☞综合保障性技术☜

一种面向自动化标检的文本分类方法*

郭泽,焦倩倩

(北京电子工程总体研究所,北京 100854)

摘要: 针对自动化标检中的段落文本分类问题,提出一种基于机器学习的改进朴素贝叶斯分类算法。该方法对朴素贝叶斯分类算法进行改进并作为分类器,采用遗传算法作为训练模型对分类器中的所有特征权重进行训练,并采用一种基于图表位置的修正算法优化分类结果。在实际的数据集中进行了实验,结果表明,该方法与传统 KNN(K-nearest neighbor)算法和朴素贝叶斯算法相比具有更好的分类结果,能够有效的处理错误样本较多的情况,可大幅提升自动化标检的准确性。

关键词: 机器学习; 文本分类; 朴素贝叶斯; 遗传算法; 自动化标检

doi: 10.3969/j.issn.1009-086x.2020.05.015

中图分类号: TP391 **文献标志码:** A **文章编号:** 1009-086X(2020)-05-0097-08

Text Categorization Algorithm for Automatic Document Review

GUO Ze, JIAO Qian-qian

(Beijing Institute of Electronic System Engineering, Beijing 100854, China)

Abstract: A machine learning based improved native Bayes algorithm proposed to solve the text classification problem in automatic document review field. Firstly, it improves naive Bayes algorithm and applies it as the classifier. Then a genetic algorithm is adopted to train all the feature weights. Finally, a table and figure position based identification algorithm is used to improve the results. The experimental results show that the algorithm performs better than traditional (K-nearest neighbors) KNN and naive Bayes in most cases, especially when the sample sets have more wrong samples. It can improve the accuracy of automatic document review effectively.

Key words: machine learning; text categorization; naive Bayes; genetic algorithm; automatic document review

0 引言

文档是用户与产品之间最直接的桥梁,它有助于软件人员设计程序,有助于管理人员监督和管理产品,有助于维护人员进行有效的修改和改进,更是用户对产品功能、使用方式等各方面进行了解的最主要方式,其质量十分重要。在军用领域,研试文件、设计文件、软件文件等等一系列文档贯穿整个产品周期,其质量的好坏对产品的研制、试验等过程有着极其重要甚至决定性作用[1]。同时,文档作为向用户展示成果的最直接窗口,其质量更是反映了一个企业的文化。一份完美的文档能够让人

* 收稿日期: 2020-04-09; 修回日期: 2020-05-07
基金项目: 有
第一作者简介: 郭泽(1988-),男,重庆巴南人。工程师,硕士,主要从事机器学习,指控总体设计。
通信地址: 100854 北京 142 信箱 30 分箱 E-mail: guoze0987@126.com

看出企业工作的严谨态度，而一份错漏百出的文档甚至会令用户失去对企业的信心。

文档的质量已经引起各军工企业的重视，对文档质量开展的各类评审、审查等工作使文档的质量大幅提高。然而目前对文档的格式、内容的审查均完全依靠人工进行审查，审查效率不高，且受审查人水平、劳累程度等主观因素影响较大。文档的质量即使经过审查，也往往出现质量参差不齐的情况。开展自动化标检技术研究，降低人力资源消耗，提高文档产品质量十分重要。对文档的自动化标检实际是一种大规模文本的处理技术，其过程可分解为文本识别、文本标检和文本处理，其中最为核心的技术在于对文本的识别，即文本分类技术[2]。

1 基于机器学习的文本分类方法

文本分类是处理和组织大规模文本数据的关键技术，目前正广泛的应用于搜索引擎、快速资料分检、自动文摘、信息资料推送等领域[3]。自 20 世纪 90 年代以来，随着信息存储技术和计算机网络的飞速发展，机器学习逐渐取代了传统的知识工程，成为文本分类的主流技术。基于机器学习的文本分类方法一般采用向量空间模型[4]，该模型包含 3 个关键技术：特征选择、特征权重估算和文本分类器。特征选择是从原始特征集合中选择一部分特征组成分类集合，最终得到原始特征集合的一个真子集，从而达到降低原始特征空间维度的目的。特征的权重反映了该特征对于标识文本内容的贡献度和文本之间的区分度。分类器则用于依据特征的权重，采用一定的模型对文本实施分类。常用的分类器包括朴素贝叶斯[5]、最近邻分类算法(K-nearest neighbor，KNN)[6]和支持向量机(support vector machine，SVM)[7]，这几类分类器在特定的领域均有较好的应用。

