

用于垃圾邮件的贝叶斯过滤算法研究

曹翠玲¹, 王媛媛², 袁野¹, 赵国冬¹

(1. 哈尔滨工程大学计算机科学与技术学院, 黑龙江 哈尔滨 150001;

2. 东北林业大学机电工程学院, 黑龙江 哈尔滨 150040)

摘要: 研究了基于改进的支持向量机(SVM, support vector machine)算法结合朴素贝叶斯算法在垃圾邮件过滤中的应用。首先, SVM 对训练集样本空间中两类交界处的集合构造一个最优分类超平面; 然后, 每个样本根据其最近邻的类型是否相同进行取舍, 从而降低样本空间也提高了每个样本类别的独立性; 最后, 利用朴素贝叶斯算法对邮件分类。仿真实验结果表明, 该算法降低了样本空间复杂度, 快速得到最优分类特征子集, 有效地提高了垃圾邮件过滤的分类速度、准确率和召回率。

关键词: 朴素贝叶斯; 支持向量机; 修剪; 垃圾邮件

中图分类号: TP319

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2017.00119

Research of a spam filter based on improved naive Bayes algorithm

CAO Cui-ling¹, WANG Yuan-yuan², YUAN Ye¹, ZHAO Guo-dong¹

(1. College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China;

2. College of Mechanical and Electrical Engineering, Northeast Forestry University, Harbin 150040, China)

Abstract: In spam filtering filed, naive Bayes algorithm is one of the most popular algorithm, a modified using support vector machine(SVM) of the native Bayes algorithm :SVM-NB was proposed. Firstly, SVM constructs an optimal separating hyperplane for training set in the sample space at the junction two types of collection, Secondly, according to its similarities and differences between the neighboring class mark for each sample to reduce the sample space also increase the independence of classes of each samples. Finally, using naive Bayesian classification algorithm for mails. The simulation results show that the algorithm reduces the sample space complexity, get the optimal classification feature subset fast, improve the classification speed and accuracy of spam filtering effectively.

Key words: naive Bayes, SVM, trim, spam mail

1 引言

目前的垃圾邮件过滤技术主要有以下几种。

1) 黑白名单过滤^[1,2], 其原理是将发送方的邮箱或者 IP 放入黑名单列表中, 但当对方采用 IP 代理、动态 IP、地址隐藏、伪造等方式发送邮件时, 该方法就失效了。

2) 基于规则的过滤技术, 该技术的代表是决策树。最早的决策树学习系统要追溯到 Hunt 于 1966 年研制的一个概念学习系统 (CLS, concept learning system), 该系统第一次提出使用决策树进行概念学习, 是许多决策树学习算法的基础。随后, Quinlan 提出了迭代分类算法 ID3, 1993 年又提出 C4.5 算法^[3,4], 旨在克服 ID3 算法在应

收稿日期: 2016-10-27; 修回日期: 2016-11-25。通信作者: 曹翠玲, caocuilin0927@163.com

用中的不足。C4.5 算法对于 ID3 算法的重要改进是使用信息增益率来选择属性。2002 年, Ruggieri 提出了 EC4.5 算法^[5], EC4.5 算法采用二分搜索取代线性搜索, 还提出几种不同的寻找连续属性的局部闭值的改进策略。实验表明, 在生成同样一棵决策树时, 与 C4.5 算法相比, EC4.5 算法可将效率提高 5 倍, 但 EC4.5 算法占用内存比 C4.5 算法多。

