文章编号:1007-1423(2016)11-0055-04 DOI:10.3969/j.issn.1007-1423.2016.11.013

基于支持向量机与人工免疫系统的垃圾邮件过滤模型

蒋亚平,梅骁

(郑州轻工业学院计算机与通信工程学院,郑州 450000)

摘要:

垃圾邮件是互联网的主要问题,因为它会造成资源的浪费和网络环境的污染。因此,垃圾邮件过滤是十分必要的。经过对垃圾邮件过滤方法进行研究,给出一种支持向量机与人工免疫系统结合的过滤方法,并利用 MATLAB 对过滤方法进行仿真实验。

关键词:

垃圾邮件;合法;电子邮件;相似系数

1 垃圾邮件过滤技术的介绍

支持向量机 SVM(Support Vector Machine)的分类器分类精度优越于其他分类技术,同时提供了更高的性能与准确度,还能有效避免"维数灾难"[1]。支持向量机利用支持向量的决策,对于创建支持向量的创建,它使用训练数据集[2]。

支持向量机的线性分类器利用决策边界的限制以减少泛化误差。SVM 的关键概念是统计学习理论,统计学习理论主要是是用来确定决策边界的位置。关键概念非常简单的支持向量机如图 1 所示。

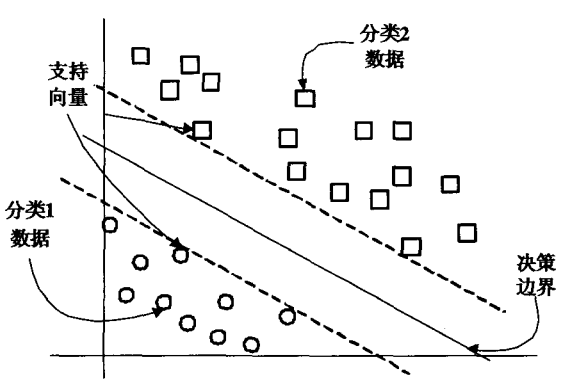

图 1 SVM 分类概述

在传统的双级分离支持向量机无限数量的决策边界被选定为支持向量,以最大限度地减少泛化误差。最接近的数据点被用来确定未知的类,这些点被称为支持向量。当一个问题到达两个类是不是线性的可区分的,SVM 方法巧妙地解决了这个难题:SVM 应用了核函数展开的相关数学定理,这样就不需要知道非线性映射的显式表达式;因为是在高维特征空间中建立的线性学习机,所以与线性模型相比,不但几乎不会增加计算的复杂性,而且在一定程度上能够有效避免"维数灾难"。选择不同类型的核函数,可以生成不同的 SVM,主要使用以下四种核函数:

①线性核函数 K(x,y)=x·y;

②多项式核函数 K(x,y)=[(x·y)+1]^d;

③径向基函数 K(x,y)=exp(-|x-y|^2/d^2)

④二层神经网络核函数 K(x,y)=tanh(a(x·y)+b).

人工免疫系统 AIS(Artificial Immune System)是一个非常复杂的对异物具有识别能力的系统[3]。生物免疫系统原理对于垃圾邮件过滤技术具有非常重要的借鉴意义,在垃圾邮件过滤系统中具有很多生物免疫系统所具备的特性,而这些特性是其他智能免疫系统几乎很少具备或完全不具备的。人工免疫系统从本质上来说是利用计算机相关技术来模拟生物的免疫系统,具有与免疫系统极其相似的功能,同时具备发现和清除

“非自体”的功能[4]。从生物免疫学的角度看,邮件过滤就是区分“自体”和“非自体”。所谓“自体”就是合法的邮件;而“非自体”就是非法的邮件。因此把人工免疫系统的相关原理和垃圾邮件过滤进行结合，是现在信息安全研究领域的一个重要热点。