与传统的文本分类问题不同，自动化标检领域的文本分类的基本单位为段落，特征向量除了文本外，段落的格式同样是决定其分类的重要特征[8]。其各类格式特征和文本特征均是相互独立的，这使得其非常适合采用朴素贝叶斯算法作为分类器[9]。由于需要进行分类的样本往往具有极强的样本倾斜性，某一类的数量(如正文)十分多，因此 *KNN* 算法不适用。此外，文本的编写中容易出现较多低级问题，使得某些特征具有一票否决的特性，支持向量机的核函数构造较为困难。综合考虑，采用朴素贝叶斯算法作为自动化标检的段落分类器。

2 分类模型与特征选取

设计一种改进的朴素贝叶斯分类算法用于段落分类。定义事件 A_i 为段落为第 i 类，事件 B_j 表示段落有特征 j，则段落可用特征向量 $\boldsymbol{X}=\{B_1,B_2,\cdots,B_j\}$ 表示。已知段落全部特征 B_1 到 B_j 时，根据贝叶斯公式，段落具有 B_1 到 B_j 特征的条件下为类型 i 的概率为

$$P(A_i\mid X)=\frac{P(A_i)\,P(X\mid A_i)}{\sum_{k=1}^{n}P(A_k)\,P(X\mid A_k)}.$$

由于各个特征相互独立，根据全概率公式，得到

$$P(A_i\mid X)=\frac{P(A_i)\,P(B_1\mid A_i),\cdots,P(B_j\mid A_i)}{P(X)}.$$

不失一般性，对于任意一个段落，在不添加任何前置条件的情况下，$P(X)$ 对于所有类为常数，公式进一步变为

$$P(A_i\mid B_1,B_2,\cdots,B_j)=\frac{P(A_i)\,P(B_1\mid A_i),\cdots,P(B_j\mid A_i)}{c},c\text{ 为常数.}$$

可以看出，任意段落为某一类型的概率与以下 2 类概率直接相关。

(1) 段落为类型 i 的先验概率[10] $P(A_i)$；

(2) 段落为类型 i 时具有特征 B_j 的概率 $P(B_j\mid A_i)$。对于任意一个段落，在书写过程中均可能出现特征与预期不符的情况。将 $P(B_j\mid A_i)$ 拆分为类型 i 的特征符合要求和不符合要求 2 种情况。

$$P(B_j\mid A_i)=\begin{cases}P, & \text{特征符合},\\ 1-P, & \text{特征不符合}.\end{cases}$$

定义 P_0 表示先验概率，P_j 表示特征 j 符合类型 i 的值。假设某段落的特征 2 不符合类型 i，其余特征均符合，则段落为类型 i 的概率为

$$P(A_i\mid B_1,B_2\cdots,B_j)=\frac{P_0\times P_1\times(1-P_2)\times\cdots\times P_j}{c},c\text{ 为常数.}$$

根据上述公式，段落的分类概率与 P_0 到 P_j 直接相关，选取合理的特征将大幅提升识别的准确性。将特征分为格式特征和文本特征 2 类，其中格式特征表示段落的格式，文本特征表示段落文字中隐含的特征属性。段落为类型 i 的概率为

$$P(A_i \mid B_1, B_2, \cdots, B_j) = \frac{\text{先验概率} \times \text{格式特征概率} \times \text{文本特征概率}}{c},$$

c 为常数.

