N\times n}$，$y_i\in\mathbf{Y}_{N\times 1}$。特征集随机分割的块数 K，基学习算法 W，Rotation Forest 的迭代次数 S，MultiBoost 的迭代次数 T，ReliefF 的迭代次数 P，近邻样本个数 H，特征权重阈值 δ，用于分类的样本点 $\mathbf{x}$；

迭代过程

第 1 步　在训练集 D 上执行 ReliefF 算法，记预处理后的训练集为 $D'=[\mathbf{X}'\quad\mathbf{Y}]$，特征集合为 F'；

第 2 步　For s=1,2,…,S

(1) 生成旋转矩阵 R_s^a，则对于基分类器 C_s 的训练集可表示为 $D^a=[R_s^a\mathbf{X}'\quad\mathbf{Y}]$；

(2) 初始化训练集 D^a 上的权重分布为 $D_1(x_i)=1,i=1,2,\cdots,N$,计数器 k=1;

(3) For t=1,2,…,T

(a) 根据权重分布 D_t,从训练集 D^a 中随机有放回的抽取 N 个样本,组成新的训练集 D_t^a;

(b) 如果 $I_k=t$,随机重置 D_t^a 的权重分布为连续型泊松分布并对数据集进行归一化处理,k=k+1;

(c) 在 D_t^a 上用基学习算法 W 训练基分类器 C_t^a,计算 C_t^a 的训练误差 $\varepsilon_t=\frac{\sum_{x_j\in D_t^a, C_t^a(x_j)\neq y_j} D_t(x_j)}{N}$;

(d) 如果 $\varepsilon_t>0.5$,随机重置 D_t^a 的权重分布为连续型泊松分布并对数据集进行归一化处理,k=k+1,转向(c);

(e) 如果 $\varepsilon_t=0$,设 $\beta_t=10^{-10}$,随机重置 D^t 的权重分布为连续型泊松分布并对数据集进行归一化处理,k=k+1;

(f) 否则,$\beta_t=\frac{\varepsilon_t}{(1-\varepsilon_t)}$,对于每一个 $x_j\in D_t^a$ 更新样本权值:

$$D_{t+1}(x_j)=\begin{cases} D_t(x_j)/2\varepsilon_t, & C_t(x_j)\neq y_j \\ D_t(x_j)/2(1-\varepsilon_t), & C_t(x_j)=y_j \end{cases},$$

如果 $D_{t+1}(x_j)<10^{-8}$,令 $D_{t+1}(x_j)=10^{-8}$;

(4) End For

(5) $C_s(x)=\underset{y\in Y}{\operatorname{argmax}}\sum_{t:C_t^a(x)=y}\log\frac{1}{\beta_t}$;

第 3 步 End For

测试阶段

对于一个新的数据点 x,集成分类器 C* 将其类标签预测为(其中 I(·)为指示函数)

$$C^*(x)=\underset{y\in Y}{\operatorname{argmax}}\sum_{s=1}^{S} I(C_s(x)=y)$$

3 实验结果及分析

3.1 实验数据及参数设置

实验数据均来自 UCI 数据库,选择了其中 10 组数据(特征维数范围为 4～60,样本范围为 150～10992),详细描述如表 1 所列。实验前,对数据进行了归一化处理。

表 1 UCI 数据集各数据描述

数据集	训练集	维数	类别
Iris	150	4	3
Sonar	208	60	2
Glass	214	10	7
Soybean	307	35	19
Ecoli	336	8	8
Ionosphere	351	34	2
Bcw	699	9	2
German	1000	24	2
Segment	2310	19	7
Pendigits	10992	36	6

文献[19]通过实验分析了 Rotation Forest 用于分类器集成时将特征集合随机分割成 K 个特征子集的必要性,并且证明了集成错误率与特征子集的分割块数没有一定的关系。唯一的规则就是当 $K=1$ 和 $K=n$(其中 n 为特征维数)时,集成的错误率较大。只要取得不是太大或太小,识别精度就与 K 值无关。在文献[19]的实验中,$M=3$($M=n/K$,即为分割的特征子集中特征的个数)时集成效果较好,因此文中建议在以后的试验中取 $M=3$。本文采取文献[20]中的方法,特征维数小于 10 时,取 $M=2$;特征维数大于 10 时,取 $M=3$。

