采用半随机特征采样算法的中文书写纹识别研究

黎冬媛[1] 刘 智[2] 刘三妍[2] 孟文婷[3]
(电子科技大学中山学院计算机学院 中山 528402)[1]
(华中师范大学国家数字化学习工程技术研究中心 武汉 430079)[2]
(华中师范大学计算机科学系 武汉 430079)[3]

摘 要 N-gram字符序列能有效捕捉文本中作者的个体风格信息,但其特征空间稀疏度高,且存在较多噪音特征。针对该问题,提出一种基于半随机特征采样的中文书写纹识别算法。该算法首先采用一种离散度准则为每个作者选取一定粒度的个体特征集,然后将个体特征集以一种半随机选择机制划分成多个等维度的特征子空间,并基于每个子空间训练相应的基分类器,最后采取多数投票法的融合策略构造集成分类模型。在中文真实数据集上与基于随机子空间和Bagging算法的集成分类器进行了对比试验,结果表明,该算法在正确率和差异度方面优于随机子空间和Bagging算法,并且取得了比单分类模型更好的识别性能。

关键词 书写纹,半随机特征采样,个体特征集,集成分类器,差异度
中图法分类号 TP391 **文献标识码** A

Research of Chinese Writeprint Recognition Using Semi-random Feature Sampling Algorithm

LI Dong-yuan[1] LIU Zhi[2] LIU San-ya[2] MENG Wen-ting[3]
(Deptartment of Computer, Zhongshan Institute, University of Electronic Science and Technology of China, Zhongshan 528402, China)[1]
(National Engineering Research Center for E-Learning, Huazhong Normal University, Wuhan 430079, China)[2]
(Department of Computer Science, Huazhong Normal University, Wuhan 430079, China)[3]

Abstract Character N-gram can be used to effectively capture individual-author stylistic information in texts. To deal with the problems of high-sparsity and high-redundancy in the feature space, an ensemble classification algorithm based on semi-random feature sampling was proposed in this study. Firstly, the whole feature space is divided into several individual-author feature sets by a divergence rule. Then each of them is divided into equally sized subspaces by a semi-random selection method, and a base classifier is trained on each random subspace. Finally, these base classifiers are combined to construct an ensemble via the majority voting method. To examine the algorithm, the experiment was conducted on a real-life dataset. It is observes that the algorithm achieved a considerable improvement in accuracy and robustness compared with the benchmark technique in Chinese writeprint identification (random subspace method, bagging and support vector machine).

Keywords Writeprint, Semi-random feature sampling, Individual feature set, Ensemble classifier, Diversity

1 引言

网络通信技术的不断发展促使了网络上交流方式的多样化,为人们提供了更为丰富和便捷的交流工具。很多人更愿意通过网络与其他人进行互动,如传递电子邮件、视频聊天、实时评论等。其中文字类信息以其成本代价低、影响范围广、完成时间短、机动能力强等特征成为互联网用户使用最广泛的一种信息载体,如微博、论坛、博客、电子邮件等Web应用的流行充分表明了这一点。但互联网的开放性和匿名性等特点也引起了各种网络诚信和犯罪问题,例如在线课堂中学生作业的抄袭与剽窃,网站中虚假谣言、色情信息和恐怖言论的发布、恐吓邮件以及其它不良信息的传递等。这些信息的肆意传播损害了个人以及社会的利益。因此,采取有效的身份鉴别方法以锁定不良信息的真实发布者,对于提高用户的诚信度和互联网的和谐发展具有重要的研究价值和实际意义。

传统的采用IP探测器和防火墙过滤等方法已不能有效地解决网络中作者身份的追踪问题。而利用文本中蕴含的风格信息以识别匿名用户的身份成为当今侦查网络犯罪的一种有效手段。相关研究[1]认为,从人的书写特征中传递出来的言语信息能反映个人的意志和行为心理现象。因此通过对书

到稿日期:2012-04-13 返修日期:2012-07-25 本文受国家"核高基"重大专项基金项目(2010ZX01045-001-005),国家"十二五"科技支撑计划项目(2011BAK08B03),广东省教育部产学研结合示范基地项目(2011B090500017),教育部-英特尔信息技术专项科研基金项目(MOE-INTEL-11-02)资助。

