基于集成学习理论的文本情感分类①

方 丁[1]，王 刚[2,3]

[1](上海机场集团有限公司，上海 200335)
[2](合肥工业大学 管理学院，合肥 230009)
[3](过程优化与智能决策教育部重点实验室，合肥 230009)

摘 要：随着 Web2.0 的迅速发展，越来越多的用户乐于在互联网上分享自己的观点或体验。这类评论信息迅速膨胀，仅靠人工的方法难以应对网上海量信息的收集和处理，因此基于计算机的文本情感分类技术应运而生，并且研究的重点之一就是提高分类的精度。由于集成学习理论是提高分类精度的一种有效途径，并且已在许多领域显示出其优于单个分类器的良好性能，为此，提出基于集成学习理论的文本情感分类方法。实验结果显示三种常用的集成学习方法 Bagging、Boosting 和 Random Subspace 对基础分类器的分类精度都有提高，并且在不同的基础分类器条件下，Random Subspace 方法较 Bagging 和 Boosting 方法在统计意义上更优，以上结果进一步验证了集成学习理论在文本情感分类中应用的有效性。

关键词：文本情感分类；集成学习；Bagging；Boosting；Random Subspace

Text Sentiment Classification Based on Ensemble Learning

FANG Ding[1], WANG Gang[2,3]

[1](Shanghai Airport Authority Group Limited Company, Shanghai 200335, China)
[2](School of Management, Heifei University of Technology, Hefei 230009, China)
[3](The Ministry of Education Key Laboratory of Process Optimization and Intelligent Decision, Hefei 230009, China)

Abstract：With the development of Web 2.0, more and more users are happy to share their opinions and experiences on the internet. Subsequently, it is increasingly difficult for people to collect and process the huge information from the network. Therefore, text sentiment classification based on the computer is proposed to tackle this problem. And one of the most important research directions is to enhance the classification accuracy for text sentiment classification. In addition, ensemble learning is an effective approach to enhance the classification accuracy and has shown better performance than base classifiers in many fields. Based on these considerations, text sentiment classification based on ensemble learning is proposed to enhance the performance of classifiers. Experimental results reveal that three ensemble methods, i.e., Bagging, Boosting and Random Subspace, enhance the classification accuracy of different base classifiers. Compared with Bagging and Boosting, Random Subspace gets more significant improvement of the classification accuracy. All these results demonstrate the effectiveness and feasibility of application of ensemble learning in text sentiment classification.

Key words：text sentiment classification; ensemble learning; bagging; boosting; random subspace

1 引言

随着 Web2.0 的迅速发展，人们越来越习惯于在网络上表达自己的观点和意见，由普通用户主动发布的文本越来越多，如新闻、博客文章、产品评论、论坛帖子等。面对这些越来越多表达情感信息的文本，传统基于主题的文本分类系统已不能满足人们的需求，

① 基金项目:国家自然科学基金(71101042);高等学校博士学科点专项科研基金(20110111120014);中国博士后科学基金(2011M501041);合肥工业大学博士学位专项资助基金(2010HGBZ0607)
收稿时间:2011-11-08;收到修改稿时间:2011-12-19

迫切需要对这些情感本文进行研究和分析[1-3]。文本情感分类就是对这些信息进行有效的分析和挖掘，识别出其情感倾向——高兴、悲伤，或得出其观点是“赞同”还是“反对”。这样就可以更好地理解用户的消费习惯，分析热点事件的舆情，为企业、政府等机构提供重要的决策依据[1-3]。

文本情感分类是指通过挖掘和分析文本中的立场、观点、情绪等主观信息，对文本的情感倾向作出类别判断。目前关于文本情感分类主要的方法有(1)使用有监督的机器学习的方法对文本进行情感分类；(2)使用情感词提取文本中与情感相关的元素作为情感分类的依据。从目前最新的研究进展来看，由于自然语言理解领域还存在一些关键技术尚待研究，方法(2)相比方法(1)，性能并无明显优势[4]，因此本文关注的是有监督的机器学习的方法。目前应用于文本情感分类的机器学习方法主要有支持向量机、朴素贝叶斯、K-近邻、决策树等，并且研究的重点很大一部分集中在如何提高文本情感分类的精度上[4-6]。与此同时，集成学习理论是提高分类精度的一种有效途径，已在许多领域显示出其优于单个分类器的良好性能[7]。集成学习通过训练多个分类器并将其结果进行合成，从而显著地提高分类精度，已成为近年来机器学习领域一个重要的研究方向[7-10]。而将集成学习理论应用于文本情感分类的研究还比较少，关于集成学习理论在文本情感分类中应用的有效性还有待验证。为此本文提出基于集成学习理论的文本情感分析方法，并通过实验对三种常用集成学习方法 Bagging、Boosting 和 Random Subspace 在文本情感分类中应用的有效性进行检验。