设 Rotation Forest 的迭代次数 $S=10$,MultiBoost 的迭代次数 $T=10$,ReliefF 的迭代次数 $P=20$,近邻样本个数 $H=10$,特征权重阈值 $\delta=0.02$。参数的设置参照文献[20,21]。

实验中以 SVM 为分类器,来自 PRTool(http://www.prtools.org)的工具箱,采用径向基核函数(Radial Basis Function,RBF)的 SVM。RBFSVM 有高斯宽度 σ 和规则化参数 C 两个参数,任何一个的改变都将导致分类器性能的改变。通过选择合适的 C 和 σ 可以有效避免过拟合。通过对 RBFSVM 的性能分析发现[22],C 值过小,分类器学习能力不好,但当 C 在一个合适的范围内取值时,RBFSVM 的性能可以简单地通过调整 σ 值而改变,且 σ 对分类器的影响更大。文献[23]分析了在 RBFSVM 中如果 σ 取相同值而带来的一些问题,提出了通过将训练每个基分类器的样本集的标准差作为该基分类器的 σ 值,来控制基分类器的分类精度,避免参数 σ 在所有基分类器中取值相同带来的问题。因此在本文中采用文献[23]的做法,把训练每个基分类器的样本集的标准差作为该基分类器的 σ 值。惩罚因子 $C=1000$。实验机器配置为 2G 内存,2.80G CPU,算法基于 Matlab7.10(R2010a)实现。

在估计分类正确率时,利用双边估计 t 检验法来计算置信水平为 0.95 的分类正确率置信区间作为最终结果。计算公式如下:

$$\frac{|\bar{x}-u|}{\sigma/\sqrt{n}}\geqslant t_{0.025}(n-1) \qquad (2)$$

式中,u,σ 分别表示三重交叉验证的均值和标准差,$t_{0.025}(2)=4.3027$。

3.2 实验结果和分析

为了验证所提方法的性能,在 UCI 数据集上进行了实验,并与单分类器(Single)及 MultiBoost、RotForest、RotBoost(文献[20]中所提方法)3 种集成方法进行了对比。实验从 3 个部分进行讨论:第一部分比较 5 种方法的预测误差;第二部分从偏差方差角度对每种方法进行比较;第三部分比较每种方法的差异性。

3.2.1 预测误差的比较

在实验中,由于所选实验数据没有单独的训练集和验证集,因此采用 3 折交叉验证法来估计预测误差。对每个数据集,先将它随机地分为大小基本一致的 3 个集合,其中 1 个作为验证集,另外两个合在一起作为训练集。在每个数据集上进行 10 次实验,实验结果取 10 次实验的平均值,如表 2 所列。

表 2 5 种方法的预测误差及置信水平为 0.95 的置信区间(%)

Dataset	Single	MultiBoost	RotForest	RotBoost	RotMBoost
Iris	5.65± 1.89	5.82± 1.17	**4.41** ± 1.04	4.55± 0.93	4.45± 1.05
Sonar	30.18± 2.25	18.59± 2.27	17.29± 2.26	17.26± 2.19	**17.19** ± 2.15
Glass	24.55± 0.86	24.43± 0.83	24.45± 0.62	24.48± 0.87	**24.4** ± 0.78
Soybean	11.29± 1.35	**6.54** ± 0.64	6.97± 0.64	**6.56** ± 0.78	**6.49** ± 0.78
Ecoli	19.76± 1.59	**14.93** ± 1.18	16.25± 1.35	15.34± 1.07	14.98± 1.09
Ionosphere	12.43± 1.28	6.35± 0.88	5.60± 0.77	5.64± 0.76	**5.58** ± 0.71
Bcw	5.58± 0.76	**3.19** ± 0.30	2.93± 0.35	**3.27** ± 0.37	**2.92** ± 0.32
German	34.86± 1.34	24.79± 0.70	28.45± 0.82	24.61± 0.85	**24.32** ± 0.79
Segment	4.93± 0.36	1.85± 0.21	2.14± 0.28	**1.80** ± 0.19	1.81± 0.18
Pendigits	4.79± 0.21	0.82± 0.04	0.64± 0.03	0.69± 0.07	**0.60** ± 0.03

