

一种基于神经网络和主动反馈的反垃圾邮件技术的研究

蒙海涛

(广西工业职业技术学院, 广西 南宁 530003)

摘 要: 在分析传统垃圾邮件过滤技术的基础上, 提出了一种基于神经网络和用户反馈的反垃圾邮件技术. 通过引入用户反馈机制, 使用神经网络分类方法, 构建面向特定用户的过滤器, 从而进行垃圾邮件过滤. 实验证明, 该方法能够有效地降低误报率, 提高反垃圾邮件系统的可用性, 具有较好的实用效果.

关键词: 垃圾邮件过滤; 机器学习; 神经网络; 主动反馈

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2011)06-0135-03

Research on a Spam Filtering Technique Based on Active Feedback and Neural Networks

MENG Hai-tao

(Guangxi Vocational & Technical Institute of Industry, Guangxi 530003, China)

Abstract: On the basis of analyzing the traditional spam filtering techniques, this paper presents a novel anti-spam technique based on the users' feedback. By introducing the novel users' feedback mechanism, the technique adopts a neural networks approach to construct classifiers for specific users to fulfill spam filtering. Experiments on well-known mail corpus and prototype demonstrate that the technique is able to reduce false positives and improve the availability of anti-spam system.

Key words: spam filtering; machine learning; neural networks; active feedback

1 引言

随着 Internet 的发展, 电子邮件在给人们带来巨大便利, 同时也导致了互联网中垃圾邮件的泛滥. 垃圾邮件问题已经消耗了大量的网络资源, 并给人们带来了极大的不便. 如何有效地防范垃圾邮件, 已经成为网络信息安全领域的一个难题.

目前, 反垃圾邮件技术大致可以分成如下几类^[1]: 接入控制方法, 包括黑名单、白名单、灰名单以及延迟技术; 身份验证方法, 包括 DNS MX 查询、反向 DNS 解析; 内容过滤技术, 包括基于规则的 RIPPER 方法、基于统计的贝叶斯方法、ME(最大熵)模型、SVM(支持向量机)方法等; 以及, 2005 年最新出现的基于 URL 精确匹配过滤和基于 URL 聚类的过滤方法^[2,3]等.

针对上述传统反垃圾邮件技术的不足, 研究如何能够针对基于机器学习的垃圾邮件过滤方法有效地减少其误报率, 并能针对用户的倾向来合理地地区分垃圾邮件, 从而提高反垃圾邮件技术的正确性以及可用性. 因此, 文中提出了一种新型的基于用户反馈和神经网络分类方法的反垃圾邮件技术.

2 基于神经网络的垃圾邮件建模

神经网络^[4]由于具有良好的非线性映射能力和对任意函数的准确逼近能力, 在分类领域得到了广泛的应用. 神经网络可以不对处理数据的概率分布函数做任何假设, 所以不受统计学方法中前提性假设条件的限制^[5]. 特别是当可以获得的信息仅仅由训练数据来提供时, 神经网络能更为有效地进行分类. 在分类领域用的比较普遍的是 BP^[6]神经网络, 这里介绍

收稿日期: 2010-09-02; 修回日期: 2010-10-22

基金项目: 国家自然科学基金项目(60573134)

一个三层前馈型的神经网络:网络输入层、网络隐含层、网络输出层. 假设 $(x_1, x_2, \dots, x_N)$ 为网络的输入层, N 为原始特征集中元素个数;网络隐含层的神经元个数为 n_h ;最后一层网络输出层具有 n_L 个节点, 这一参数主要是由实际问题中类别总数来确定, 实验中该值为 2. 网络输出层和隐含层的传输函数均为对数—S 形函数, 即 $f(x) = \frac{1}{1 + e^{-x}}$. 首先利用 BP 算法进行网络学习, 使用了如式(1)所示的扩展互熵误差函数来训练网络, 其中互熵误差函数 E_0 (Cross-Entropy Error Function) 为标准误差项, 具体取值可以由式(2)计算, 另外二项均为正则化项.