本文其余部分安排如下：第 2 节介绍基于集成学习理论的文本情感分类方法，第 3 节介绍具体的实验设计，第 4 节对实验结果进行分析和讨论，最后一节是本文结论和未来工作的展望。

2　基于集成学习理论的文本情感分类方法

2.1　文本情感分类的特征表示方法[4,11]

使用集成学习理论对文本情感进行分类的第一步需要把数据集中的文本表示成特征向量，这样才能使用集成学习理论对文本情感进行分类。目前向量空间表示模型是文本表示的主要方法，相关研究集中在以什么语意单元作为特征，以及如何计算特征的权重两个问题上，通常以特征的出现频率作为基础计算权重，假定集合 $\{f_1, f_2, \cdots, f_m\}$ 是文档 d 中出现的 m 个特征，令 $n_i(d)$ 是特征 f_i 在文档 d 中出现的次数，则文档 d 可以有一个特征向量来表示：$d=(n_1(d), n_2(d), \cdots, n_m(d))$。

也有一些文本表示法希望借鉴自然语言处理技术，考虑被“词袋”忽略的语义单元间的联系，将词义及短语等复杂的特征应用到分类方法的文本表示中。不过这些方法在分类效果上还没有明显的优势，而且往往需要比较复杂的语言预处理，在分类时会影响分类器的速度[4]。并且考虑到本研究的重点在于探讨集成学习理论在文本情感分类中应用，为此本研究中采用上文中提到的向量空间表示模型来表示文本情感分类的特征。

2.2　基于集成学习理论的文本情感分类方法

构造集成学习的方法有很多，主要分为基于数据划分的方法（Data Partitioning Methods）和基于特征划分的方法（Attribute Partitioning Methods）[7]。其中基于数据划分的方法通过处理训练样本产生多个样本集，分类器运行多次，每次使用一个样本集。基于数据划分的方法主要有 Bagging 和 Boosting 等[8,9]。基于特征划分的方法把输入特征划分成子集，用作不同分类器的输入向量，每次使用一个特征子集。基于特征划分的方法主要有 Random Subspace 等[10]。下面我们就对 Bagging、Boosting 和 Random Subspace 做一介绍。

2.2.1 Bagging

1996 年，Breiman 提出了 Bagging（Bootstrap Aggregating）方法[8]。各成员分类器的训练集由从原始训练集中自助选取的若干样本组成，训练集的规模通常与原始训练集相当，训练样本允许重复选取。这样原始训练集中某些样本可能在新的训练集中出现多次，而另外一些样本可能一次不出现。Bagging 方法通过重新选取训练集增加了集成的差异度，从而提高了泛化能力。Bagging 具体流程及算法如图 1 和 2 所示。

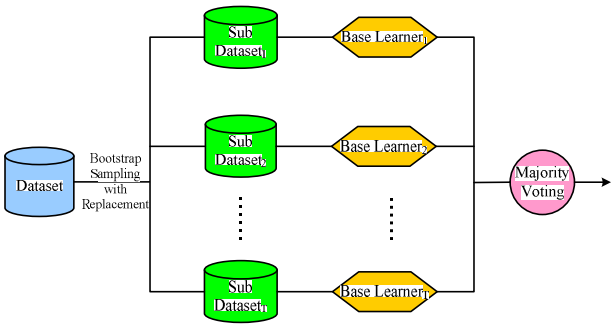

图 1　Bagging 流程图

Input: Data set $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$;

Base learning algorithm L;

Number of learning rounds T.

Process:

For $t=1,2,\cdots,T$:

$D_t = Bootstrap(D)$; % Generate a bootstrap sample from D

$h_t = L(D_t)$ % Train a base learner h_t from the bootstrap sample

end.