表 2 中的黑体表示在每个数据集上预测误差最小的值。从表 2 可以看出，在所有数据集上 RotMBoost 方法的预测误差都小于 Single。同时可以看出 RotMBoost 在大多数数据集上的分类性能都优于 MultiBoost、RotForest 和 RotBoost，尽管有些优势并不显著。在数据集 Iris、Ecoli 和 Segment 上，RotMBoost 虽然不是最好的，但仅次于最好的方法。

此外我们知道没有一种方法在所有可能的分类任务中的泛化性能都优于其它方法，但是在现实世界的分类任务中或至少在其中特定的子集上，某种方法还是有它的相对优势。实验中选择了 UCI 数据集中 10 个不同性质并且来自不同领域的数据集，它们在一定程度上代表了 UCI 数据集。在这里用统计的观点对文中所涉及的分类方法的相对性能进行分析。在文中我们利用文献[10]中提出的一些统计量进行讨论。表 3 给出了在所有数据集上，每种方法的误差比较。表的第一行是每种方法的误差在所有数据集上的平均值。如果用“row”表示表的每一行所列方法的误差，“col”表示表的每一列所列方法的误差，则表 3 中“$\dot{r}$”一行的值表示“row/col”的几何平均值。“s”对应的行给出的是 win/tie/loss 统计量，其中的 3 个值分别表示 col<row，col=row，col>row 的数据集个数。

表 3　每种方法在各个数据集上的误差(Error)比较

Algorithm		Single	Multi-Boost	Rot-Forest	Rot-Boost	RotM-Boost
Mean error (%)		15.40	10.73	10.91	10.42	**10.27**
Single	$\dot{r}$		0.571	0.550	0.538	0.521
	s		9/0/1	10/0/0	10/0/0	10/0/0
MultiBoost	$\dot{r}$			0.961	0.943	0.912
	s			5/0/5	6/0/4	9/0/1
RotForest	$\dot{r}$				0.979	0.947
	s				5/0/5	9/0/1
RotBoost	$\dot{r}$					0.967
	s					9/0/1

从表 3 的实验结果可以看出，RotMBoost 方法具有最小的平均误差。分析平均误差、“$\dot{r}$”和“s” 3 个统计量可以看出，5 种方法按照分类效果由好到差依次为 RotMBoost、RotBoost、RotForest、MultiBoost、Single。单独考虑 RotMBoost 的分类效果，它与 Single 的误差比率的几何平均值比其它方法相对于 Single 的误差比率的几何平均值都小，同时，与 RotBoost、RotForest、MultiBoost 的误差比率的几何平均值相比，本文所提方法的效果比较好。此外，与其它 4 种方法相比，RotMBoost 在更多的数据集上有较小的分类误差。实验结果表明该方法是有效的。

3.2.2　误差的偏差和方差分解

为了进一步比较每种方法的分类效果，我们对误差从偏差和方差分解的角度进行讨论。从理论上讲，1 个分类器的误差应该分解为 3 部分，即内部误差或不可减少的误差(irreducible error)、偏差和方差。而在实际的学习任务中，类的真实分布一般是未知的，很难估计出内部误差。在一个学习任务中，内部误差对几个学习算法是不变的，它不会影响到学习算法之间的相对效率，在文献[10]定义的误差的偏差和方差分解中，内部误差被融入到了偏差和方差中。令 T 为训练集的分布，D 为来自分布 T 的训练样本，L 为基学习算法，$L(D)$ 表示用 L 训练的分类器，则对于验证样本点(x,y)，其偏差和方差的定义为：

$$\begin{cases} Bias(\boldsymbol{x}) = P_{D\sim T}(L(D)(\boldsymbol{x}) \neq y \ \&\ L(D)(\boldsymbol{x}) = y^*) \\ Var(\boldsymbol{x}) = P_{D\sim T}(L(D)(\boldsymbol{x}) \neq y \ \&\ L(D)(\boldsymbol{x}) \neq y^*) \end{cases} \tag{3}$$

式中，y^* 是由来自 T 的不同训练集 D 训练的分类器对样本 $\boldsymbol{x}$ 所预测的类标签的中心趋势(central tendency).

