基于集成学习与遗传算法的网络书写纹识别研究

孙建文[1] 杨宗凯[1] 刘三妍[1] 王 佩[2]
(华中师范大学国家数字化学习工程技术研究中心 武汉 430079)[1]
(武汉大学信息管理学院 武汉 430072)[2]

摘 要 N-gram字符是网络书写纹识别最有效的特征类型之一。针对其特征维数高、冗余特征多且无关特征少等特点,提出一种基于特征空间划分来构造集成学习分类器的网络书写纹识别方法。该方法首先根据一定的划分粒度,将初始特征集划分为等维度、无交叉的特征子集,然后基于每一个特征子集训练生成对应的基分类器(多元朴素贝叶斯),最后采用算术与几何平均相结合的融合策略完成集成学习分类器的构造。特征空间的划分(即特征子集的选择)采用遗传算法进行优化。实验在一个真实数据集上开展,其结果表明该方法有效地提高了网络书写纹的识别性能。

关键词 网络书写纹,集成学习,遗传算法,特征子集
中图法分类号 TP391 **文献标识码** A

Research of Online Writeprint Identification Based on Ensemble Learning and Genetic Algorithm

SUN Jian-wen[1] YANG Zong-kai[1] LIU San-ya[1] WANG Pei[2]
(National Engineering Research Center for E-learning, Huazhong Normal University, Wuhan 430079, China)[1]
(School of Information Management, Wuhan University, Wuhan 430072, China)[2]

Abstract Online writeprint identification is a technique to identify individuals based on textual identity cues people leave behind online messages. Character N-gram is one of the most effective approaches to identify writeprint according to previous research. To deal with the high dimensional and redundant feature problems and the property of each feature being valuable for the task of writeprint identification, an ensemble learning approach based on feature subspacing was proposed in this study. The essence of this method is to partition the features into distinct subsets. Firstly, the whole feature set is split into equally sized and disjoint subsets. Then each of them is used to train a base classifier using Multinomial Naive Bayes. Finally, these individual classifiers are aggregated to construct the ensemble via an appropriate combination rule which is a simple average of arithmetic mean and geometric mean. Additionally, genetic algorithm was used to optimize the feature subspacing (i. e. feature subsets selection). To examine the approach, experiment was conducted on a real world test bed. Performance results showed the proposed approach was quite effective and obtained a considerable improvement in accuracy compared with the benchmark technique in writeprint identification (Support Vector Machine).

Keywords Writeprint identification, Ensemble learning, Genetic algorithm, Feature subset

1 引言

互联网的迅速发展与广泛应用,给人们带来诸多好处,同时也为各种网络信誉或犯罪问题(如远程教育中在线作业或论文的抄袭与剽窃,电子商务中产品介绍与客户评论等信息的弄虚作假,以及垃圾邮件、色情信息、反动言论、威胁恐怖信息等在网络上的泛滥)的产生提供了新的空间和手段。这些现象的产生很大程度上是由于互联网的开放性与可匿名性。用户可采用匿名或冒用他人名义的方式发布信息,以躲避追踪,难以被取证,这样容易从心理上造成人们对网络空间责任感的缺失。因此,如何有效鉴别与锁定信息的真实发布者,提高人们在互联网上的责任感,具有重要的研究价值与实际意义。

相关研究[1]认为,人们在自己书写的文字作品中会无意识地留下自己独特的风格(如遣词用字、语调、文法习惯等),并由此诞生了风格学或文体学理论。网络书写纹即是用户在网络留言中所留下的风格,是能够识别其身份的特征集合,是用户在互联网上留下的"网络指纹"。因此,通过研究与分析网络书写纹来鉴别与锁定不良信息的真实作者,可以从心理上提高人们在网络空间的责任感,并且可从技术上对取证提

到稿日期:2010-07-29 返修日期:2010-12-13 本文受国家863计划项目(2008AA01Z131),华中师范大学中央高校基本科研业务费项目(CCNU09A02006)资助。