黎冬媛(1977—),女,硕士,高级实验师,主要研究方向为计算机应用技术;刘 智(1986—),男,博士生,主要研究方向为智能软件与知识服务、数据挖掘,E-mail:liuzhi8673@gmail.com(通信作者);刘三妍(1973—),男,教授,博士生导师,主要研究方向为计算机应用。

写风格特征进行统计与分析，可判别作者身份，起到类似生物指纹识别的效果。书写纹[2]即是指作者在电子文档中留下的独特的文字风格特征，是标识其身份的特征组合。

书写纹识别实际上是解决一种在高维和稀疏特征空间中的多类单标签的文本分类问题，而该研究的一个子任务是抽取代表作者风格的特征集。根据相关研究[3-5]知，通常用于构建识别模型的特征主要包括词汇、句法、结构以及语义等几类。其中变长的 N-gram(使用 $N=1\sim5$ 多种粒度的字符组合特征集)字符特征能充分利用文本中的每一个字符[6,7]，最大程度地捕捉鉴别性信息，统计的稳定性更好，因此相对而言，其更加适用于网络中短文本类型的文本表示。但 N-gram 在捕捉丰富字符信息的同时，也带来较多冗余特征。特征空间中过多的噪音特征会为识别模型的构建带来很多错误的混淆信息，特别是在中文的环境下，抽取的 N-gram 特征集维数越大，稀疏度越高。这种情况下，传统的单分类模型较难利用特征空间中分散的鉴别信息，并会引起在高维空间中的过训练问题，识别鲁棒性难以保证。

为了解决书写纹识别中的过训练问题，国内外相关学者研究了多种集成分类技术，并取得了很大的进展。代表性研究如 Stamatatos 将集成学习用于英语和希腊语的书写纹识别[8]；并提出了通过随机划分特征空间构建集成分类模型的方法，在划分粒度较小的情况下取得了比支持向量机更好的识别效果，验证了基于特征空间划分的集成学习技术在网络书写纹识别中的有效性。但该方法在特征空间较大的情况下，特征子集的选择粒度与基分类器个数难以权衡，构造的基分类器稳定性并不能得到保证。Koppel 将特征空间随机划分方法与相似性度量结合以识别大规模作者集[9]，该方法将抽取的特征子集分别与待定作者的匿名样本相匹配，并计算匹配得分；同时为每个作者设定阈值，累积匹配得分超过该阈值，则确定该匿名样本的身份归属。该方法能为每个作者统计相似度得分，但难以为每个作者设定一个合理的阈值。国内方面，孙建文将遗传算法与集成分类算法相融合[10]，利用遗传算法搜索不同的特征子集来划分特征空间，形成一种基于搜索算法的集成分类系统。该方法能通过遗传算法自适应地选取具有差异性的特征子空间，但在高维的特征空间中，特别是作者数量较大时，遗传算法的搜索代价太大，识别效率较低，并且容易出现过拟合问题。与以上研究相比，基于作者个体特征选择的书写纹识别技术仍然研究较少，实际上，作者个体特征集中包含较多的鉴别性特征，能为区分不同作者的写作风格提供更多个体差异度信息。

本文提出一种半随机特征采样算法以构造集成分类器，其充分利用特征空间中的个体鉴别信息的分布，首先提取不同作者的个体特征集，这些特征集能充分代表每个作者的独特书写纹风格且为基分类器的构造提供了差异度信息；然后结合一种半随机特征采样方法划分不同的特征子集并构造对应的基分类器；最后采取多数投票法的融合策略结合这些基分类器的预测结果形成一个强分类器以提高中文书写纹识别的性能。