表 1 生物免疫系统与垃圾邮件过滤系统的实体映射关系

生物免疫系统	垃圾邮件过滤系统
抗原	待测试邮件
自体	合法邮件
非自体	垃圾邮件
抗体	检测器
抗原识别	邮件分类
细胞亲和度	邮件相似度
记忆抗体	已知的垃圾邮件检测器
抗原提呈	邮件异常特征提取
疫苗	已知异常的预恢复策略
B 细胞、T 细胞、抗体	用特征向量表示的抗体
抗体和抗原的绑定	模式匹配算法
细胞因子	邮件通信系统

2 过滤模型设计

在模型中共设计了五个阶段：①数据集预处理；②特征选择;③特征提取;④分类;⑤结果计算。流程图如图 2 所示。

图 2 改进后的系统

(1)数据集预处理

在这个阶段对数据进行预处理，实时传入的电子邮件经过垃圾邮件过滤器进行数据集预处理。通过使用字符串编译器创建一个字典，一些无关紧要的单词在此阶段被丢弃。之后系统处理的数据传递给下一个阶段。

(2)特征选择

这一阶段的通过选择策略对上一阶段的数据进行特征选择,并将其传递到特征提取阶段。为了得到更好的选择效果,选用合适的垃圾邮件过滤策略十分重要。

(3)特征提取

在这个阶段的垃圾邮件过滤系统分析了选定的特征提取条件，并通过选取的关键词从本地语料库进行特征提取，从而为判断电子邮件是否为垃圾邮件做好分类的准备。

(4)分类

在这个阶段中，支持向量机和人工免疫系统两个分类器通过在平行工作的方式进行垃圾邮件过滤,旨在获得更高的精度和更短的响应时间。然而，一个系统，构成了过滤器的串行组合可能需要更高的平行时间比。

(5)结果计算

结果存放在一个二进制数数组中，存储在数组中的元素(0 或 1)指定分类的结果。用 0 代表垃圾邮件，用 1 代表合法电子邮件。然后加权平均计算[5]。假设权重是准确的。使用下面的这个公式进行加权平均计算：

$$M=\frac{\left\{\frac{\alpha_1\times F_1+\alpha_2\times F_2}{F_1+F_2}\right\}}{\left\{\frac{\beta_1\times \overline{F_1}+\beta_2\times \overline{F_2}}{\overline{F_1}+\overline{F_2}}\right\}}$$

在上述公式中 α 是垃圾邮件的合作效率,β 是合法电子邮件合作效率,F_1 和 F_2 是对应于 SVM 过滤规则和 AIS 过滤规则,M 是垃圾邮件的意思。

3 仿真实验

在实验中采用了一台计算机进行仿真实验，它的 CPU 频率是 2.4GHz,内存为 4GB,其操作系统环境是 Windows7,实验软件是 MATLAB。

我们在四个基准测试语料库 PU1,PU2,PU3,PU4 进行我们的实验。这些语料库进行预处理消除了 HTML 标签,附件和报头域。在 PU1 的 1099 封邮件中，480 封邮件是垃圾邮件和 619 封是合法的。在 PU2 的

表 2 仿真实验结果统计

方法 / 数据集	SVM		AIS		改进后的系统	
	垃圾邮件	合法邮件	垃圾邮件	合法邮件	垃圾邮件	合法邮件
PU1（1099）	434	665	428	671	480	619
PU2（721）	149	572	156	565	149	572
PU3（4139）	1828	2311	1809	2330	1826	2313
PU4（1142）	570	572	569	573	572	570

721 封邮件中，149 封邮件是垃圾邮件和 572 是合法的。在 PU3 的 4139 封邮件中，1826 封邮件是垃圾邮件，2313 封是合法的。在 PU4 的 1142 封邮件中，572 邮件是垃圾邮件和 570 是合法的。

4 实验结果分析

对该模型提出的过滤方法进行仿真实验，将数据集分为训练集和测试集，其平均召回率与平均精准率即为模型的正确率与平均精准率[6]。对于所有的过滤方法来说关注的主要焦点是精度问题，支持向量机人工免疫系统的比较结果示于图 3。