3) 基于统计的智能学习技术, 支持向量机 (SVM)、朴素贝叶斯 (NB, native Bayes) 等都是智能学习技术。比较 SVM 和 NB 及其改进算法, 实验结果表明, 在召回率和准确率上, SVM 算法有较大优势, 但是在分类速度和训练集、测试集大小上, 朴素贝叶斯算法有明显优势。马小龙^[6]提出了 SVM-EM 朴素贝叶斯算法, 该算法先利用 SVM 算法将数据集分成完整集和缺失集, 计算缺失属性数据项与完整属性数据项的相关度, 利用 EM 算法对数据不完整属性进行修补处理, 最后利用朴素贝叶斯算法分类。SVM-EM 算法主要是根据修补不完整属性来分类的, 缺点是随着邮件数量的增多, 属性也随着增多, 其中的冗余属性也相应增加, 该算法并没有处理冗余属性, 随着邮件数量和样本集的增加, 分类速度和吞吐量就会降低。本文提出的改进的朴素贝叶斯 (TSVM-NB) 算法有效地解决了冗余属性, 提高了分类速度、准确率和召回率。该算法首先利用 SVM 对训练集样本空间中两类交界处的集合构造一个最优分类超平面, 明确每个样本根据其最近邻的类型是否相同进行取舍, 舍去冗余属性, 从而降低样本空间也提高了每个样本类别的独立性, 最后利用朴素贝叶斯算法对邮件分类, 在分类速度和准确率上都有所提高。

2 朴素贝叶斯算法及其改进

2.1 垃圾邮件过滤流程

电子邮件是基于文本形式的, 而且本身是一种无结构的文本, 为了使计算机能够对邮件进行学习和处理, 一般采用空间向量模型, 将电子邮件集用向量集合表示, 所以需要对邮件预处理。预处理包括文本分词、文本标注、特征选择、特征词权重计算等。

预处理完成后就是邮件分类, 现有的主流文

本分类方法是朴素贝叶斯算法和支持向量机算法, 两者的分类原理、使用场合、效率等各方面都有所不同。图 1 为垃圾邮件过滤的简单流程。

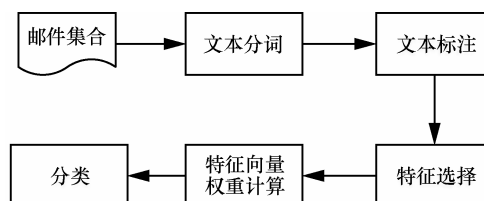


图 1 垃圾邮件过滤的简单流程

1) 文本分词是将一段连续的中文句子按照一定的规则拆分成具有一定语义的词, 想要对一句中文进行处理, 必须要把这句中文拆分成不同的词来进行处理, 这是对中文信息处理的基础。

2) 文本标注是对分词词性标注, 以便后续的特征选择, 即要确定每个词是名词、动词、形容词或其他词性, 除此之外, 还需要在集合中使用停用词表删除助词、虚词等无意义或者贡献不大的词语。

3) 电子邮件内容经过分词处理后, 形成一个代表电子邮件内容的特征向量, 这个特征向量包含了邮件内容所有被划分的词, 特征项提取是指从分词结果集中选择具有代表文章内容信息的分词。

4) 对于不同的特征选择方法, 其特征向量权重的计算方法不同, 权重代表的意义也不一样。比如, TF-IDF^[7]是根据一篇文档词如果出现频率高, 但是在其他文档出现频率低, 则说明该词具有很好的区分文档的能力, 词频方法是根据某个词出现的频率, 将出现频率小的删除。

5) 本文的重点就是分类, 下文详细介绍分类方法以及在传统的分类方法上的改进算法。

2.2 朴素贝叶斯算法模型

朴素贝叶斯文本分类原理^[8~10]是求解向量 $X(x_1, x_2, \dots, x_n)$ 属于类别 $C(c_1, c_2, \dots, c_j)$ 的概率值 $(P_1, P_2, \dots, P_n)$, 其中, P_n 为 $X(x_1, x_2, \dots, x_n)$ 属于 c_j 的概率, 则 $\max(P_1, P_2, \dots, P_n)$ 所对应的类别就是文本 X 所属的类别, 因此, 分类问题被描述为求解方程式(1)的最大值。

$$P(c_j | x_1, x_2, \dots, x_n) = \frac{P(x_1, x_2, \dots, x_n | c_j)}{P(c_1, c_2, \dots, c_n)} \quad (1)$$

其中

1) $P(c_j)$ 是训练文本中, 文本属于类别 c_j 的概率。

2) $P(x_1, x_2, \dots, x_n | c_j)$ 指的是如果待分类文本属于类别 c_j , 则类别 c_j 中包含向量 $\mathbf{X}(x_1, x_2, \dots, x_n)$ 的概率。