2 相关概念

2.1 随机子空间

文本中包含的 N-gram 字符特征维度较大，经过特征提取后仍然达到上千维，若直接在原始空间再进行降维，可能会丢失某些重要的风格特征。而随机子空间 RS(Random Subspace)算法能较好地保持原始空间的信息完整度。传统的 RS[11]选择是根据均匀分布 U 从原始特征空间 F_d 中随机抽取 k 个不同的子集 $S=\{S_1,S_2,\cdots,S_k\}$，每个子集的维数为 r，对每个子空间都定义一个映射 $P_S:F_d\rightarrow F_k$，在此基础上得到每个训练子集 $T_i=\{(P_S(t_j),y_j)\mid 1\leqslant j\leqslant N\}$；再由某种分类算法 Φ 在该子集上构造基分类器并输出待测样本的决策结果 h_i，此过程重复 m 次；最后利用多数投票法[12]得到最终决策，其中子空间维数 r 和基分类器个数 m 作为参数可自动确定。

2.2 差异性度量

在集成分类器中，不同基分类器之间的差异度被视为衡量集成分类器有效性的一个关键因素。通常，多个完全一致的基分类器(即对待测样本集分类错误率的分布完全相同)不会提高集成系统的分类性能，其分类性能等同于一个单分类模型。由于集成系统中的基分类器的分类能力都是有限的，因此基分类器之间必须存在差异性。

定义 1 在集成分类器中，存在某些分类器对其他分类器分类错误的样本做出正确的判断，这种基分类器之间的互补能力称为差异性。衡量这种差异性的方法称为差异性度量。

目前已有不少度量分类器差异性的方法，但如何定义和度量差异性仍没有一个统一的标准，现有的差异性度量方法主要分为两类：成对差异性度量和非成对差异性度量。本文采用非成对差异性度量中的信息熵度量方法[13]来计算基分类器之间的差异度。

定义 2(信息熵差异性度量方法) 首先度量所有基分类器在一个样本上分类结果的离散度，然后计算所有样本离散度的均值。如式(1)所示。

$$entropy=\frac{1}{T}\sum_{i=1}^{T}\sum_{j=1}^{L}-\frac{N_j^i}{k}\log_L\left(\frac{N_j^i}{k}\right) \tag{1}$$

T 为待测样本的总数，L 为类别个数，k 为基分类器个数，N_j^i 为将第 i 个样本分为第 j 类的基分类器的个数。集成分类器的信息熵越大，基分类器输出结果之间的差异度就越大。

2.3 多分类器融合

定义 3 将所有基分类器对待测样本的识别结果进行投票统计并做决策的过程，称为多分类器融合。

3 基于半随机特征采样的集成分类方案

3.1 问题描述

在书写纹识别中，训练和测试样本均表示为 N-gram 字符组成的向量，设训练集中出现频率最高的 d 个 N-gram 字符的降序排列集合为 $G_d=\{g_1,g_2,\cdots,g_d\}$，$f_{ij}$ 为第 i 个样本中出现 G_d 中第 j 个 N-gram 字符的频率，则每个样本 t_i 可表示为($f_{i1},f_{i2},\cdots,f_{id}$)。

设作者 a_i 的个体风格特征集为 IFS_i^m(Individual Feature Set)，其中包含 m 个 N-gram 字符，且 $IFS_i^m\in G_d(m<d)$。另设 S_{ij}^n 为在 IFS_i^m 上选取的第 j 个随机子空间并包含 n 个 N-gram 字符(其中 $n<m$)。C_{ij} 为在子空间 S_{ij}^n 上训练的基分类器，则基于 $K\times L$ 个特征子集(K 为作者个数，L 为每组个体特征集上选取的子空间个数)的集成分类器可表示为：

$$Ens=ensemble\{C_{ij}(S_{ij}^n),comb,1\leqslant i\leqslant K,1\leqslant j\leqslant L\} \quad (2)$$

式中，$comb$ 为基分类器的融合策略。

3.2 具体方案

根据以上描述，在基于半随机特征采样 SRFS(Semi-Random Feature Sampling)的集成学习方案中，需要确定的参数包括个体特征集的选择粒度 m、随机子空间的大小 n、个体特征集上构造的基分类器个数 L。另外，需要制定随机子空间的选取策略、基分类器的类别以及融合策略。

对于上述问题，本文拟采取以下方案：

(1)将个体特征集的选择粒度 m 作为参数进行考量，分析 m 值对识别不同数量作者集的影响。

(2)计算 SRFS 构造的基分类器之间的差异度，评价其集成分类性能。

(3)随机子空间的选择粒度为 $floor(m/2)$，即个体特征集大小的 50%。对随机子空间中的特征选取设定阈值 η，分为$[0,\eta]$和$[\eta,1]$两个区间分别随机选取，得到 $K\times L$ 个等维度的特征子空间。