孙建文(1982-),男,博士生,主要研究方向为智能软件与知识服务、机器学习,E-mail:sunjw.work@gmail.com;杨宗凯(1963-),男,教授,博士生导师,主要研究方向为现代信息网络、数字信号处理;刘三妍(1973-),男,教授,主要研究方向为人工智能、计算机应用;王 佩(1984-),女,博士生,主要研究方向为信息资源管理。

供支持，对于上述问题的解决也提供了一条新的思路。

从机器学习角度，网络书写纹识别是一个多类别单标签的文本分类问题，核心问题是特征选取与分类模型构建。根据已有研究[2,3]，通常用于构建网络书写纹识别分类模型的特征主要包括词汇、语法、结构及语义等几类。其中，N-gram字符特征是最有效的特征类型之一[4]。从语言学角度，N-gram字符作为一种底层特征，将其用于网络书写纹识别的优点包括与语种无关、预处理简单、含上层特征信息(如语法、结构与语义)等。这对于中文环境尤其适合，可避免在如分词、词性标注等环节因自然语言处理技术的不成熟所带来的误差。因此，本文拟采用N-gram字符类型特征来构建用于网络书写纹识别的分类模型。

但另一方面，采用N-gram字符特征通常会出现特征维度高、冗余特征多且无关特征少等情况。传统的单分类器较难处理，构造的模型过于复杂且难以获得令人满意的识别性能。集成学习被称为机器学习四大方向之一，在多种分类任务中都显著提高了学习系统的泛化能力[5]。划分样本空间和特征空间是构造集成学习的两种主要方法。根据上述N-gram字符特征的特点，本文采用划分特征空间的方法，基于特征子集对特征空间进行划分，并采用遗传算法进行优化，以提高网络书写纹的识别性能。

2 基于特征子集的集成学习方案

2.1 问题描述

与词袋表示法类似，用于网络书写纹识别的训练或测试文本可表示为N-gram字符的向量。设 $G_n=\{g_1,g_2,g_3,\cdots,g_n\}$ 为训练集出现频率最高的 n 个N-gram字符的降序集合，f_{ij} 为第 j 个N-gram字符(g_j)在第 i 个文本中的频率，则任一文本 x_i 可表示为 $\{f_{i1},f_{i2},f_{i3},\cdots,f_{in}\}$。

设 $G_{m,n}$ 为 G_n 的子集，包含 m 个N-gram字符(其中 $m<n$)。另设 $C_i(G_{m,n})$ 为在特征子集 $G_{m,n}$ 上训练生成的基分类器 C_i，则基于 k 个特征子集(即 k 个基分类器)的集成学习分类器EnsFS (Ensemble based on Feature Subsets)可表示为：

$$EnsFS=E(\sum_{i=1}^{k}C_i(G_{m,n}),comb) \tag{1}$$

式中，$comb$ 表示基分类器的融合方法。

2.2 具体方案

根据以上描述，对于一个具体的基于特征子集的集成学习方案，需要进一步确定的问题主要包括：特征子集数量 k(即基分类器数量)的确定、特征子集的选择(包括每个特征子集的大小、能否包含相同的特征等)、基分类器的确定以及基分类器的融合策略。

对于以上问题，本文拟采用如下方案：

(1)将特征子集数量 k 本身作为一个参数进行考察。

(2)对于特征子集的选择，为使问题简单，本文每个特征子集包含相同数目的特征，并且所有特征子集之间无交集，则每个特征子集包含特征的个数为 $floor(n/k)$。特征子集的选择采用随机选择与遗传算法两种方法。

(3)基分类器采用多项朴素贝叶斯(Multinomial Naive Bayes，MNB)，因其广泛应用于文本分类，在分类性能与处理速度上具有较好的平衡。此外，MNB分类结果中包含对所有类别的后验概率，可为集成学习的融合提供全信息。