3) $P(c_1, c_2, \dots, c_n)$ 是给定所有类别的联合概率。

显然, 对于给定的所有类别, 分母 $P(c_1, c_2, \dots, c_n)$ 是一个已知的常数, 所以, 将式(1)简化为求解式(2)的最大值。

$$c_{NB} = \arg \max_{c_j \in C} P(x_1, x_2, \dots, x_n | c_j) P(c_j) \quad (2)$$

又根据朴素贝叶斯假设, 文本特征向量属性 $x_1, x_2, \dots, x_n$ 独立同分布, 其联合概率分布等于各个属性特征概率分布的乘积, 即

$$P(x_1, x_2, \dots, x_n | c_j) = \prod P(x_i | c_j) \quad (3)$$

所以

$$c_{NB} = \arg \max_{c_j \in C} P(c_j) \prod P(x_i | c_j) \quad (4)$$

4) 在前文提到的朴素贝叶斯算法及其改进算法利用的都是朴素贝叶斯的基本原理, 只是放松了独立性假设条件, 但是那些实际上相互不独立的属性都还是存在于训练样本集中。从式(4)中可以看出, 最后计算文本类别概率时, 用到的还是条件独立的假设, 那么实际上相互不独立的属性还是限制了算法的性能, 特别是在准确率和召回率方面, 这些算法都遇到了一定的瓶颈。那么, 有没有一种算法, 可以将独立性假设条件应用到现实世界中? 如果某个算法将所有参与到计算中的样本集属性根据其是否相关联处理, 即如果 2 个属性之间是有关系、不独立的, 就能确定这 2 个属性所属类别是否相同, 然后根据算法来处理这 2 个属性, 这就是本文提出的改进的朴素贝叶斯算法 TSVM-NB。

2.3 基于 SVM 算法的改进朴素贝叶斯算法

2.3.1 支持向量机

支持向量机^[11,12]因为显著的泛化能力而倍受人们的青睐, 原理是在特征空间内构造出一个超平面, 使两类之间的宽度达到最大, 即距离构造的超平面最远, 但还必须使类别的错分惩罚达到最小, 所以 SVM 的本质就是二次寻优问题。

在训练集可分的情况下, SVM 构造一个最优超平面

$$(\mathbf{w}^T \mathbf{x}) + \mathbf{b} = 0 \quad (5)$$

使样本集 (x_i, y_i) ($i = 1, 2, \dots, n; \mathbf{x} \in R^d, y \in \{+1, -1\}$), 满足约束条件

$$y_i [(\mathbf{w}^T x_i) + \mathbf{b}] - 1 \geq 0, i = 1, \dots, n \quad (6)$$

并且边界平面最优化, 即最小化倒数,

$$\min : \varphi(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}\|^2 = \frac{1}{2} \sqrt{\mathbf{w}^T \mathbf{w}} \quad (7)$$

通过求解最优化问题可得到最优的超平面如下。

$\sum_{i=1}^n y_i a_i (\mathbf{x}^T x_i) + \mathbf{b} = 0$, 其中, a_i 是拉格朗日乘子。

当训练集线性不可分时, 引进松弛因子 $\varepsilon_i \geq 0$ 及惩罚参数 C , 在约束条件 $y_i [(\mathbf{w}^T x_i) + \mathbf{b}] \geq 1 - \varepsilon_i (i = 1, \dots, n)$ 下最小化函数 $\varphi(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \varepsilon_i$,

分类规则只需取 $I(\mathbf{x}) = \text{sgn} \left(\sum_{i=1}^n y_i a_i (\mathbf{x}^T x_i) + \mathbf{b} \right)$ 。

核函数的引入是 SVM 算法的一大特点, 低维空间向量集往往很难划分, 那就自然想到将低维空间映射到高维空间, 但随之就会增加计算复杂度, 而核函数很巧妙地解决了这个问题。

$K(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{x}) \varphi(\mathbf{y})$, 其中, φ 表示某种映射, 只要适当选择核函数, 就可以得到对应的高维空间的分类函数

$$I(\mathbf{x}) = \text{sgn} \left(\sum_{i=1}^n y_i a_i [\varphi(\mathbf{x}), \varphi(x_i)] + \mathbf{b} \right)$$