(4)基分类器采用基于线性核函数的支持向量机 LSVM (Linear Support Vector Machine)[14]，因其在文本分类领域应用广泛，且在稀疏特征空间中具有的良好泛化能力。

(5)采取多数投票法作为基分类器的融合策略。

$$votes(a_1,a_2,\cdots,a_K)=\mathrm{sgn}\{\sum_{i=1}^{K}\sum_{j=1}^{L}C_{ij}\} \quad (3)$$

式中，基分类器 $C_{ij}\in\{0,1\}$，符号 sgn 代表对基分类器输出结果的投票统计。

4 半随机特征采样算法

4.1 个体特征集选择

在面向较大规模作者集的识别中，传统的基于全局特征选择的方法难以有效地选择出对所有作者都具有区分度的特征集合。从空间复杂度的角度看，全局特征选择的计算复杂度较高，泛化能力差，难以解决高维特征空间中的冗余性问题。本文采取面向作者个体的特征选择方法为每个作者 a_i 选取一组个体特征集 IFS_i^m，使 IFS_i^m 最大程度地将作者 a_i 与其他作者集 $\bar{a}_i$ 在样本空间上的样本点达到距离最大化，其中 $\bar{a}_i\cap a_i=\varnothing$，$\bar{a}_i\cup a_i=A(1\leqslant i\leqslant K)$，$A$ 为所有作者的集合。在对每个作者选择个体特征集时，将整个训练集均看作 a_i 和 $\bar{a}_i$ (除 a_i 外的所有作者)两类作者的训练集。因此关于 a_i 与 $\bar{a}_i$ 的类内离散度矩阵之和如式(4)所示。

$$S_w=\sum_{j=1}^{2}P_j\Sigma_j \quad (4)$$

该式表示作者训练样本空间内的样本点聚合程度，其中 P_i 为作者 a_i 的样本空间的先验概率值，Σ_1 为作者 a_i 样本空间的协方差矩阵，而 Σ_2 为 $\bar{a}_i$ 样本空间的协方差矩阵。

此外，关于 a_i 与 $\bar{a}_i$ 的类间离散度矩阵如式(5)所示。该式表示不同作者之间样本空间的分散程度。

$$S_b=\sum_{i=1}^{2}P_i(\mu_i-\mu_0)(\mu_i-\mu_0)^{\mathrm{T}} \quad (5)$$

式中，μ_0 表示全体作者集的训练样本均值，如式(6)所示。

$$\mu_0=\sum_{i}^{2}P_i\mu_i \quad (6)$$

将 S_w 与 S_b 相加得到混合离散矩阵 S_m，如式(7)所示。

$$S_m=S_w+S_b=E[(t-\mu_0)(t-\mu_0)^{\mathrm{T}}] \quad (7)$$

式中，混合离散矩阵 S_m 表示所有作者的训练样本空间内样本点的离散程度。显然，属于 a_i 的个体特征集必须使作者 a_i 与作者集 $\bar{a}_i$ 的训练样本间距离 S_w 最大化，而训练集内部样本距离 S_b 最小化，因此构造准则 $J(a_i,\bar{a}_i)$以计算每个作者的个体特征集，如式(8)所示。

$$J(a_i,\bar{a}_i)=tr(S_w^{-1}\cdot S_m)\propto\frac{(\mu_{a_i}-\mu_{\bar{a}_i})^2}{\sigma_{a_i}^2+\sigma_{\bar{a}_i}^2} \quad (8)$$

由于对矩阵 $S_w^{-1}\cdot S_m$ 求秩的计算复杂度较高，并且类间离散矩阵 S_b 通常为奇异矩阵，因此直接求解 $tr(S_w^{-1}\cdot S_m)$是不可行的。但不难发现，两类作者训练样本空间的方差的平方和 $\sigma_{a_i}^2+\sigma_{\bar{a}_i}^2$ 越小，则说明类内距离$|S_w|$越小；两作者样本集的均值之差$|\mu_{a_i}-\mu_{\bar{a}_i}|$越大，则类间距离$|S_b|$越大。因此可得到式(8)的转换关系。对每个作者 a_i，选择使该准则函数值较大的 m 个特征作为其个体特征集 IFS_i^m。