$$E = \frac{E_0}{n_L} + \alpha_1 \frac{1}{P n_h} \sum_{p=1}^P \sum_{k=1}^{n_h} f'(a_{kp}^h) + \alpha_2 \frac{1}{P n_L} \sum_{p=1}^P \sum_{j=1}^{n_L} f'(a_{jp}^L) \tag{1}$$

可以看出, 所添加的二项有效地限制了传输函数的导数, 使得隐含层和输出层的神经元工作在饱和和区域内. 利用限制隐含层传输函数导数的正则化神经网络具有很好的泛化能力. 在用网络来解决问题时, 要想使网络获得最低的泛化误差, 必须要求隐含层和输出层中各个节点具有不同的敏感度, 所以在误差函数中使用了 α_1, α_2 这 2 个正则化参数. 式(2)中 P 为训练样本的个数, a_{kp}^h, a_{jp}^L 分别为输入为第 p 个数据时隐含层中第 k 个神经元和输出层第 j 个神经元的输入值, 计算公式为:

$$a_{kp}^h = \sum_{i=0}^N \omega_{ip}^o x_{ip}, \quad a_{jp}^L = \sum_{i=0}^{n_h} \omega_{ij}^h O_{ip}^h \tag{2}$$

其中, ω_{ik}^o 为网络第 i 个输入与输入层第 k 个神经元之间的权值, x_{ip} 为输入数据样本中第 p 个数据时网络的第 i 个输入值, ω_{ij}^h 为隐含层第 i 个神经元与输出层第 j 个神经元之间的权值, O_{ip}^h 是指网络输入为第 p 个数据时隐含层中第 i 个神经元的输出值, 计算公式为 $O_{ip}^h = f(a_{ip}^h)$. 式(2)中, $f'(a_{kp}^h), f'(a_{jp}^L)$ 分别为网络输入为第 p 个数据时传输函数在隐含层中第 k 个神经元和输出层中第 j 个神经元处的导数.

$$E_0 = -\frac{1}{2P} \left[\sum_{p=1}^P \sum_{j=1}^{n_L} (d_{jp} \log O_{jp}^L + (1 - d_{jp}) \log(1 - O_{jp}^L)) \right] \tag{3}$$

其中, d_{jp} 为网络输入为第 p 个样本数据时第 j 个神经元处的输出期望值.

3 基于反馈的面向特定用户的过滤器

传统的垃圾邮件过滤系统都是面向所有用户而

言的, 其过滤标准是针对用户的特定需求来进行区分, 这将导致系统的误报率较高. 因此, 文中提出了采用用户反馈机制, 构建面向特定用户的过滤器的方法, 来对用户的邮件分别进行过滤.

该过滤器的构造及维护包括如下两个阶段:

(1) 构造阶段. 系统需要根据用户的反馈(对系统中的邮件进行标记, 确认为“Spam”或者“Ham”), 采用上述神经网络方法对用户确认的垃圾邮件进行机器学习, 并形成智能的垃圾邮件过滤器. 这样, 系统中的每个用户都有其独立的垃圾邮件过滤器.

(2) 维护阶段. 系统为了动态地保证该过滤器的完备性, 需要定期(可设定)向用户“征求”反馈信息, 并再次使用神经网络方法进行自学习, 完善该过滤器. 同时, 如果在此限定期间用户需提交反馈信息, 也可以主动向用户反馈收集器发送反馈信息, 并由系统自动进行再学习. 另外, 为了优化过滤器的结构, 系统能够对用户提供的相同的反馈信息进行提取和归纳, 以学习生成各用户通用的垃圾邮件过滤器. 因为据我们统计, 一个邮件系统的用户在对垃圾邮件的判定上, 对 85% 以上的垃圾邮件有相同的判定结果. 因此, 通过生成通用的垃圾邮件过滤器, 可以减少各用户过滤器的部分学习压力, 提高系统的处理效率, 且丝毫不会影响到系统的功能.

该过滤器的整体框架如图 1 所示.