其中, $\varphi(\mathbf{x})$ 是比 $\mathbf{x}$ 高维的向量 (无需知道 φ 的具体形式), 由于 $K(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{x}) \cdot \varphi(\mathbf{y})$ 只涉及 $\mathbf{x}$ 、 $\mathbf{y}$, 并没有涉及高维运算, 所以没有增加计算复杂度。

2.3.2 改进的朴素贝叶斯 TSVM-NB

前文提到, 朴素贝叶斯算法的使用前提条件是训练集样本中的属性是相互独立的, 利用支持向量机中的原理, 可以找到完美的一个超平面, 将两类之间的距离达到最大即两类边界处的混叠情况不会出现, 但是在实际应用中, 这种独立性假设条件是不成立的, 这就严重影响了朴素贝叶斯算法分类的召回率与正确率, 本文利用支持向量机修剪技术^[13]降低属性之间的交叉重叠, 增强其独立性, 并结合朴素贝叶斯算法分类速度快的优

垃圾邮件与正常邮件的全域范围内, 计算每个特征 X 的 IG 值, 然后按照 IG 值大小排序, 依次选择所需数量作为特征。选择完成之后构成特征向量, 特征向量代表该邮件。

3) 特征向量构成之后先用朴素贝叶斯算法对特征向量初次训练, 得到初始的特征向量训练集合及其类别。

4) 用 TSVM 对 3) 中的特征向量修剪, 修剪的目的是降低特征属性之间的独立性约束, 即降低维度, 使特征向量集合减少冗余。修剪之后得到修剪后的训练集。

5) 朴素贝叶斯算法根据修剪后的训练集对邮件分类。

4 实验结果分析

本文所有实验都是在普通 PC (Intel CORE 7i, 2.60 GHz CPU, 8.0 GB RAM), 软件为 MyEclipse 8.5, 算法语言为 Java 和 Matlab 实现提出的 TSVM-NB 算法, 实验数据来自数据堂 DATAMALL 的 5 000 封正常邮件和 5 000 封垃圾邮件, 其中, 4 000 封垃圾邮件和 4 000 封正常邮件作为训练集, 其余的作为测试集。表 1 是对数据集的基本描述。

编号	数据集	例子数量	属性数量	类数
1	垃圾广告	2 245	27	3
2	黄色暴力	1 520	15	2
3	特殊乱码字符	836	10	2
4	特殊言论	399	16	3
5	正常工作邮件	2 546	31	4
6	正常交流邮件	2 454	28	5

垃圾邮件和正常邮件都来自不同的领域, 并且涉及的垃圾类别也不一样。例如, 垃圾广告中就包含很多不正常营销、推销、培训等垃圾信息; 特殊乱码字符类垃圾邮件往往是在一些乱码字符中夹杂一些上述垃圾广告或者黄色暴力广告; 特殊言论是包含一些敏感词汇, 宣传不正当宗教, 威胁国家安全等的一些言论信息。正常的工作和交流邮件就是人与人之间基本沟通的邮件, 当然这些邮件当中可能也包含正常广告、营销类等内容。

最后从召回率、正确率以及在不同训练集数

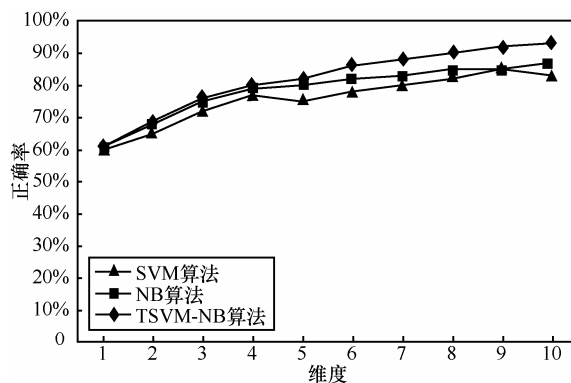
量下的运行速度等指标来评估比较朴素贝叶斯算法、支持向量机算法以及利用 SVM 改进的朴素贝叶斯算法。