4.2 算法设计方案

根据上述方法，在选择出作者 $a_i(i=1,2,\cdots,K)$的个体特征集 IFS_i^m 后，可得到 IFS_i^m 中每个特征对于作者的风格隶属度。通过对隶属度大小排序，定义前 k 个具有最大隶属度的特征集贡献率 $d_{a_i}^k(1<k<d,a_i\in A)$，如式(9)所示。

$$d_{a_i}^k=\frac{\sum_{u=1}^{k}J_u(a_i,\bar{a}_i)}{\sum_{v=1}^{d}J_v(a_i,\bar{a}_i)} \quad (9)$$

根据式(9)，随机子空间中所包含的特征集贡献度越大，则其构造的基分类器稳定性越高，但基分类器的性能不可能完全稳定。为了构造具有差异度的基分类器，仍需要利用隶属度较小的特征集。因此结合标准的随机子空间算法框架，在子空间选择时设定一个阈值 η，分为两部分随机选取特征子集。SRFS 具体流程如下：

输入：训练集 T，特征集 F_d，匿名样本 μ，作者集 $A=\{a_1,a_2,\cdots,a_K\}$，个体特征集维数 m，贡献率阈值 $\eta\in[0,1]$；

输出：匿名样本 μ 的识别结果。

(1) 训练阶段

Step 1 将训练样本中每一维特征向量 f_i 归一化到[0,1]区间；

Step 2 对作者 a_i 选择粒度为 m 的个体特征集 $IFS_i^m(i=1,2,\cdots,K)$；

Step 3 按特征贡献率对 IFS_i^m 中的特征降序排列；

Step 4 对 IFS_i^m 中贡献率小于 η 的特征，选取随机特征序号 v_1^i，对贡献率大于 η 的特征，选取随机特征序号 v_2^i；

Step 5 选取特征子空间 $S_{ij}(v_1^i\cup v_2^i)\subset IFS_i^m$，并在该子空间上构造对应的基分类器 C_{ij}；

Step 6 直到 $K\times L$ 个基分类器构造结束。

(2) 识别阶段

Step 1 将匿名样本 μ 的特征向量归一化到[0,1]区间；

Step 2 将 μ 的特征向量划分到对应于训练阶段选取的 $K\times L$ 个特征子集中；

Step 3 应用构造的 $K\times L$ 个基分类器分别对 μ 进行分类；

Step 4 统计基分类器的识别结果，并采用多数投票法的融合策略得到 μ 的身份归属，如下式所示：

$$\mathrm{author}(\mu)=\arg\max\{\mathrm{votes}(a_1,a_2,\cdots,a_K)\} \quad (10)$$

与传统的随机子空间方法不同，首先，SRFS 并不是在原始特征空间内完全随机划分，而是在选取随机子空间之前引入个体特征选择，充分利用了不同作者的书写纹信息，提高局部子空间选取的差异度。其次，子空间选取过程中，根据个体特征集的贡献率 $d_{a_i}^k$，设置阈值 η，将子空间的随机抽取分为$[0,\eta]$与$[\eta,1]$两个区间进行。显然，$[0,\eta]$区间内随机选取的

特征鉴别度更高，能保证基分类器的分类精度和稳定性，而$[\eta,1]$区间内随机选取的特征虽对作者个体的鉴别度不高，但具有较强的扰动性，为基分类器的构造提供了差异性的保证。

5 实验研究

5.1 数据集

本文所使用的实验数据采集于华中师范大学校园BBS——博雅论坛。该论坛采取用户实名制，发帖量较大，数据真实可靠，适合于做中文书写纹识别研究。数据集的相关信息如表1所列。

表1 实验数据集信息

作者数量	每个作者帖子数	帖子平均长度	时间跨度
40	30	74.5字符	1年

数据集的预处理包括两部分：1)去除文本中与内容无关的字符，包括html网页代码以及其他灌水字符等；2)在整个训练样本集中统计所有N-gram字符(为了减少特征空间的稀疏度，取$0<N<5$)的出现频率，并取频率位于前6000的N-gram字符作为初始特征集。实验中，训练集与测试集按照1:1的比例分割，分别包含600个样本。