为了计算偏差和方差的值，首先需要知道训练数据的真实分布 T。由于在实际问题中，T 一般是未知的，因此只能用某种方法来估计偏差和方差。在此采用 Webb[10] 所提方法来估计偏差和方差。具体步骤如下：首先将数据 D 随机地分为 3 个大小基本一致的集合 f_1, f_2, f_3，并将该过程重复进行 10 次得到 30 个不同的集合 $f_1^1, f_2^1, f_3^1, f_1^2, f_2^2, f_3^2, \cdots, f_1^{10}, f_2^{10}, f_3^{10}$。在每次实验中，将 3 个集合中的每个集合都用作一次验证集，与其对应的另外两个集合用作训练集。为了简化记号，令 $D_1^i = f_2^i \cup f_3^i$，$D_2^i = f_1^i \cup f_3^i$，$D_3^i = f_1^i \cup f_2^i$，则对于 $x \in D$，其类标签的中心趋势可以估计为：

$$\operatorname{argmax}\left(\sum_{i=1}^{10}\sum_{j=1}^{3} I(\boldsymbol{x} \in f_j \ \&\ T(D_j^i)(\boldsymbol{x}) = y)\right) \tag{4}$$

表 4 给出了每个数据集上各种方法所对应的偏差和方差。在某些情况下，偏差和方差之和并不严格等于相应的误差，主要是因为对一些值进行了四舍五入。

表 4　每种方法在各个数据集上的方差和偏差

Dataset	偏差					方差				
	Single	MulB	RotF	RotB	RotMB	Single	MulB	RotF	RotB	RotMB
Iris	3.49	4.72	3.52	3.71	3.69	2.16	1.10	0.89	0.84	0.76
Sonar	14.45	11.71	10.93	10.14	11.05	15.74	6.88	6.36	7.13	6.14
Glass	23.32	23.18	23.37	23.10	22.99	1.23	1.25	1.08	1.38	1.05
Soybean	5.81	4.53	4.62	4.46	4.38	5.48	2.01	2.36	2.10	2.11
Ecoli	11.74	10.95	11.46	11.76	11.50	8.02	3.98	4.79	3.58	3.48
Ionosphere	7.58	5.21	4.69	4.03	4.05	4.85	1.14	0.91	1.61	1.53
Bcw	3.02	2.64	2.51	2.56	2.50	2.57	0.55	0.42	0.71	0.42
German	21.45	19.23	24.11	19.14	19.10	13.41	5.56	4.34	5.47	5.22
Segment	2.10	1.12	1.27	1.15	1.11	2.83	0.73	0.87	0.65	0.70
Pendigits	1.54	0.52	0.50	0.44	0.44	3.25	0.30	0.14	0.25	0.16

为了更直观地比较每种方法对偏差和方差的减小程度，我们也采用统计量进行分析。表 5 和表 6 给出了每种方法在各个数据集上的偏差和方差比较。表中各行各列代表的意义与表 3 相同。

分析表 5 和表 6 可以看出，RotMBoost 对偏差的减小优于 Single、MultiBoost、RotForest，与 RotBoost 相差不多。5 种方法对偏差的减小程度由好到差依次为 RotBoost、RotMBoost、RotForest、MultiBoost、Single。RotMBoost 对方差的减小是 5 种方法中最好的，对方差的减小程度由好到差依次为 RotMBoost、RotForest、MultiBoost、RotBoost、Single。基于表 5 和表 6 的实验结果可以得出结论，RotMBoost 比 RotForest、MultiBoost 更好地降低了偏差和方差。尽管在改善基

分类器偏差方面 RotMBoost 不如 RotBoost，但 RotMBoost 对方差的减小要优于 RotBoost，因而在几种方法中，RotMBoost 的预测误差是最小的。

表 5　每种方法在各个数据集上的偏差(Bias)比较

Algorithm		Single	Multi-Boost	Rot-Forest	Rot-Boost	RotM-Boost
Mean error (%)		9.45	8.38	8.70	**8.05**	8.08
Single	$\dot{r}$		0.774	0.764	0.717	0.715
	s		9/0/1	8/0/2	8/0/2	9/0/1
MultiBoost	$\dot{r}$			0.987	0.926	0.924
	s			6/0/4	8/0/2	9/0/1
RotForest	$\dot{r}$				0.937	0.937
	s				7/0/3	8/0/2
RotBoost	$\dot{r}$					0.998
	s					7/1/2