$$\text{召回率} = \frac{\text{检测出的垃圾邮件数量}}{\text{样本中垃圾邮件总数}}$$

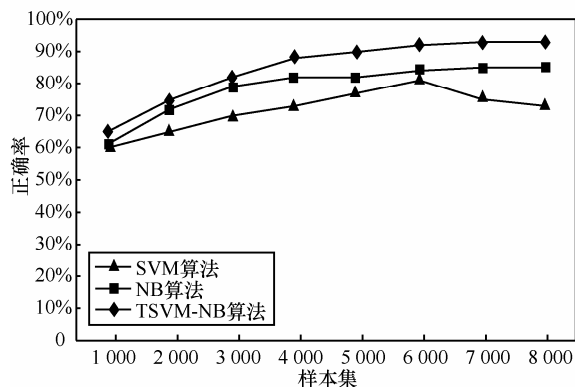
$$\text{正确率} = \frac{\text{正确识别出的邮件数}}{\text{邮件总数}}$$

本文实验特征选择采用信息增益 (IG) 方法, 多维度分析算法在不同的过滤阈值、不同样本集数量实验结果, 并从召回率、正确率、分类速度、支持向量个数等方面比较 3 种算法。

如图 3 所示, 越多的训练样本集结果越精确, 但是过多的训练样本集使向量个数增加, 而且过多样本集使代表性向量增加的同时, 冗余向量、无用向量也增加, 这使计算量跟着增加, 大大降低了分类速度, 利用 TSVM-NB 算法修剪向量, 减掉冗余和无用向量, 降低向量个数从而增加计算速度 (图 4 与图 3 类似)。图 5 是 3 种算法在不同的样本集下支持向量个数比较。改进的算法支持向量个数减少, 计算速度就会明显提高, 3 种算法的速度比较如图 6 所示。



(a) 不同维度下3种算法正确率比较 (样本集4000)



(b) 不同样本集下3种算法正确率比较 (维度5)