Output: $H(x)=\arg\max_{y\in Y}\sum_{t=1}^{T}1(y=h_t(x))$ % the value of $1(\alpha)$ is 1 if α is true % and 0 otherwise

图 2 Bagging 算法

2.2.2 Boosting

Boosting 方法最早由 Schapire 提出，其思想是对那些容易被错分的训练样本进行强化学习，首先给每个训练样本赋予相同的权重，然后使用训练的基分类器进行测试，对于那些被错判的样本提高其权重，对于那些正确判决的样本降低其权重。通过这种方法可以产生一系列分类器，各分类器的训练集决定于在其之前产生的分类器的性能，被已有分类器错误判断的样本将以较大的概率出现在新分类器的训练集中，这样新分类器将能够很好的处理对已有分类器来说很困难的样本[9]。

另一方面，虽然 Boosting 方法能够增强集成学习的泛化能力，但是同时也有可能使集成过于偏向某些特别困难的样本，因此该方法不太稳定，对噪声数据较为敏感。Boosting 是一类集成学习算法的总称，它有许多变种，其中 AdaBoost 是最为流行的方法。AdaBoost 的流程和算法如图 3 和 4 所示。

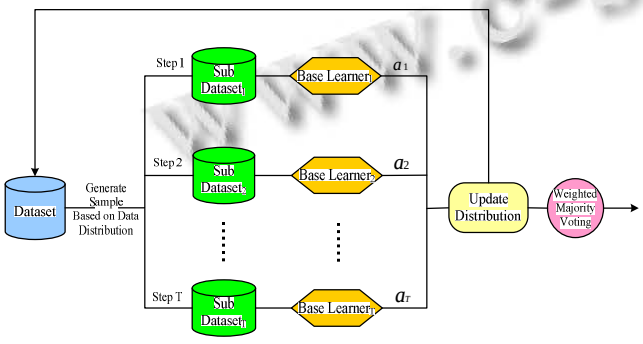

图 3 AdaBoost 流程图

2.2.3 Random Subspace

Bagging 和 Booting 都属于基于数据划分的方法，而 Random Subspace 属于基于特征划分的集成学习方法[10]。与基于数据划分的方法不同，Random Subspace 首先随机选择一定数目的特征得到不同的特征子集，然后在经过不同特征子集过滤后的数据上训练不同的分类器，最后再进行集成学习。由于 Random Subspace 需要对分类特征进行划分，因此较适用于特征空间较大的分类问题，比如文本情感分析。Random Subspace 的流程和算法如图 5 和 6 所示。

Input: Data set $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$;

Base classifier algorithm L;

Number of learning rounds T.

Process:

$D_1(i)=1/m$. % Initialize the weight distribution

For $t=1,2,\cdots,T$:

$h_t = L(D,D_t)$; % Train a base classifier h_t from D using distribution D_t

$\varepsilon_t = \Pr_{i\sim D}[h_t(x_i)\neq y_i)]$; % Measure the error of h_t

$\alpha_t = \frac{1}{2}\ln\frac{1-\varepsilon_t}{\varepsilon_t}$; % Determine the weight of h_t

$D_{t+1}(i)=\frac{D_t(i)}{Z_t}\times\begin{cases}\exp(-\alpha_t) & if\ h_t(x_i)=y_i\\ \exp(\alpha_t) & if\ h_t(x_i)\neq y_i\end{cases}$ % Update the distribution, where Z_t is a

$=\frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t}$ %normalization factor with enables D_{t+1} to be a distribution

end.

Output: $H(x)=\text{sign}\,(f(x))=\text{sign}\sum_{t=1}^{T}\alpha_t h_t(x)$

图 4 AdaBoost 算法

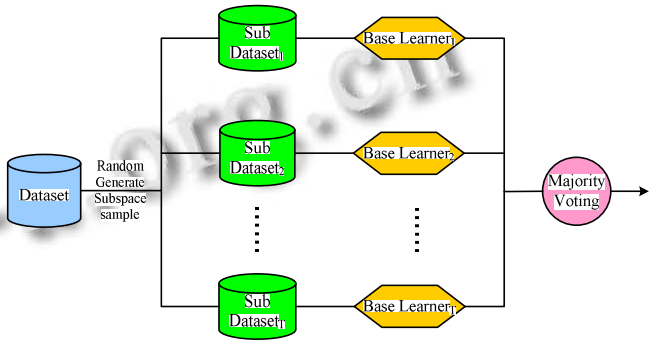

图 5 Random Subspace 流程图

Input: Data set $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$;

Base learning algorithm L;

Number of selected features rate k;

Number of learning rounds T.