表 6　每种方法在各个数据集上的方差(Variance)比较

Algorithm		Single	Multi-Boost	Rot-Forest	Rot-Boost	RotM-Boost
Mean error (%)		5.95	2.35	2.22	2.37	**2.16**
Single	$\dot{r}$		0.340	0.295	0.343	0.294
	s		9/0/1	10/0/0	9/0/1	10/0/0
MultiBoost	$\dot{r}$			0.856	1.009	0.853
	s			7/0/3	5/0/5	8/0/2
RotForest	$\dot{r}$				1.163	0.997
	s				4/0/6	6/1/3
RotBoost	$\dot{r}$					0.887
	s					8/0/2

3.2.3　差异性比较

为了得出更具统计意义的实验结论，利用秩和检验法对上面的结果进行分析。其中秩水平计算如下：

$$R_j=\frac{1}{J}\sum_i r_i^j \tag{5}$$

式中，r_i^j 为在第 i 个数据集上用第 j 种方法所得到的秩大小，J 为每种方法所进行的实验次数，文中 $L=10$。表 7 给出了每种方法在误差、偏差和方差 3 个方面所得到的秩和平均数。

表 7　各种方法秩和平均数比较

Algorithm	Single	Multi-Boost	Rot-Forest	Rot-Boost	RotM-Boost
Error	4.9	3.1	3.0	2.8	**1.3**
Bias	4.3	3.3	3.2	2.4	**1.7**
Variance	4.8	3.1	2.4	3.0	**1.7**
Global Rank	4.67	3.17	2.87	2.73	**1.57**

从表 7 可以看出 RotMBoost 所得到的秩和平均数最小为 1.57，RotBoost 次之，Single 最大。为了验证这 5 种方法的分类效果具有统计意义上的显著差别，我们采用了 Nemenyi 检验方法，即两种方法具有显著性差异，当此两种方法的秩和平均差大于临界值 CD(critical difference value)[24]：

$$CD=q_\alpha\sqrt{\frac{k(k+1)}{6J}} \tag{6}$$

式中，q_α 可通过查询“The Studentized Range Statistic”表得到，k 为所要验证的方法数，J 为每次实验的次数。在本实验中比较了 5 种方法在置信水平为 $\alpha=0.05$ 下的分类效果，如表 7 所列，即 $k=5$，$q_{0.05}=1.860$，代入式(6)可得差异临界值(CD)为 1.315。观察表 7 可知 RotMBoost 的秩平均数比其余方法秩平均数都要小且差值都大于差异临界值，因此可以说 RotMBoost 在 95%的置信区间都要好于其它方法。

结束语　本文提出了一种新的结合 RotForest 和 MultiBoost 的集成方法 RotMBoost，其主要是利用 Rotation Forest中旋转变换的思想来增加基分类器间的差异性，利用 MultiBoost 增加基分类器的准确度。该方法首先利用 ReliefF 对数据集进行预处理，有效地降低了算法的计算复杂度。由于 SVM 是一种稳定的分类器，对样本扰动不敏感，因此在 Rotation Forest 中，作者利用 RP 取代 PCA 生成旋转矩阵，试图利用 RP 的随机性提高 SVM 集成的性能。为了验证本文所提方法的有效性，在 UCI 数据集上进行实验，利用多数投票法融合各基分类器的输出结果。实验结果表明该方法能够获得低于 RotForest 和 MultiBoost 的预测误差，并且能够有效地减小偏差和方差。因此，本文提出的方法是有效的。

参 考 文 献

[1] Dietterich T G. Machine learning research: four current directions [J]. AI, Magazine, 1997, 18(4): 97-136

[2] Dettling M. BagBoosting for tumor classification with gene expression data [J]. Bioinformatics, 2004, 20(18): 3583-3593

[3] Liu Kun-hong, Huang De-shuang. Cancer classification using Rotation Forest [J]. Computers in Biology and Medicine, 2008, 38: 601-610

[4] Ceamanos X, Waske B, Benediktsson J A, et al. Ensemble strategies for classifying hyperspectral remote sensing data [C]// Benediktsson J A, Kittler J, Roli F, eds. Multiple Classifier Systems, Lecture Notes in Computer Science. 2009, 5519: 62-71