图 3 3种算法正确率对比

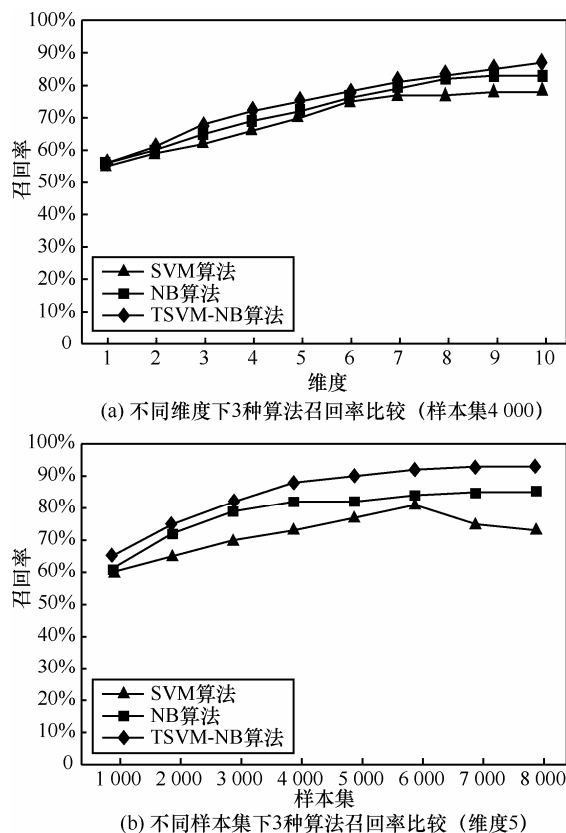


图4 3种算法召回率对比

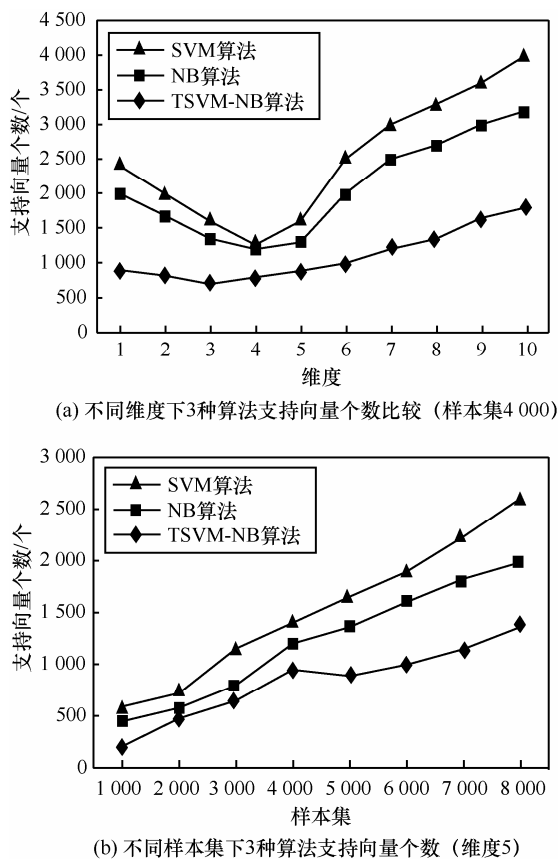


图5 3种算法支持向量个数对比

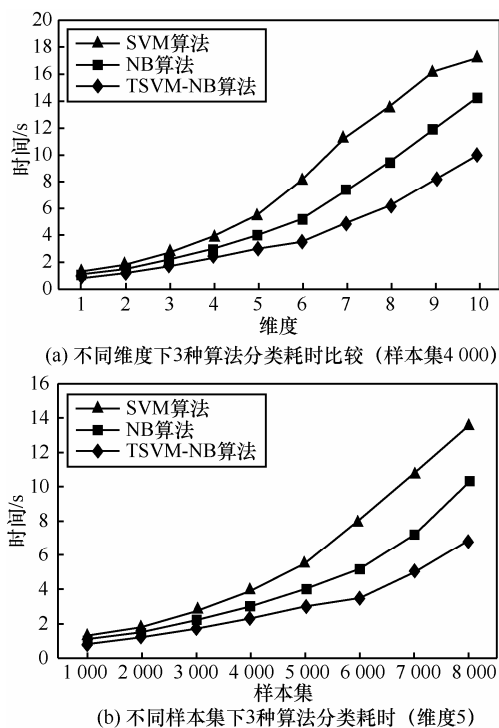


图6 3种算法分类耗时比较

从图6中可以看出,在样本集数量较小的情况下,3种算法的分类速度没有太大的区别,但是随着样本集的增加,SVM算法和NB算法的分类所用时间上升很快。利用SVM算法改进的朴素贝叶斯算法耗时虽然增加(随着样本集数量的增加,耗时增加这是必然的),但是耗时增加比较慢,所以相对来说,该算法一定程度上降低了耗时增长的速率,从而提高了分类速度。

将训练集分为1 000封、2 000封、3 000封、4 000封、5 000封、6 000封、7 000封、8 000封等8个阶段,每个阶段分别用SVM算法、NB算法计算TSVM-NB的正确率和召回率,从图3和图4可以看出,在训练集比较大的情况下,SVM算法不管是正确率还是召回率都不如其他2种算法,并且达到一定量之后2种指标反而下降,这是因为SVM算法不适合在大量邮件集下应用,朴素贝叶斯算法比SVM效果好,但实验中当邮件训练集超过4 000封时,召回率和正确率也有所下降,改进的朴素贝叶斯算法TSVM-NB算法不管是正确率还是召回率在一定程度上都有所提高。

5 结束语

本文在支持向量机算法和朴素贝叶斯算法的基础上,针对朴素贝叶斯算法的限制——属性相

互条件独立, 用 SVM 寻找最优平面, 修剪重叠属性, 增强属性独立, 提出了改进的朴素贝叶斯算法 TSVM-NB, 并根据垃圾邮件系统的评价指标正确率和召回率评估该算法, 经过大量实验, 证明该算法可以在一定程度上提高垃圾邮件处理的正确率、召回率以及分类速度。