5 结语

本文给出了基于支持向量机与人工免疫系统的垃圾邮件过滤模型并利用 MATLAB 实现了该过滤模型的仿真实验，该模型可用于解决垃圾邮件过滤中的一些相关问题，具有一定的实际意义。

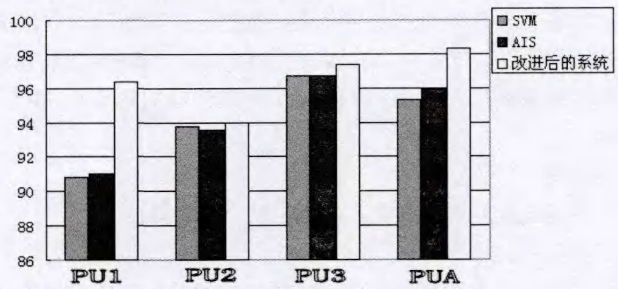

图 3 垃圾邮件过滤方法精确度比较

参考文献：

[1] Qing J J, Mao R L, Bie R F, et al. An AIS-based E-mail Classification Method[C]. The 2009 International Conference on Intelligent Computing, Ulsan, Korea, 2009:492-499.

[2]Secker A, Freitas A, Timmis J. AISEC:an Artificial Immune System for E-mail Classification[C]. The 2003 Congress on Evolutionary Computation, California, USA, 2003:131-138.

[3]刘凤玲，杨广明，王欣艳，刘莹. ARTIS 人工免疫模型在邮件过滤中的研究与应用[J]. 小型微型计算机系统，2007，28(7):1293-1296.

[4]梁刚，刘晓洁，李涛，蒋亚平，杨进，龚勋 .NSC:一种新型的垃圾邮件过滤器[J]. 小型微型计算机系统，2008，29(1):158-161.

[5]黄珏，陈兵，廖常武. 改进的人工免疫垃圾邮件过滤算法[J]. 计算机工程与应用，2011，47(30):72-74.

[6]李霞，蒋盛益. 一种垃圾邮件快速识别方法[J]. 小型微型计算机系统，2013，34(3):498-502.

作者简介：

蒋亚平(1970-)，男，河南永城人，博士副教授，硕士研究生导师，研究方向为网络技术、信息安全

梅骁(1990-)，男，河南南阳人，硕士研究生，研究方向为信息安全

收稿日期：2016-01-15　　修稿日期：2016-03-01

(下转第 80 页)

Comparative Study of SIFT and SURF in UAV Image Matching

LUO Liang, XIONG Zhu-guo

(College of Engineering of Surveying and mapping, East China Institute of Technology, Nanchang 330013)

Abstract:

The characteristics of the UAV image, in combination with the characteristics of SIFT and SURF operators, the two operators are compared in terms of feature extraction efficiency and feature extraction speed, according to the experiments on two kinds of operators in evaluating the pros and cons of UAV image matching, the experiments show that SURF has a speed advantage, also in green vegetation cover surface can match is better than SIFT operator.

Keywords:

UAV; Image Matching; SIFT; SURF

(上接第 57 页)

A Model for Spam Filtering Using Support Vector Machine and Artificial Immune System

JIANG Ya-ping[1], MEI Xiao[2]

(1. Computer and Communication Engineering, Zhengzhou Institute of Light Industry, Zhengzhou 450000;
2. Computer and Communication Engineering, Zhengzhou Institute of Light Industry, Zhengzhou 450000)

Abstract:

Spam is a major problem of the Internet, because it can cause pollution and waste of resources in the network environment. Therefore, spam filtering is necessary. After spam filtering methods study, presents a support vector machine and artificial immune system combines filtration methods and filtration methods using MATLAB simulation experiments.

Keywords:

Spam; Legitimate; E-mail; Similarity Coefficient