[5] Kim T K, Arandjelovic O, Cipolla R. Boosted manifold principal angles for image set-based recognition [J]. Pattern Recognition, 2007, 40(9): 2475-2484

[6] Breiman L. Bagging predictors [J]. Machine Learning, 1996, 24 (2): 123-140

[7] Freund Y, Schapire R E. Experiments with a new boosting algorithm[C] // Proc. 13th International Conference on Machine Learning. Morgan Kaufmann, Bari, Italy, 1996: 148-156

[8] Ho T K. The random subspace method for constructing decision forests [J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 1998, 20(8): 832-844

[9] Rodríguez J J, Kuncheva L I, Alonso C J. Rotation Forest: A New Classifier Ensemble Method [J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2006, 28(10): 1619-1630

[10] Webb G I. MultiBoosting: A Technique for Combining Boosting and Wagging [J]. Machine Learning, 2000, 40: 159-196

[11] Kim H C, Pang S, Je H M, et al. Pattern classification using support vector machine ensemble[C] // Proceedings of the 16th International Conference on Pattern Recognition. Los Alamitos CA, 2002: 1276-1283

[12] Chali Y, Hasan S A, Joty S R. A SVM-Based Ensemble Approach to Multi-Document Summarization [C] // Advances in Artificial Intelligence. Canadian, 2009: 199-202

[13] Ye Yan-fang, Chen Li-fei, Wang Ding-ding, et al. SBMDS: an interpretable string based malware detection system using SVM ensemble with bagging [J]. Journal in computer, 2009, 5: 283-293

[14] Wang Xiao-dan, Zheng Chun-ying, Yao Xu, et al. Multi-polarized HRRP classification by SVM ensemble [C] // ISCIDE: Lecture Notes in Computer Science. volume 7202/2012, Xi'an, 2012: 184-192

(下转第 290 页)

的隐私保护 k-means 聚类方法相比，IDP k-means 聚类方法在聚类可用性和隐私保护级别两方面，取得了更好的平衡。下一步将深入研究在相同的隐私保护级别下，如何进一步提高聚类的可用性，特别是如何提高小数据集差分隐私保护下的聚类可用性，并将继续研究其它聚类算法的差分隐私保护方法。

参 考 文 献

[1] Blum A, Dwork C, McSherry F, et al. Practical Privacy: The SuLQ Framework[C]//24th ACM SIGMOD International Conference on Management of Data / Principles of Database Systems, Baltimore (PODS 2005). Baltimore, Maryland, USA, June 2005

[2] Dwork C. Differential Privacy[C]//33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006). Venice, Italy, Springer Verlag, July 2006

[3] Dwork C. Differential Privacy: A Survey of Results[C]//Theory and Applications of Models of Computation (TAMC2008). Xi'an, China, Springer Verlag, April 2008

[4] Dwork C. The Differential Privacy Frontier[C]// 6th Theory of Cryptography Conference (TCC 2009). San Francisco, CA, Springer Verlag, March 2009

[5] Dwork C. Differential Privacy in New Settings[C]//Symposium on Discrete Algorithms (SODA), Society for Industrial and Applied Mathematics. Austin, TX, January 2010

[6] Dwork C. A Firm Foundation for Private Data Analysis [J]. Communications of the ACM, 2011, 54(1): 86-95

[7] Dwork C. The Promise of Differential Privacy. A Tutorial on Algorithmic Techniques[C]//52nd Annual IEEE Symposium on Foundations of Computer Science. Palm Springs, CA, October 2011

[8] Agrawal R, Strikant R. Privacy-preserving data mining [C]// Proceedings of the 2000 ACM SIGMOD International Conference on Managementof Data. Dallas, Texas, May 2000: 439-450

[9] Sweeney L. K-anonymity: A Model for Protecting Privacy[J]. International Journal on Uncertainty[J]. Fuzziness and Knowledge-based Systems, 2002, 10(5): 557-570

[10] Lindell Y, Pinkas B. Privacy preserving data mining[C]// Proceedings of the 20th Annual International Cryptology Conference on Advances in Cryptology. Santa Barbara, California, August 2000: 36-54