5.2 实验设计

为了验证SRFS的识别性能以及在书写纹识别中的鲁棒性，实验设计方法如下：

(1)考察个体特征集的选取粒度m对SRFS算法识别性能的影响。粒度值m分别取值为1000，2000，3000，4000和5000，测试的作者数量分别为5，10，20，30和40，对不同数量的作者集识别时，从每组个体特征集上分别构造5个基分类器。

(2)对比SRFS与RS、Bagging算法以及单分类器LSVM的识别性能，考察4种方法对40个作者集的识别结果。其中LSVM基于原始特征空间进行分类，RS与SRFS采取对特征空间进行划分，而Bagging算法对样本空间进行随机划分[15]。由于LSVM是在之前多个关于书写纹识别的实验研究中取得最好识别性能的单分类器，因此所有的集成方法均采用LSVM作为底层分类器。其中，3种集成分类算法分别使用不同的基分类器个数做实验($L=120,200,400$)。SRFS中个体特征集粒度为原始特征集的50%，即3000维，随机子空间大小为个体特征集的50%。

(3)在以上两种实验中，为了保证基分类器的识别精度以及差异度，SRFS中的贡献率阈值η均取0.8。采用识别正确率Accuracy以及Kappa统计值作为识别结果的评价指标：

$$Accuracy=\frac{\text{测试集中正确分类的样本数}}{\text{测试集中样本总数}} \tag{11}$$

$$Kappa=\frac{P_o-P_e}{1-P_e} \tag{12}$$

式中，P_o为分类混淆矩阵主对角线上观测值的总和，P_e为分类混淆矩阵主对角线上期望值的总和。Kappa系数的含义是度量分类输出值与真实值的一致性。Kappa值越大，说明一致性越好，Kappa值越小，则一致性越差。同时使用差异性度量式(1)计算3种集成分类器的差异度并作对比分析。由于随机空间的划分具有不确定性，因此3种集成分类算法均采取5次独立运行的平均值，所有实验都在MATLAB7.1平台上完成。

5.3 实验结果与分析

实验结果如图1与表2所示。图1表示在不同个体特征集选择粒度m下，采用SRFS算法对不同规模作者集识别的结果图。表2描述4种不同识别方法对40个作者集识别的比较情况，并利用差异度衡量各集成分类算法的性能。

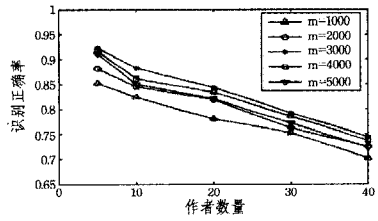

图1 识别正确率与个体特征集选择粒度的关系

表2 不同算法在40个作者集上的识别结果比较

算法	基分类器个数	识别正确率(%)	Kappa值(%)	差异度(%)
SRFS	120	74.6	72.8	36.8
	200	75.2	73.6	38.5
	400	77.8	75.6	39.1
RS	120	72.1	70.2	26.4
	200	73.4	71.6	28.2
	400	74.2	71.9	29.1
Bagging	120	73.1	71.6	25.6
	200	74.6	72.2	27.8
	400	75.2	72.6	28.2
LSVM	N/A	72.6	70.4	N/A

注：SRFS表示半随机特征采样算法，RS表示标准的随机子空间算法，Bagging算法是采取在样本空间随机划分的算法，LSVM表示带有线性核函数的支持向量机。

个体特征选择是SRFS算法的关键环节。由图1可知，个体特征选择的粒度m越小，子空间内关于作者的个体关键特征覆盖率越低；相反，m越大，随机子空间的选择范围将随之扩大，鉴别信息在不同子空间内分布越均匀，但m的取值过大也会造成算法性能的下降，不同子空间内的特征子集所包含的共现特征将增多。从集成学习中差异度的角度看，m越大，基分类器之间差异度越低，极端情况下会造成集成分类器等同于一个单分类器的性能。可以观察到，当$m=1000$时，因为选取粒度太小，识别准确率最低。当$m=3000$(原始特征集的50%)时，识别性能较好，识别40个作者时，精度超过了74%，在识别5到30个作者集时，识别准确度均超过了80%。而当m的取值提高到5000时，选取粒度接近于原始空间的维度，准确率不再提升，反而有所下降。