格式特征为通用特征,即每个段落都具备的特征,是进行分类的基础特征。文本特征为特有特征,当某些段落具备特殊的文本特征时,该段落属于某一类型的概率提升,属于其他类型的概率降低。任意段落具备类型 k 的文本特征时,属于不同类型的概率进一步分解为

$$P(A_i \mid B_1, B_2, \cdots, B_j) = \begin{cases} \dfrac{\text{先验概率} \times \text{格式特征概率} \times \text{文本特征 } k \text{ 概率}}{c}, \\ c \text{ 为常数}, i = k, \\ \dfrac{\text{先验概率} \times \text{格式特征概率} \times \dfrac{1 - \text{文本特征 } k \text{ 概率}}{\text{类型个数}}}{c}, \\ c \text{ 为常数}, i \neq k. \end{cases}$$

根据上述推导,我们选取了 19 个特征的概率值作为训练参数,选取参数如表 1 所示。

3 基于遗传算法的参数训练模型

各个特征对于最终文本分类结果的贡献度由其权重直接决定,单纯的依赖经验难以获取较好的分类结果,直接影响最终的标检质量。本文采用一种基于遗传算法的参数训练模型对 19 个特征的权重(概率)进行训练,采用一种有监督[11]的机器学习的算法,使得机器的分类结果尽可能的接近人工分类结果,各个特征的权重由样本数据决定,随着样本量的增大,其分类的准确性将有效提升。

3.1 基因设计

由于 19 个特征相互独立,且均为概率值,本文采用一种一维线性基因,每个特征的权重作为其中的一个编码,可以较为便捷的进行交叉和变异操作。

表 1 训练参数选择情况

Table 1 The choice of training parameters

序号	特征名称	特征说明	特征类型
1	字体	特征与类型要求的一致时,该段落属于该类型的概率	格式特征(通用特征)
2	字号		
3	对齐方式		
4	行距		
5	特殊缩进		
6	左侧缩进		
7	右侧缩进		
8	段前间距		
9	段后间距		
10	一级标题	起始文字为“X. ”时为一级标题的概率,X 为数字	文本特征(特有特征)
11	二级标题	起始文字为“X. X”时为二级标题的概率,X 为数字	
12	三级标题	起始文字为“X. X. X”时为三级标题的概率,X 为数字	
13	四级标题	起始文字为“X. X. X. X”时为四级标题的概率,X 为数字	
14	一级列项	起始文字为“X) ”时为一级列项的概率,X 为英文字母	
15	二级列项	起始文字为“X) ”时为二级列项的概率,X 为数字	
16	表题	起始文字为“表 X ”时为表题的概率,X 为数字	
17	图题	起始文字为“图 X ”时为图题的概率,X 为数字	
18	正文	没有特殊文本特征时默认为正文的概率	
19	先验概率	段落属于某类型的先验概率	先验概率

3.2 算子设计

选择算子采用锦标赛算子[12]，交叉算子[13]采用单点交叉和两点交叉算子，变异算子采用单点变异和位置变异算子[14]。

3.3 适应度设计

考虑到文档的段落类型的倾斜度，适应度函数以文档为单位计算分类参数的准确度，机器分类的结果与人工分类的结果越接近，则适应度越高。设 p 表示单份文档中的段落个数，q 表示机器分类与人工分类相同的段落个数，则适应度计算函数为

$$f = \frac{q}{p}.$$

依据以上设计，本文采用传统遗传算法，在适应度计算阶段将交叉、变异后的基因解析为特征权重并带入到文本分类算法中，对样本进行分类计算，将分类结果与人工结果进行自动比对，计算适应度并执行选择操作，判断是否满足准确度要求或迭代次数要求，不满足则继续进行下一代遗传，满足则输出特征权重至文本分类模型中作为最终参数。基于遗传算法的分类模型如图 1 所示。