Process:

For $t=1,2,\cdots,T$:

$D_t = RS(D,k)$; % Random generate a subspace sample from D

$h_t = L(D_t)$ % Train a base learner h_t from the subspace sample

end.

Output: $H(x)=\arg\max_{y\in Y}\sum_{t=1}^{T}1(y=h_t(x))$ % the value of $1(\alpha)$ is 1 if α is true % and 0 otherwise

图 6 Random Subspace 算法

3 实验设计

为了验证集成学习理论在文本情感分析领域应用的有效性，本文选取经典的语料库 Movie Reviews[12] 作为实验原始语料进行实验。该语料库包括 1000 个正面评论和 1000 个负面评论，分别存储在 POS 和 NEG 两个文件夹下。

实验的评价指标采用目前文本情感分类领域常用的评价指标：分类精度（Classification Accuracy）[1, 4-6]，定义如下：

$$Classification\ Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

本研究采用的实验环境——计算机 CPU：Intel Core 2 Duo，内存 2GB，操作系统 Microsoft Windows XP，软件 WEKA3.7.0。首先使用 WEKA 自带的 StringToWordVector 函数，剔除停用词，并将原始的文本语料转化为 WEKA 所识别的 ARFF 文件格式。然后在实验中选取了 NB、DT、KNN、SVM 作为基础分类器，对 Bagging、Boosting 和 Random Subspace 的有效性进行验证。选取 WEKA 下的 NavieBayes 模块，J48 模块（WEKA 下的 C4.5 实现）、IBk 模块和 SMO 模块来具体实现 NB、DT、KNN、SVM 算法，选取 Bagging 模块、ADBoostM1 模块和 RandomSubSpace 模块来具体实现 Bagging、Boosting 和 Random Subspace 算法，模型中各算法参数无特殊说明均取默认值。

为了提高实验结果的可信性和有效性，实验过程使用 10 次 10 倍交叉验证法，即将初始样本集划分为 10 个近似相等的数据集，每个数据集中属于各分类的样本所占的比例与初始样本集中的比例相同，在每次实验中用其中 9 个数据集组成训练集，用剩下的 1 个数据集作为测试集，轮转一遍进行 10 次试验，因此本文的实验结果为 10 次 10 倍交叉验证的平均值。

4 结果分析

本研究的主要目的在于验证集成学习理论在文本情感分析中应用的有效性，根据上节的实验设计，主要实验结果如表 1 所示。根据表 1 的实验结果，我们可以看到三种集成学习方法 Bagging、Boosting 和 Random Subspace 在分类精度上较基础分类器都有了显著提高，这也说明了集成学习理论在文本情感分类中应用是有效的。并且从分类精度上看基础分类器 DT 和 SVM 的提升效果较 NB 和 KNN 更为明显。

表 1　不同方法在 Movie Reviews 数据集上的分类精度

	NB	*DT*	*KNN*	*SVM*
Base Learner	81.16%	65.55%	55.95%	79.21%
Bagging	81.19%	74.39%	56.39%	81.26%
Boosting	81.39%	73.35%	56.00%	80.21%
Random Subspace	81.85%	74.61%	60.79%	82.54%

进一步为了分析三种集成学习理论在文本情感分析中应用的不同效果，我们分别计算不同集成学习方法在不同基础分类器条件下分类精度提高的百分比，得到图 7。如图 7 所示，在三种集成学习方法中，Random Subspace 方法较 Bagging 和 Boosting 方法应用效果更为明显，主要原因在于文本情感分类本质上也是属于文本分类的范畴，由于文本分类中存在大量冗余分类特征[13]，相对于 Bagging 和 Boosting，Random Subspace 方法主要在使用基于特征的划分方法，可以得到多样性更好的基础分类器，所以也可以得到更好的集成效果。