(4)采用简单算术平均与几何平均相结合的方式作为基分类器的融合策略，来弥补算术平均易受较大后验概率值影响以及几何平均易受多个较小后验概率值影响的缺陷。

$$P(ens,x,c)=\frac{\frac{1}{k}\sum_{i=1}^{k}P_i(C_i(G_{m,n}),x,c)+k\sqrt{\prod_{i=1}^{k}P_i(C_i(G_{m,n}),x,c)}}{2} \tag{2}$$

式中，$P_i(C_i(G_{m,n}),x,c)$ 表示给定输入文本 x，采用基分类器 $C_i(G_{m,n})$ 所得到的属于类 c 的后验概率估计。$P(ens,x,c)$ 表示按以上规则融合全部 k 个基分类器的结果，即采用集成学习分类器 ens，所得到的关于输入样本 x 属于类 c 的概率估计。

3 基于遗传算法的特征子集选择

与传统的基于遗传算法的特征子集选择方法相比[6,7]，本方法在编码方案、遗传算子以及适应度函数设计等方面有所不同，下面将分别从这几个方面进行阐述，最后给出整个算法的流程。

3.1 编码方案

传统的基于遗传算法的特征子集选择方法一般采用二进制编码，每条染色体表示一个特征子集。而在本文，每条染色体表示多个特征子集，其数量与所采用的基分类器个数相等。因此，本方案采用实数编码，染色体长度等于特征的维数，基因位的取值使用特征编号。染色体被均分为与基分类器个数相同的基因片段，基因片段的长度等于特征子集大小，并且每个基因片段的基因位不包含相同的取值，即一条染色体所有基因位的取值不重复，并能完整表达所有特征。该编码方案可较好地满足以上所提出的采用等维度、无交叉的特征子集构造集成学习分类器的方案。

3.2 遗传算子

选择操作采用轮盘赌方法与精英保留策略，以保证最优个体始终存活于下一代种群。对于交叉和变异操作，由于采用基于特征编号的实数编码，且个体所有基因位取值不重复，标准遗传算法中的交叉和变异算子难以直接应用，易产生无意义的个体。为此，本文采用Grefenstette[8]编码策略。首先对个体进行Grefenstette编码，然后执行交叉与变异操作(分别采用单点交叉与简单随机变异方法)，最后对个体进行Grefenstette译码，还原基于特征编号的编码，以计算适应度值。

3.3 适应度函数

与大多数基于遗传算法的Wrapper型特征子集选择方法类似，个体适应度采用构建于个体所表示的多特征子集的集成学习分类模型进行评价。具体来说，首先基于个体所表示的多特征子集对原特征空间进行降维，形成新的样本集，并在其上构建多分类器模型；然后输入测试集，通过式(2)计算其对于所有类别的后验概率并对测试结果进行评价，最后以分类正确率作为其目标函数值，并通过线性尺度变换得到其适应度值。

3.4 算法流程

根据以上讨论，基于标准遗传算法框架，该算法具体流程如下：

输入：数据集 D，特征集 F，种群大小 $Size$ 与最大进化代数 $MaxGen$。

输出：优化的特征空间划分方案。

步骤：

(1)根据 3.1 节的编码方案，随机生成大小为 *Size* 的初值种群；

(2)根据 3.3 节的适应度函数，计算当前种群的个体适应度；

(3)判断是否达到最大进化代数 *MaxGen*，若是则输出当前种群的最优个体，否则执行以下步骤；

(4)根据个体适应度执行选择操作；

(5)对所选个体执行 Grefenstette 编码；

(6)执行交叉操作；

(7)执行变异操作；

(8)执行 Grefenstette 译码；

(9)返回步骤(2)。

4 实验研究

4.1 实验数据集及预处理

本文研究所使用的实验数据采集自华中师范大学校园 BBS——博雅论坛。该论坛采用实名制，发帖量大，数据充分、可靠，适于做网络书写纹识别研究。实验数据集的相关信息如表 1 所列。