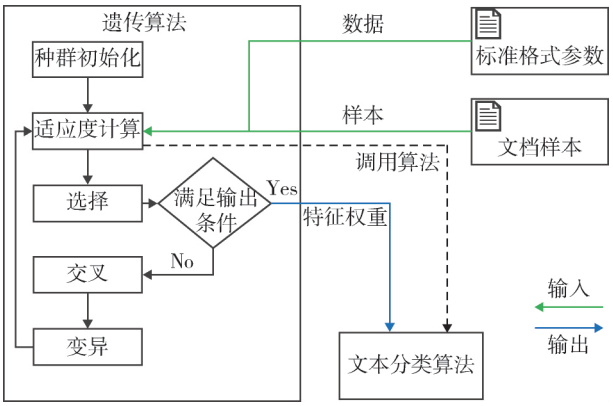

图 1 基于遗传算法的分类模型

Fig. 1 Classification model based on genetic algorithm

4 基于识别结果的自动化标检模型

自动化标检的目的是找出用户编写的文档中格式错误或文本错误的文本，其关注的重点是用户编写错误的情况。因此在文本分类时需要考虑到错误较为严重的例子，例如用户将图题、表题的格式完全写错的时候，由于段落紧跟图或表，仍应当识别为图题或表题，否则将直接影响后续标检结果。为了解决该类问题，在上面的训练和分类模型的基础上补充一种基于图表位置的图题表题识别算法优化文本分类结果。本文采用的标检流程如下。

(1) 检查文件载入：将参数配置文件载入模型中；

(2) 特征提取与筛选：提取段落的主要格式特征，剔除空段落、无效段落等干扰数据；

(3) 段落分类：为了进一步提高识别准确率，本方法加入了基于经验的先验识别算法；

图题表题识别算法(算法 1)。首先利用文字处理程序提供的 api 函数获取其中所有的图和表位置，初步识别出为表题和图题的段落；

通用识别算法(算法 2)。然后采用基于改进贝叶斯算法的分类算法计算所有段落的分类结果，记录概率最高的 3 个类型；

最后将 2 种识别算法结果进行融合。由于图题和表题通常紧跟图或表，因此通过 api 函数获取的图题表题结果可信度较高。因此，算法 1 识别为图题或表题时，直接采用算法 1 结果。算法 1 识别为非图题或非表题时，从算法 2 的结果中选取与不违背算法 1 结果的概率最高的结果。

(4) 错误检查：基于识别结果对各段落进行错误比对，记录所有的错误位置，并生成错误提示字串；

(5) 错误输出：自动统计错误情况，将所有错误在对应的位置直接以批注的形式输出错误提示字串[15]。

标检完成后将自动打开文档便于标检人员查看错误情况，同时还将在文档中标注出错误统计情况，用于直观判断文档的编写质量。

5 实验

定义文档的识别准确率如下：

$$\text{识别准确率} = \text{识别正确的段落数} / \text{总段落数} \times 100\%.$$

为了验证本文算法的效果，分别采用传统 KNN 算法、朴素贝叶斯算法和本文的改进朴素贝叶斯算法进行实验。选取质量技术处提供的实际文档作为样本，共计 5 150 个段落，样本主要选取了最常见的需要标检的 4 类文档，包括

(1) 设计文件：正确样本率 100%；

(2) 研试文件：正确样本率 80%；

(3) 软件文档：正确样本率：60%；

(4) 三大规范：正确样本率：40%。

除了模板，针对这4类文档，各随机选取了一份真实文件进行检查。

1) 模板文件识别准确率分析

各算法的模板文件的识别比较情况如图2～5所示。

可以看出，传统的KNN和朴素贝叶斯算法在处理正确率较低的样本效果较差，而本文提出的算法在各个不同正确率的样本集中均取得了95%以上的识别准确率。

2) 随机选取文件识别准确率分析

对4类文档随机选取的样本进行识别准确率分析，结果如表2所示。

可以看出，随机选取的文档识别准确率均能达到95%左右的水平。

为了验证错误提示的正确性，设计《测试文档.doc》，植入不同的错误格式。植入的错误包括：段前行距错误、字号错误、首行缩进错误、段后行距错误、字体错误、对齐方式错误、右侧缩进错误、左侧缩进错误。将部分缩进进行组合放在同一自然段，且最后2个自然段为正确格式，用于检查是否误报。