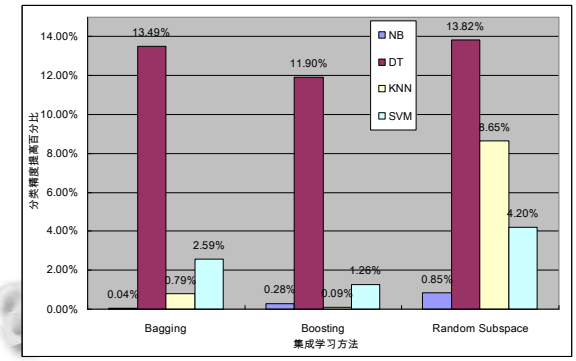

图 7　分类精度提高结果分析

通过上述分析我们看到集成学习理论在文本情感分析中应用的效果，为了确保以上分析不是偶然得到的，我们使用配对 t 检验对上述结果进行统计检验，表 2 至 5 为统计检验结果。在使用 DT 和 SVM 作为基础分类器时，Bagging、Boosting 和 Random Subspace 方法在统计上都较基础分类器有显著性提高；在 KNN 作为基础分类器时，Bagging 和 Random Subspace 方法在统计上都较基础分类器有显著性提高；在 NB 作为基础分类器时，Random Subspace 方法在统计上都较基础分类器有显著性提高。在三种集成学习方法间比较时，Random Subspace 方法较 Bagging 和 Boosting 方法

有显著性提高。

表 2 显著性统计检验结果（NB）

	NB	Bagging NB	Boosting NB	Random Subspace NB
NB	-	0.403	0.991	4.236**
Bagging NB		-	1.116	4.368**
Boosting NB			-	1.732*
Random Subspace NB				-

Notes: *P-values significant at alpha=0.1; **P-values significant at alpha=0.01.

表 3 显著性统计检验结果（DT）

	DT	Bagging DT	Boosting DT	Random Subspace DT
DT	-	22.048**	21.612**	22.296**
Bagging DT		-	-2.697**	0.650
Boosting DT			-	3.086**
Random Subspace DT				-

Notes: *P-values significant at alpha=0.1; **P-values significant at alpha=0.01.

表 4 显著性统计检验结果（KNN）

	KNN	Bagging KNN	Boosting KNN	Random Subspace KNN
KNN	-	3.139**	0.223	11.833**
Bagging KNN		-	-3.139**	10.425**
Boosting KNN			-	11.833**
Random Subspace KNN				-

Notes: *P-values significant at alpha=0.1; **P-values significant at alpha=0.01.

表 5 显著性统计检验结果（SVM）

	SVM	Bagging SVM	Boosting SVM	Random Subspace SVM
SVM	-	11.050**	2.367*	13.572**
Bagging SVM		-	-5.647**	5.290**
Boosting SVM			-	9.490**
Random Subspace SVM				-

Notes: *P-values significant at alpha=0.1; **P-values significant at alpha=0.01.

5 结语

Web2.0 环境下越来越多的用户乐于在互联网上分享自己的观点或体验，这类评论信息迅速膨胀，仅靠人工的方法难以应对网上海量信息的收集和处理，因此迫切需要计算机帮助用户快速整理和分析这些相关评价信息，文本情感分类技术在这样的背景下应运而生。

对于基于机器学习的文本情感分类技术，分类的准确性是评价机器学习方法的一个重要标准，本文提出基于集成学习理论的文本情感分类方法，并通过标准的文本情感分类数据集对三种常用的集成学习方法 Bagging、Boosting 和 Random Subspace 进行了检验。实验结果显示三种集成学习方法对基础分类器的分类精度都有提高，并且在不同的基础分类器条件下，Random Subspace 方法较 Bagging 和 Boosting 方法在统计意义上更优。在进一步的研究中，首先，本文结论仅是建立在 Movie Reviews 数据集上的，在未来研究中还需要在其它数据集以及实践中对本文结论进行验证。其次，要考虑特征选择技术在文本情感分类中的应用。本文的结论已经显示基于特征划分的集成学习方法较基于数据划分的方法要好，在未来的研究中结合特征选择技术的集成学习方法是文本情感分类技术的一个重要研究方向。

参考文献

1 Pang B, Lee L. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval. 2008,2(1-2):1–135.

2 赵妍妍,秦兵,刘挺.文本情感分析.软件学报,2010,21(8):1834–1848.

3 周立柱,贺宇凯,王建勇.情感分析研究综述.计算机应用,2008,28(11):2725–2728.

4 唐慧丰,谭松波,程学旗.基于监督学习的中文情感分类技术比较研究.中文信息学报,2007,21(6):88–108.