该算法主要是适用于属性向量之间的交错重叠特别严重的数据集中, 即类别划分不是特别容易的情况, 如果数据集之间混叠性较弱, 该算法的优势就体现不出来。

随着科技的发展, 垃圾邮件不仅局限于文本形式, 还存在垃圾图片、垃圾视频、垃圾音频等各种形式, 本文研究算法只是针对文本形式的垃圾邮件, 如何高效过滤图片、视频、音频将会在以后工作中进行研究。

参考文献:

- [1] [EB/OL].<http://www.anti-spam.org.cn/>.
- [2] JI W Y, KIM H, HUH J H. Hybrid spam filtering for mobile communication[J]. Computers & Security, 2009, 29(4):446-459.
- [3] HAIBO H, GARCIA E A. Learning from imbalanced data[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.
- [4] WU X, KUMAR V, ROSS QUINLAN J, et al. Top 10 algorithms in data mining[J]. Knowledge and Information Systems, 2008, 14(1): 1-37.
- [5] RUGGIERI S. Efficient C4.5[J]. IEEE Transactions on Knowledge & Data Engineering, 2002, 14(2):438-444.
- [6] 马小龙. 一种改进的贝叶斯算法在垃圾邮件过滤中的研究[J]. 计算机应用研究, 2012, 29(3):1091-1094.
MA X L. Research of spam-filtering based on optimized native Bayesian algorithm[J]. Application Research of Computer, 2012, 29(3): 1091-1094.
- [7] SCHOLKOPF B, MIKA S, BURGESS C, et al. Input space versus feature space in kernel-based methods[J]. IEEE Transactions on Neural Network, 1999, 10(5):1000-1017.
- [8] FRIEDMAN N, GEIGER D, GOLDSZMIDT M. Bayesian network classifiers[J]. Machine Learning, 1997, 29(2/3):131-163.
- [9] 石洪波, 王志海, 黄厚宽, 等. 一种限定性的双层贝叶斯分类模型[J]. 软件学报, 2004, 15(2):193-199.
SHI H B, WANG Z H, HUANG H K, et al. A restricted double-level Bayesian classification model[J]. Journal of Software, 2004, 15(2): 193-199.
- [10] 王双成, 杜瑞杰, 刘颖. 连续属性完全贝叶斯分类器的学习与优化[J]. 计算机学报, 2012, 35(10):2129-2138.

WANG S C, DU R J, LIU Y. The learning and optimization of full Bayes classifiers with continuous attributes[J]. Chinese Journal of Computer, 2012, 35(10):2129-2138.

- [11] 曾志强, 高济. 基于向量集简约的精简支持向量机[J]. 软件学报, 2007, 18(11): 2719-2727.
ZENG Z Q, GAO J. Simplified support vector machine based on reduced vector set method[J]. Journal of Software, 2007, 18(11): 2719-2727.
- [12] 李晓黎, 刘继敏, 史忠植. 基于支持向量机和无监督聚类相结合的中文网页分类器[J]. 计算机学报, 2001, 24(1):62-68.
LI X L, LIU J M, SHI Z Z. A Chinese Web page classifier based on support vector machine and unsupervised clustering[J]. Chinese Journal of Computer, 2001, 24(1): 62-68.
- [13] 李红莲, 王春华, 袁保宗. 一种改进的支持向量机: NN-SVM[J]. 计算机学报, 2003, 26(8): 1015-1020.
LI H L, WANG C H, YUAN Z B. A improved SVM: NN-SVM[J]. Chinese Journal of Computer, 2003, 26(8): 1015-1020.

作者简介:



曹翠玲 (1990-), 女, 河北邯郸人, 哈尔滨工程大学硕士生, 主要研究方向为网络信息安全、嵌入式系统。



王媛媛 (1995-), 女, 黑龙江哈尔滨人, 东北林业大学本科生, 主要研究方向为信息安全。



袁野 (1995-), 男, 黑龙江北安人, 哈尔滨工程大学本科生, 主要研究方向为嵌入式系统。

赵国冬 (1978-), 黑龙江大庆人, 博士, 哈尔滨工程大学讲师, 主要研究方向为机器人、信息安全。