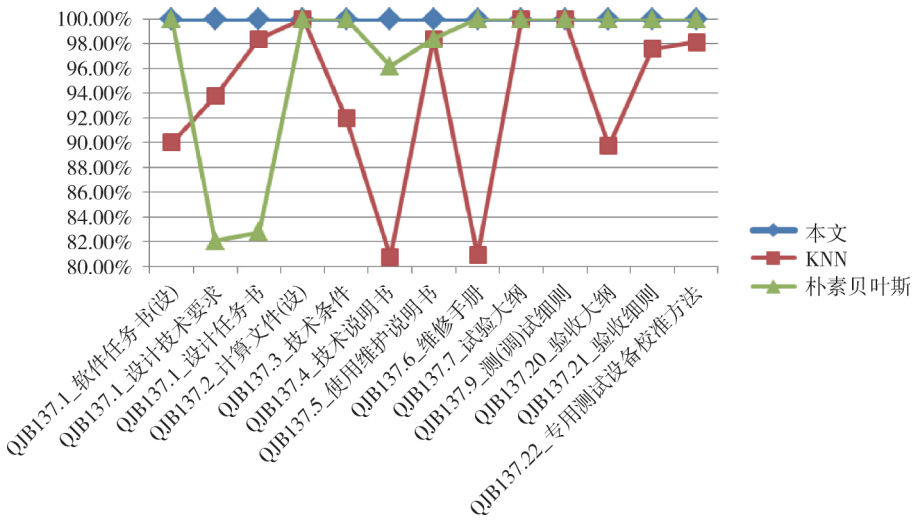

图2 设计文件识别准确率

Fig. 2 Identification accuracy result of design documents

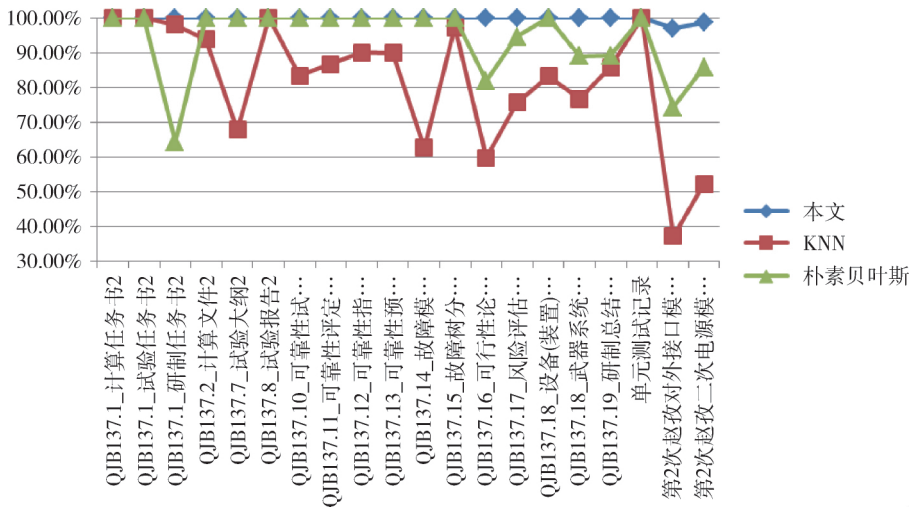

图3 研试文件识别准确率

Fig. 3 Identification accuracy result of research & experiment documents

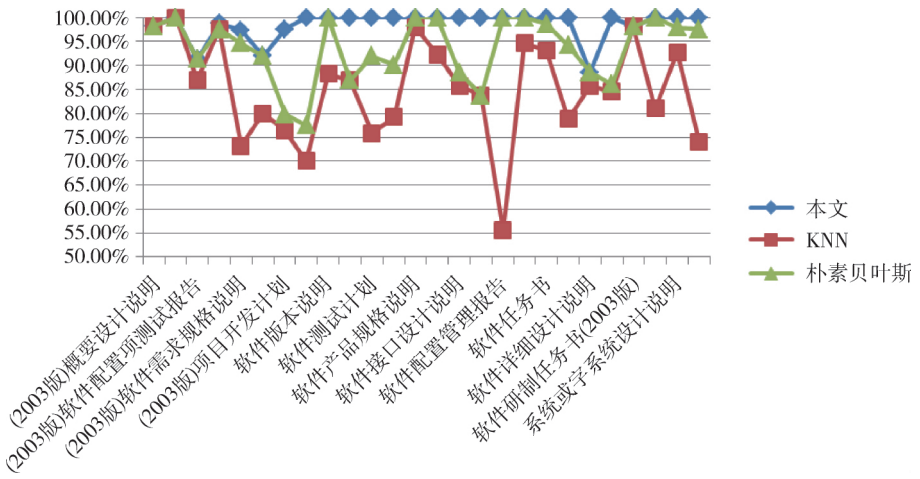

图 4　软件文档识别准确率

Fig. 4　Identification accuracy result of software documents

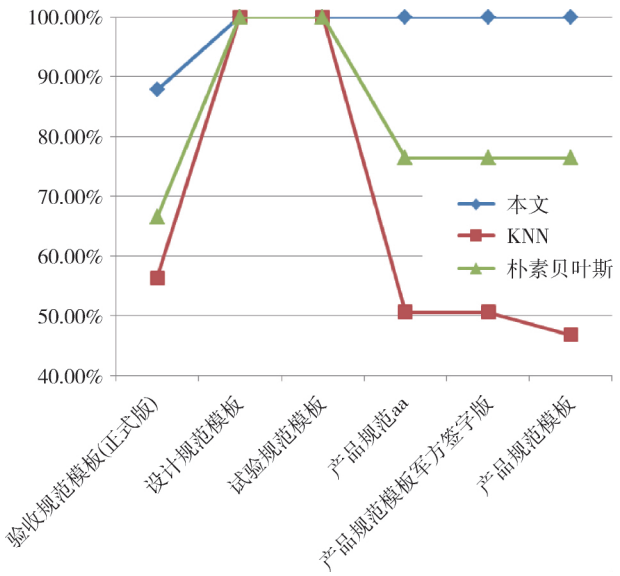

图 5　三大规范识别准确率

Fig. 5　Identification accuracy result of standards

表 2　随机文件识别准确率

Table 2　Identification accuracy result of random documents

文档类型	文档名称	页数	总段落数	识别准确率(%)