在表2中，从识别正确率和Kappa值统计结果可观察到，与基于特征全集的单分类器LSVM相比，3种集成分类算法在书写纹的识别上表现出更高的性能。采用完全随机式选择的RS性能略高于LSVM，随着基分类器个数的增加，识别正确率也有所提升，表明RS能在一定程度上提高集成分类的效果，但这种完全随机的选择机制并不能保证子空间内特征子集的鉴别度，会造成随机选取的子空间内出现大量冗余特征，降低基分类器识别的稳定性，进而制约了其识别性能的提升。而采用Bagging算法从样本空间进行随机划分时，取得了比RS更好的识别性能，说明不同作者的样本空间内也能提供一定粒度的个体鉴别信息，但由于该算法构造的基分类器均在特征全空间上进行分类，对于冗余特征的处理无能为力。从差异度的统计结果来看，Bagging所构造的基分类

(下转第152页)

开放。

结束语 网络安全风险量化评估对网络系统的安全保障和主动防护具有重要的现实意义。针对现有风险评估技术存在的自主性不足等问题，本文提出了一种基于攻击图的多 Agent 风险评估模型——MREMBAG。通过在风险评估过程中引入多 Agent 技术并采用全局攻击图生成算法自动生成网络攻击图，依据该攻击图计算攻击路径、组件、主机、网络的风险指数和漏洞及其关联关系风险量化指标，通过计算和分析获取目标网络的安全风险指标。实验结果表明，MREMBAG 模型为解决网络安全风险的量化评估问题提供了一个可行、有效的方法。

在未来的研究中，将以 MREMBAG 模型为基础并综合考虑已有安全措施及管理因素对网络风险的影响，通过利用网络数据对评估模型和评估方法进行改进，从而进一步完善评估效果。

参 考 文 献

[1] 江常青. 信息安全评估需要研究和解决的几个关键问题[J]. 国家信息安全测评认证，2007(5)：1-4

[2] Phillips C, Laura S P. A graph-based system for network vulnerability analysis[C]// Proceedings of the 1998 workshop on New security paradigms. VA, USA: ACM Press, 1998: 71-79

[3] Ammann P, Pamula J, Ritchey R, et al. A host based approach to network attack chaining analysis[C]// Proceedings of the 21st Annual Computer Security Applications Conference. Tucson, Arizona, USA: IEEE Computer Society Press, 2005: 72-84

[4] Ramakrishnan C, Sekar R. Model-based vulnerability analysis of computer systems[C]// Proceedings of the 2nd International Workshop on Verification. Pisa, Italy: Model Checking and Abstract Interpretation Press, 1998: 1-8

[5] Ritchey R, Ammann P. Using model checking to analyze network vulnerabilities[C]// Proceedings of the 2000 IEEE Symposium on Security and Privacy. Berkeley, California, USA: IEEE Computer Society Press, 2001: 156-165

[6] Sheyner O. Scenario Graphs and Attack Graphs [D]. School of Computer Science, Carnegie Mellon University, Pittsburgh, USA, 2004

[7] Sheyner O, Haines J, Jha S, et al. Automated Generation and Analysis of Attack Graphs[C]// Proceedings of the 2002 IEEE Symposium on Security and Privacy. Oakland, California, USA: IEEE Computer Society Press, 2002: 254-265

[8] Jha S, Sheyner O, Wing J. Two Formal Analyses of Attack Graphs[C]// Proceedings of the 15th Computer Security Foundations Workshop. Beijing, China: Chinese Academy of Sciences Press, 2002: 49-63

[9] 李冠君. 基于安全代理的网络自保护系统模型研究[D]. 天津：中国民航大学，2009

[10] Roschke S, Cheng F, et al. Towards Unifying Vulnerability Information for Attack Graph Construction[J]. Computer Science, Information Security, 2009(5735): 218-233