[11] 杨维嘉. 在数据挖掘中保护保护隐私信息的研究[D]. 上海：上海交通大学，2009：13-29

[12] Fienberg S E, Mclntyre J. Data swapping: Variations on a theme by Dalenius and Reiss[C]// Proceedings of the Privacy in Statistical Databases(PSD). Barcelona, Spain, 2004: 14-29

[13] Kifer D, Gehrke J. Injecting utility into anonymized data-sets[C]// Proceedings of the ACM SIGMOD Conference on Management of Data (SIGMOD). Atlanta, Georgia, USA, 2006: 217-228

[14] Agrawal R, Srikant R. Privacy preserving data mining[C]//Proceedings of the ACM SIGMOD Conference on Management of Data(SIGMOD). Dalias, Texas, 2000: 439-450

[15] Du W, Zhan Z. Using randomized response techniques for privacy-preserving data mining[C] // Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington D C, USA, August 2003: 505-510

[16] Clifton C, Kantarcioglou M, Lin X, et al. Tools for privacy preserving distributed data mining[J]. ACM SIGKDD Explorations, 2002, 4(2): 28-34

[17] Oliveira S R M, Zaiane O R. Achieving privacy preservation when sharing data for clustering[C]// Secure Data Management Proceedings. Toronto, Canada, Berlin: Springer, 2004: 67-82

[18] Mukherjee S, Chen Zhi-yuan, Gangopadhyay A. A privacy preserving technique for Euclidean distance-based mining algorithms using Fourier-related transforms [J]. The International Journal on Very Large Data Bases, 2006, 15(4): 293-315

[19] Parameswaran R, Blough D M. Privacy preserving data obfuscation for inherently clustered data[J]. International Journal of Information and Computer Security, 2008, 2(1): 1744-1765

[20] 崇志宏，倪巍伟，刘腾腾，等. 一种面向聚类的隐私保护数据发布方法[J]. 计算机研究与发展，2010，47(12)：2083-2089

[21] Witten I H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques[M]. Morgan Kaufmann, 2005

[22] van Rijsbergen C J. Information Retrieval (2nd edition)[M]. London: Butterworths, 1979

(上接第 270 页)

[15] Liu Man-hua, Zhang Dao-qiang, Shen Ding-gang. Ensemble sparse classification of Alzheimer's disease[J]. Neurolmage, 2012, 2(60): 1106-1116

[16] 王晓丹，高晓峰，姚旭，等. SVM 集成研究与应用[J]. 空军工程大学学报：自然科学版，2012，13(2)：84-89

[17] Alonso-Gonzalez C J, Moro-Sancho Q I, Ramos-Munoz I, et al. Rotation Forest on Microarray Domain: PCA versus ICA[C]// IEA/AIE 2010, Part II. LNAI 6097, 2010: 96-105

[18] Maudes J, Rodrlguez J J, Garcla-Osorio C, et al. Random Projections for SVM Ensembles[C]//Garcla-Pedrajas N, et al., eds. IEA/AIE 2010, Part II. LNAI 6097, 2010: 87-95

[19] Kuncheva L I, Rodrlguez J J. An Experimental Study on Rotation Forest Ensembles[C]// Haindl M, Kittler J, Roli F, eds. MCS 2007. LNCS 4472, 2007: 459-468

[20] Zhang Chun-xia, Zhang Jiang-she. RotBoost: A technique for combining Rotation Forest and AdaBoost [J]. Pattern Recognition Letters, 2008, 29: 1524-1536

[21] Yang Fei-hu, Cheng Wei-qing, Dou Ren-fu, et al. An Improved Feature Selection Approach Based on ReliefF and Mutual Information[C] // International Conference on Information Science and Technology. Nanjing, China, 2011: 246-250

[22] Valentini G, Dietterich T G. Bias-variance Analysis of Support Vector Machines for the Development of SVM-Based Ensemble Methods [J]. Journal of Machine Learning Research, 2004, 5: 725-775

[23] 王晓丹，孙东延，郑春颖. 一种基于 AdaBoost 的 SVM 分类器[J]. 空军工程大学学报：自然科学版，2006，7(6)：54-57

[24] Demsar J. Statistical Comparisons of Classifiers over Multiple Data Sets[C]//J. Machine Learning Research. 2006, 7: 1-30