设计文件	某显示控制软件任务书	23	334	99.1
研试文件	某对外接口模块研制任务书	31	291	99.0
软件文档	某模拟器软件设计说明	152	2 104	94.5
三大规范	某设计标准	15	270	98.1

植入的错误在各段落末尾标注出设计测试文档，植入错误的分布情况如图 6 所示。使用工具进行格式检查后，自动生成错误批注，检查结果如图 7 所示。

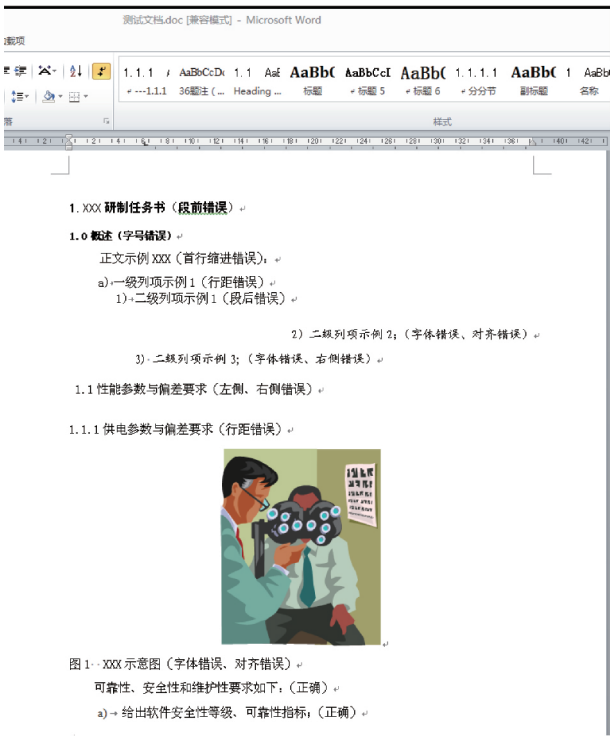

图 6　测试文档设计情况

Fig. 6　Design of test document

所有植入的错误均被工具自动识别且标注出，标注的段落位置正确。正确的段落未出现误报，预埋错误的识别率达到 100%，工具的基本格式检查

图 7 格式检查结果

Fig. 7 Result of format check

功能满足设计要求。

6 结束语

本文首先对基于机器学习的文本分类算法进行了介绍，在此基础上选取了面向自动化标检的特征向量，进而提出改进的朴素贝叶斯分类算法和基于遗传算法的分类模型。然后，在实际的数据集中分别采用 KNN 算法、传统朴素贝叶斯算法和本文的算法进行了分类。实验结果表明，本文提出的分类模型能够有效处理段落数多、错误多的情况，正确的将段落进行分类。能够有效地提高自动化标检的正确率，从而提高标检质量。