[11] 陈天平，许世军，等. 基于攻击检测的网络安全风险评估方法[J]. 计算机学报，2010，37(9)：94-96

[12] 张永铮，方滨兴，迟悦，等. 网络风险评估中网络节点关联性的研究[J]. 计算机学报，2007，30(2)：234-240

(上接第 123 页)

器差异度仍然较低。基分类器个数 $L=400$ 时，差异度仅为 28.2%。最后采用 SRFS 算法，大大提高了基分类器之间的差异度，$L=200$(对每个作者构造 5 个基分类器)时，差异度为 36.8%，当 L 提高到 400(对每个作者构造 10 个基分类器)时，差异度达到了 39.1%，比其他方法最高高出了约 10%，Kappa 值最高能达到 75.6 的最高点，说明识别结果的可信度较高。但值得注意的是，基分类器的数量太大也会降低集成系统的运行效率，虽然增加基分类器个数能一定程度提高集成分类性能，但超过某个阈值会使性能降低，并减小基分类器之间的差异度。

以上实验结果说明了，通过从特征空间中挖掘作者个体书写纹特征并结合随机子空间的划分方法能显著地提高书写纹识别的正确率和鲁棒性，另一方面也体现了基分类器之间的差异度对集成分类器性能具有重要的影响。

结束语 书写纹识别对于匿名网络主体行为的规范化和网络安全的维护具有重要的实际意义，但在中文语境下，目前仍存在较多问题需要进一步解决。针对 N-gram 字符特征集的高稀疏度和噪音数据较多等问题，提出一种基于作者个体特征选择和随机子空间的集成分类方法，其充分利用特征空间中的个体鉴别信息来提高基分类器之间的差异度，降低集成分类系统对噪音数据的敏感度。实验结果表明，该算法能有效提高书写纹的识别性能。下一步将就作者个体特征选择对识别性能的理论依据展开进一步的研究，并结合对样本空间的划分机制设计更加优化的中文书写纹识别算法。

参 考 文 献

[1] 索绪尔 F，等. 普通语言学教程[M]. 高名凯，译. 北京：商务印书局，1980

[2] Li J, Zheng R, ChenH. From fingerprint to writeprint[J]. Communications of the ACM, 2006, 49(4): 76-82

[3] Zheng R, Li J, Chen H, et al. A framework for authorship identification of online messages: writing style features and classification techniques[J]. Journal of the American Society of Information Science and Technology, 2006, 57(3): 378-393

[4] Zhao Y, Zobel Y. Effective and scalable authorship attribution using function words[C]// Proceedings of the 2nd Asian Information Retrieval Symposium. Berlin: Springer, 2005: 174-189

[5] 武晓春，黄萱菁，吴立德. 基于语义分析的作者身份识别方法研究[J]. 中文信息学报，2006，20(6)：61-68

[6] Stamatatos E. A survey of modern authorship attribution methods[J]. Journal of the American Society of Information Science and Technology, 2009, 60(3): 538-556

[7] Houvardas J, Stamatatos E. N-gram feature selection for authorship identification[C]// Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, Applications. Berlin: Springer, 2006: 77-86

[8] Stamatatos E. Authorship attribution based on feature set subspacing ensembles[J]. International Journal on Artificial Intelligence Tools, 2006, 15(5): 823-838

[9] Koppel M, Schler J, Argamon S. Authorship attribution in the wild[J]. Language Resources and Evaluation, 2010, 45(1): 83-94

[10] 孙建文，杨宗凯，刘三妍，等. 基于集成学习与遗传算法的网络书写纹识别研究[J]. 计算机科学，2011，38(6)：242-245

[11] Ho T. The random subspace method for constructing decision forests[J]. IEEE Trans. on PAMI, 1998, 20(8): 832-844

[12] 韩宏，杨静宇. 多分类器组合及其应用[J]. 计算机科学，2000，27(1)：58-61

[13] Kuncheva L, Wgitaker C. Measures of diversity in classifier ensembles[J]. Machine Learning, 2003, 51(2): 181-207

[14] 顾亚祥，丁世飞. 支持向量机研究进展[J]. 计算机科学，2011，38(2)：14-17

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