表 1 实验数据集的相关信息

作者数量	每个作者帖子数	帖子平均长度	时间跨度
20	50	73.4 字符	1 年

数据集的预处理主要包括两方面：首先在整个数据集上统计所有 N-gram 字符(为减少冗余，N 值小于 5)的频率，并取频率位于前 1000 的 N-gram 字符作为初始特征集；其次按 4∶1 将数据集划分为训练集与测试集，分别含 800 和 200 个样本。

4.2 实验设计

为验证基于特征子集的集成学习的分类性能以及在网络书写纹识别应用中的有效性，实验设计如下：

首先，考察特征空间的划分粒度对集成学习分类性能的影响。划分粒度值 m 分别取值为 5，10，50，100 和 200。

其次，分别采用随机选择与遗传算法两种方法对特征空间进行划分，并通过实验对比两种方法对于集成学习分类性能的影响。

最后，与两种单分类器：多元朴素贝叶斯(MNB)与支持向量机(SVM)进行比较。其中，MNB 是作为本文所提集成学习方案的基分类器，SVM 则是在之前多个关于网络书写纹识别的实验研究中都取得最好识别性能的单分类器，可作为网络书写纹识别实验对比的基准分类器。

在所有实验中，均采用识别正确率作为比较指标，即测试集中被正确分类的样本占测试集样本总量的比例。随机特征子集选择方法实验执行 10 次，结果取平均值。

本文所有实验都在 WEKA[9] 平台上完成。SVM 采用适合于文本处理的 LIBLINEAR[10]，C 值设为 1。遗传算法初始种群大小为 32，最大进化代数设为 500，交叉与变异概率分别取值为 0.6 与 0.03。

4.3 实验结果与分析

部分实验结果如图 1 与表 2 所示。图 1 表示在不同的划分粒度 m 下，采用遗传算法对特征子集进行优化选择的运行状况图。表 2 是 4 种不同识别方法的识别正确率，包括两种单分类器方法与两种集成学习识别方法。其中，两种集成学习方法均使用了 5 种划分粒度(m=5，10，50，100 和 200)。

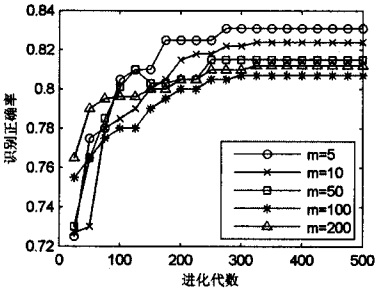

图 1 基于遗传算法的特征子集选择运行图

表 2 不同方法的识别正确率比较

识别方法	MNB	SVM	RAFSENS					GAFSENS				
			m=5	10	50	100	200	m=5	10	50	100	200
识别正确率	74.2	77.5	71.8	70.6	71.6	71.9	72.2	83.1	82.4	81.5	80.7	81.2

注：MNB 表示多元朴素贝叶斯分类器，SVM 表示支持向量机分类器，RAFSENS 表示采用随机方法产生特征子集的集成学习识别方法，GAFSENS 表示采用遗传算法进行特征子集选择的集成学习识别方法，m 表示特征空间的划分粒度。

由图 1 可知，在 5 种不同的划分粒度下，算法均在第 250 代左右开始收敛，体现出遗传算法的良好设计以及对特征空间进行划分的有效性。此外，划分粒度对集成学习识别性能的影响呈现出一定的规律性，m 越小其识别正确率越高。并且当 m 较小时，其初期的识别正确率较低，但随着进化代数的增加，识别正确率逐渐提高，并超过 m 取较大值时的识别正确率。这个现象可在一定程度上用集成学习的多样性理论来解释。在遗传算法的初期，由于基分类器之间的多样性不够明显，集成学习的性能主要由基分类器的性能决定。当 m 较小时，基于 m 个特征训练生成的基分类器的识别正确率较低，故集成学习的整体性能偏低。但随着遗传算法对特征空间划分的不断优化，基分类器之间的多样性不断增加，其对集成学习整体性能的影响逐渐增强。从定性的角度看 m 越小其多样性越大，故集成学习的整体性能越高。