参考文献：

[1] 宁凌，韩冰洁. 基于产品数据管理系统的标准化管理[J]. 计算机与网络，2011，37(6)：42-44.
NING Lin，HAN Bing-jie. Standardization Management Based on Product Data Management System [J]. China Computer & Network，2011，37(6)：42-44.

[2] 任朋后，王芳，黄树成. 一种改进的文本分类算法[J]. 电子设计工程，2017，25(18)：1-5.
REN Peng-qi，WANG Fang，HUANG Shu-cheng. An Improved Text Classification Algorithm [J]. Electronic Design Engineering，2017，25(18)：1-5.

[3] 刘赫，刘大有，裴志利，等. 一种基于特征重要度的文本分类特征加权方法[J]. 计算机研究与发展，2009(10)：1693-1703.
LIU He，LIU Da-you，PEI Zhi-li，et al. A Feature Weighting Scheme for Text Categorization Based on Feature Importance [J]. Journal of Computer Research and Development，2009(10)：1693-1703.

[4] 吴龙峰，于璨，王峰. 向量空间模型的文本分类研究进展与应用[J]. 宿州学院学报，2019，34(12)：69-72.
WU Long-feng，YU Li，WANG Feng. Research Progress and Application of Text Categorization of Space Vector Model [J]. Journal of Suzhou University，2019，34(12)：69-72.

[5] 刘勇华. 基于朴素贝叶斯的中文段落情感分析[D]. 太原：太原理工大学，2015.
LIU Yong-hua. Analysis of Chinese Paragraphs Emotion Based on Naive Bayes [D]. Taiyuan：Taiyuan University of Technology，2015.

[6] 张孝飞，黄河燕. 一种采用聚类技术改进的 KNN 文本分类方法[J]. 模式识别与人工智能，2009(6)：936-940.
ZHANG Xiao-fei，HUANG He-yan. An Improved KNN Text Categorization Algorithm by Adopting Cluster Technology [J]. Pattern Recognition and Artificial Intelligence，2009(6)：936-940.

[7] 谢娟英. 基于 SVM 的特征选择方法研究[D]. 西安：西安电子科技大学，2012.
XIE Juan-ying. SVM Based Feature Selection Algorithms for Classificaion [D]. Xi'an：Xidian University，2010.

[8] 黄永，陆伟，程齐凯，等. 学术文本的结构功能识别——基于段落的识别[J]. 情报学报，2016，35(5)：530-538.
HUANG Yong，LU Wei，CHENG Qi-kai，et al. The Structure Function Recognition of Academic Text——Paragraph-based Recognition [J]. Journal of the China Society for Scientific andTechnical Information，2016，35(5)：530-538.

[9] 郭正斌，张仰森，蒋玉茹. 一种面向文本分类的特征向量优化方法[J]. 计算机应用研究，2017(8)：2299-2302.
GUO Zheng-bin，ZHANG Yang-sen，JIANG Yu-ru. Feature Vector Optimization Method for Text Classification [J]. Application Research of Computers，2017(8)：2299-2302.

[10] 沈宏伟，邵堃，张阳洋，等. 基于朴素贝叶斯的信任决策模型[J]. 小型微型计算机系统，2018，39(2)：275-279.
SHEN Hong-wei，SHAO Kun，ZHANG Yang-yang，et al.

Trust Decision Model Based on Naive Bayesian [J]. Journal of Chinese Computer Systems,2018,39(2):275-279.