5 Pang B, Lee L, Vaithyanathan S. Thumbs up Sentiment classification using machine learning techniques. Proc. of the EMNLP 2002. Morristown: ACL, 2002. 79?86.

6 Cui H, Mittal VO, Datar M. Comparative experiments on sentiment classification for online product reviews. Proc. of the AAAI 2006. Menlo Park: AAAI Press. 2006. 1265–1270.

7 Polikar R. Ensemble based systems in decision making. IEEE Circuits and Systems Magazine, 2006,6(3):21–45.

8 Breiman L. Bagging predictors. Machine Learning, 1996, 24(2):123–140.

(下转第 248 页)

从图 1 和图 2 可以看出，按三种不同组卷时间要求进行组卷，参数 b 对组卷成功率和组卷质量的影响是比较一致的。b 的变化范围较大从 1.2 递减到 0.4 时，效果最佳。原因是前期较大的 b 有利于算法跳出局部最小值，提高算法的搜索能力，后期较小的 b 有利于算法快速收敛。所以，b 在[1.2，0.4]范围变化内变化比在[1.0，0.4]及[0.8，0.4]范围内变化有较好的组卷质量和较高的组卷成功率。

5 结论

本文基于项目反应理论构建了组卷问题的数学模型，基于 AMQPSO 设计一种求解组卷问题算法。实验证明，本文提出算法较基本遗传组卷算法能显著地提高组卷效率，较好地完成组卷要求。.本文基于 QPSO 改进的自动组卷算法，算法性能还有待于在实际的系统中进行测试，还需要在实践中不断完善和改进。

参考文献

1 杨军.一种改进的遗传算法在自动组卷中的应用.计算机应用与软件,2009,26(12):225−227.

2 李会民,张仁津,马桂英.基于遗传算法的交规考试自动组卷方法研究.计算机工程与设计,2009,30(18):1026−1030.

3 毛秉毅.基于遗传算法的智能组卷系统数据库结构的研究.计算机工程与应用,2008,39(6):230−232.

4 刘贝贝,肖明,马晓敏.基于推理的组卷数学建模及其应用.计算机工程,2010,36(4):195−197.

5 孟朝霞.基于自适应免疫遗传算法的智能组卷.计算机工程.2008,34(14):203−205.

6 董敏,霍剑青,王晓蒲.基于 IRT 智能组卷的模型管理系统.中国科学技术大学学报,2004,34(5):612−617.

7 刘仁金.基于粒度合成计算的智能组卷策略研究.广西师范大学学报(自然科学版),2005,23(4):33−36.

8 姜伟.基于自组织映射网络的智能组卷系统.辽宁师范大学学报(自然科学版),2005,28(3):283−284.

9 张建国.智能教学系统中的自动组卷算法研究.郑州:河南大学,2009.

10 Kennedy J, Eberhart RC. Particle swarm optimization. Institute of Electrical and Electronics Engineers, 1995,(11): 1942−1948.

11 Sun J, Feng B, Xu WB. Particle swarm optimization with particles having quantum behavior. Proc. of 2004 Congress on Evolutionary Computation. Piscataway, NJ: IEEE Press. 2004. 325−331.

12 刘俊芳,高岳林.带自适应变异的量子粒子群优化算法.计算机工程与应用,2011,47(3):41−43.

13 漆书青,戴海崎,丁树良.现代教育与心理测量学原理.北京:高等教育出版社,2002.

14 Van der Linden WJ, Boekkooi-Timminga E. A maxmin model for test design with practical constraints. Psychometrika, 1989,54(2):237−247.

15 Swanson L, Stocking ML. A model and heuristic for solving vety large items selection problems. Applied Psychological Measurement, 1993,17(2):151−166.

(上接第 181 页)

9 Schapire RE. The strength of weak learnability. Machine Learning, 1990,5(2):197−227.

10 Ho TK. The random subspace method for constructing decision forests. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1998,20(8):832−844.

11 沈凤仙,朱巧明.基于特征倾向的网页特征提取方法研究.计算机工程与设计,2009,30(16):3894−3896.

12 Online movie reviews data. http://www.cs.cornell.edu/people/pabo/movie-review-data/2010-12-28.

13 Leopold E, Kindermann J. Text categorization with Support Vector Machines: How to Represent Texts in Input Space. Machine Learning, 2002,46(1-3):423−444.