与构建于特征全集的基分类器(MNB)相比，两种基于特征子集的集成学习方法在网络书写纹的识别性能上表现出很大的区别。当采用随机方法来划分特征空间时，其性能低于 MNB。但采用遗传算法来优化特征子集的选择时，却大大提高了识别性，其识别正确率比 MNB 最高高出了约 9%。这个结果说明了通过划分特征空间来构建集成学习方法的可行性及其在网络书写纹识别应用中的有效性，另一方面也体现了特征子集的选择对集成学习性能具有重要的影响。

而与网络书写纹识别的基准分类器——SVM 相比，采用随机方法产生特征子集的集成学习的识别性能远远低于 SVM，但通过采用遗传算法进行优化的方法则高于 SVM。这一方面验证了作为单分类器的 SVM 在网络书写纹识别中的优越性能，另一方面也说明了在通过划分特征空间构建集成学习分类器时，采用遗传算法进行优化的有效性。

关于特征空间的划分粒度对集成学习识别性能的影响，由前所述，在采用遗传算法划分特征空间时，划分粒度对集成学习识别性能的影响呈现出较明显的规律性，划分粒度越小其识别性能越高。但如表 2 所列，当采用随机方法对特征空间进行划分时，划分粒度对集成学习识别性能的影响未呈现出明显的规律性。对于划分粒度对集成学习识别性能影响更深层次的理论解释，还是一个需要进一步深入研究的问题。

结束语 网络书写纹识别是一个具有研究价值与实际意义的方向，特别是针对汉语环境，目前还有许多问题需要进一步解决。针对 N-gram 字符特征具有高维、冗余等特点，提出一种基于特征空间划分的集成学习方法，并通过遗传算法对特征空间的划分进行优化。实验表明，该方法可有效提高网络书写纹识别的性能。下一步将就划分粒度对识别性能影响的理论依据以及特征空间划分的其他方法等问题展开进一步研究。另外，将充分利用集成学习与遗传算法的内在并行性，设计网络书写纹识别的并行算法，以提高其处理效率。

参 考 文 献

[1] Holmes D I. The analysis of literary style — A review[J]. Journal of the Royal Statistical Society, Series A, 1985, 148(4): 328-341

[2] Abbasi A, Chen H. Applying authorship analysis to extremist-group web forum messages[J]. IEEE Intelligent Systems, 2005, 20(5): 67-75

[3] 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

[4] 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

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

[6] Goldberg D E. Genetic Algorithms in Search, Optimization and Machine Learning[M]. Boston, MA: Addison-Wesley Longman Publishing Co., Inc., 1989

[7] 任江涛，孙婧昊，黄焕宇，等. 一种基于信息增益及遗传算法的特征选择算法[J]. 计算机科学，2006，33(10)：193-196

[8] Grefenstette J J, Gopal R, Rosmaita B, et al. Genetic Algorithms for the Traveling Salesman Problem[C]// Proc. 1st Int. Conf. Genetic Algorithms and Their Applications. Lawrence Erlbaum Ass., 1985: 160-168

[9] Fan R E, Chang K W, Hsieh C J, et al. LIBLINEAR: A library for large linear classification [J]. The Journal of Machine Learning Research, 2008, 9: 1871-1874

[10] Hall M, Frank E, Holmes G, et al. The WEKA data mining software: An update[J]. ACM SIGKDD Explorations Newsletter, 2009, 11(1): 10-18

(上接第 222 页)

[2] Stoilos G, Nikos S, Stamou G, et al. Uncertainty and the semantic Web[J]. IEEE Trans on Intelligent Systems, 2006, 21(5): 83-87

[3] Straccia U. Towards a fuzzy description logic for the semantic Web (preliminary report)[C]// Proceedings of 2nd European Semantic Web Conference, LNCS 3532. Heraklion, Greece, 2005: 167-181