[11] 沈荣,张保文. 机器学习学习方式及其算法探讨 [J]. 电脑知识与技术,2017,13(23):159-160.
SHEN Rong, ZHANG Bao-wen. Machine Learning's Learning Methods and Algorithms [J]. Computer Knowledge and Technology,2017,13(23):159-160.

[12] 乔家庆,付平,孟升卫. 基于个体差异的遗传选择算子设计 [J]. 电子学报,2006(S1):2414-2416.
QIAO Jia-qing, FU Ping, MENG Sheng-wei. A Genetic Selection Operator Based on Difference Among Individuals [J]. Acta Electronica Sinica, 2006 (S1): 2414-2416.

[13] 于岩,王春雨,汪洪艳. 基于改进后的实数编码遗传算法无源测向定位 [J]. 现代防御技术,2016,44(5):116-119.
YU Yan, WANG Chun-yu, WANG Hong-yan. Passive Direction Location Based on Improved Real Encoding Genetic Algorithm [J]. Modern Defence Technology,2016,44(5):116-119.

[14] 安霆. 基于遗传算法的图像分割处理技术研究 [J]. 电子技术应用,2019,45(10):92-95.
AN Ting. Research on Image Segmentation Technology Based on Genetic Algorithms [J]. Application of Electronic Technique,2019,45(10):92-95.

[15] 李建波. 基于 VSTO 的文档审阅批注自动导出技术 [J]. 计算机与现代化,2018(5):56-59.
LI Jian-bo. Auto-Exporting Technology for Word Review Comments Based on VSTO [J]. Computer and Modernization,2018(5):56-59.

(上接第 91 页)

LIU Jia-ni. Inspiration to Our Army About American Army Field Contractor Support and Their Combat Readiness [J]. Journal of the Academy of Equipment Command & Technolog,2006,17(6):15-18.

[9] 李俊杰,姜坤,黄进进. 美空军航空装备合同商保障管理研究 [J]. 飞机设计,2017,37(4):77-80.
LI Jun-jie, JIANG Kun, HUANG Jin-jin. The Management Mechanism of Air Force of USA Contractor Logistics Support Study [J]. Aircraft Design,2017,37(4):77-80.

[10] 吕岳卿,程中华,王禄超. 美军装备合同商保障分析 [J]. 价值工程,2010,23-0246-01:246.
LÜ Yue-qing, CHENG Zhong-hua, WANG Lu-chao. Analysis on Contractor Support in USA Army [J]. Value Engineering,2010,23-0246-01:246.

[11] 张红梅,刘沃野,董良喜. 基于性能的合同商装备保障研究 [J]. 装备指挥技术学院学报,2010,21(5):37-41.
ZHANG Hong-mei, LIU Wo-ye, DONG Liang-xi. Research on Performance-based Contractor Equipment Support [J]. Journal of the Academy of Equipment Command & Technolog,2010,21(5):37-41.

[12] 王凯,肖杰,多久廷,等. 信息化装备的合同商保障 [J]. 兵工自动化,2009,28(9):4-6.
WANG Kai, XIAO Jie, DUO Jiu-ting, et al. Contractor Support of Information Equipment [J]. Ordnance Industry Automation,2009,28(9):4-6.

[13] 孔学东,陆裕东,恩云飞. 电子产品 PHM 及其关键技术 [J]. 中国质量,2010(3):15-18.
SUN Xue-dong, LU Yu-dong, EN Yun-fei. Electronic Product PHM and It's Key Technology [J]. China Quality,2010(3):15-18.

[14] 王博,仲维彬. 航空故障诊断与健康管理技术研究 [J]. 现代导航,2019(6):454-457.
WANG Bo, ZHONG Wei-bin. Technology Research on Aviation Fault Diagnosis and Health Management [J]. Modern Navigation,2019(6):454-457.

[15] 蔡丽影,孙江生,连光耀,等. 我军合同商保障模式研究 [J]. 装甲兵工程学院学报,2013,27(5):25-28.
CAI Li-ying, SUN Jiang-sheng, LIAN Guang-yao, et al. Research on Contractor Support Mode of Our Army [J]. Journal of Academy of Armored Force Engineering,2013,27(5):25-28.