[4] Quan T T, Hui S C, Fong A C M, et al. Automatic fuzzy ontology generation for semantic Web [J]. IEEE Transaction on Knowledge and Data Engineering, 2006, 18(6): 842-856

[5] Sanchez E, Yamanoi T. Fuzzy ontologies for the semantic Web [C]// Proceedings of 7th International Conference on Flexible Query Answering Systems, LNAI 4027. Milan, Italy, 2006: 691-699

[6] Lee C, Jian Z, Huang L. Fuzzy ontology and its application to news summarization[J]. IEEE Transactions on Systems, Man, and Cybernetics- Part B: Cybernetics, 2005, 35(5): 859-880

[7] Ma Zong-min, Lv Yan-hui, Yan Li. A fuzzy ontology generation framework from fuzzy relational databases [J]. International Journal on Semantic Web and Information Systems, 2008, 4(3): 1-15

[8] 李曼，王琰，赵益宇，等. 基于关系数据库的大规模本体的存储模式研究[J]. 华中科技大学学报：自然科学版，2005，33(sup.)：217-220

[9] Zhou Jian, Ma Li, Liu Qiao-ling, et al. Minerva: a scalable OWL ontology storage and inference system[C]// Proceedingis of 1st Asian Semantic Web Conference, LNCS 4185. Bejing, China, 2006: 429-443

[10] Anuradha G, Cindy X C, Kajal T C, et al. From ontology to relational databases[C]// Proceedings of 2004 International Workshop on Conceptual Model, LNCS 3289. 2004: 278-289

[11] Vysniauskas E, Nemuraite L. Transforming ontology representation from OWL to relational database[J]. Information Technology and Control, 2006, 35(3): 333-343

[12] 许卓明，黄永菁. 从 OWL 本体到关系数据库模式的转换[J]. 河海大学学报，2006，34(1)：95-99

[13] 朱姬凤，马宗民，吕艳辉. OWL 本体到关系数据库模式的映射[J]. 计算机科学，2008，35(8)：165-169，205

[14] Barranco C D, Campana J R, Medina J M, et al. On storing ontologies including fuzzy datatypes in relational databases [C]// Proceedings of 16th IEEE International Conference on Fuzzy Systems. London, UK, 2007: 1-6

[15] Lv Yan-hui, Ma Zong-min, Zhang Xu-hui. Fuzzy ontology storage in fuzzy relational database [C] //Proceedings of 6th International Conference on Fuzzy Systems and Knowledge Discovery. Tianjin, China, 2009, 2: 242-246

[16] Hitzler P, Krotzsch M, Parsia B, et al. OWL 2 web ontology language primer [M/OL]. http: // www. w3. org/TR/2009/REC-owl2-primer-20091027/, 2009-10-27

[17] Klyne G, Carroll J J. Resource description framework (RDF): concepts and abstract syntax [M/OL]. http: // www. w3. org/TR/2004/REC-rdf-concepts-20040210/, 2004-02-10

[18] Hayes P. RDF semantics [M/OL]. http: // www. w3. org/TR/2004/REC-rdf-mt-20040210/, 2004-02-10

[19] Manola F. RDF primer [M/OL]. http: // www. w3. org/TR/2004/REC-rdf-primer-20040210/, 2004-02-10

[20] Hull R. Relative information capacity of simple relational database schemata[J]. SIAM Journal of Computing, 1986, 15(3): 856-886

[21] Miller R J, Ioannidis Y E, Ramakrishnan R. The use of information capacity in schema integration and translation [C]// Proceedings of 19th International Conference on Very Large Data Bases. Dublin, Ireland, 1993: 120-133

[22] Miller R J, Ioannidis Y E, Ramakrishnan R. Schema equivalence in heterogeneous systems: bridging theory and practice [J]. Information Systems: 1994, 19(1), 3-31

[23] Qian Xiao-lei. Correct schema transformations [C]// Proceedings of 5th International Conference on Extending Database Technology (EDBT), LNCS 1057. Avignon, France, 1996: 114-128