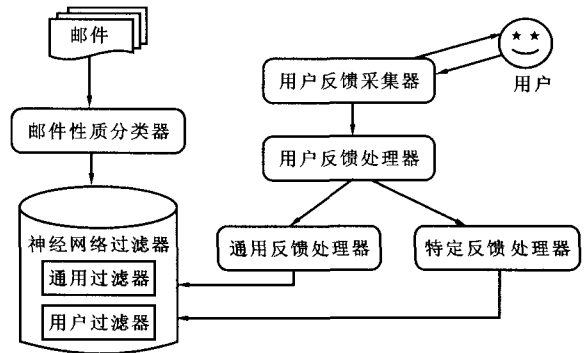


图 1 基于反馈的面向特定用户的过滤器框架示意图

该过滤器的处理流程是: 当新邮件到来时, 首先通过“邮件性质分类器”根据邮件的性质作预处理, 然后再分别传入“神经网络过滤器”进行判别. 判别的依据已形成的两类过滤器: 面向所有用户的“通用过滤器”以及面向特定用户的“用户过滤器”.

对于用户反馈的处理过程来说, 用户“反馈采集器”向用户采集反馈信息. 采集的方式可以是定时的主动采集, 也可以是由用户驱动的随机的被动采集. 提交的用户反馈经过“用户反馈处理器”的分别处理—采用“通用反馈处理器”提取用户反馈的通用信

息,传入“通用过滤器”进行再学习;采用“特定反馈处理器”提取不同用户的不同反馈信息后,将其传入“用户过滤器”进行再学习.该过滤器就是通过上述的过程周而复始地不断完善,而进行垃圾邮件的过滤工作.

4 实验与分析

4.1 实验方法及结果

为了验证神经网络过滤方法的正确性和效率,采用经典的 SpamAssassin 语料进行验证.该语料共有大约 9150 篇邮件,包括大约 2000 篇垃圾邮件和 7150 篇合法邮件.特征选择方法使用信息增益方法,过滤器算法神经网络采用后向传播模型.使用交叉验证的方法,将整个语料分成均等的 10 份,每次取 9 份做训练,剩下 1 份做测试.在实验中,将文中方法与经典的朴素贝叶斯方法^[2]进行了对比,表 1 为实验的结果.

表 1 神经网络方法测试结果

特征个数	误报率(Naive Bayes)	误报率(神经网络方法)
50	0.450712	0.261012
100	0.406958	0.225778
200	0.344755	0.230364
500	0.279916	0.183577
1000	0.241434	0.194221
2000	0.218239	0.167146

另外,为了验证面向用户反馈和神经网络方法在实际应用中的效果,基于开源的 Spamassasin^[7]系统开发了一个垃圾邮件过滤的原型系统,并将其部署在本研究室的邮件网关上.通过用户反馈及数据分析,得出表 2 所示实验数据.

表 2 垃圾邮件过滤原型系统测试结果

过滤邮件总数	93 500/封
进入邮件	41 400
发出邮件	52 100
正常邮件总数	62 500
垃圾邮件总数	31 000
误报率	0.24%
正确率	99.83%

4.2 实验分析

实验结果表明,通过使用神经网络方法根据邮件性质进行分类的预处理,可以提高该机器学习方

法的正确率,降低误报率;尤其是通过引入用户反馈机制和为每个用户构建过滤器,可以极大地满足用户的实际应用需求对垃圾邮件进行过滤,大大降低误报率,提高反垃圾邮件技术和系统的可用性.并且,用户对 85%以上的垃圾邮件都有共同的判定结果,过滤器的维护以及再学习代价比较小,占用的系统存储资源和计算资源都相对较小.

5 结束语

目前,在实验室的网关上成功地部署和使用该方法进行邮件过滤,并取得了一定的效果,用户的反映也表明该方法比较有效和实用.为了将该方法推广到大型的邮件系统中使用,对于系统为每个用户维护的过滤器所生成的开销以及对处理能力等各方面提出的应用问题作进一步的解决和优化.

参考文献:

[1] Perone M. An overview of spam blocking techniques[Z]. Cupertino, CA: Technical report, Barracuda Networks corp. 2004.

[2] Spam StopHere corp. How URL Spam Filtering Beats Bayesian/Heuristics Hands Down[Z]. Parkland Plaza: Greenview Data Inc. 2005.

[3] Hershkop S, Stolfo S J. Combining email models for false positive reduction[C]// The Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA: ACM, 2005.

[4] Castellano G, Fanelli A M. Variable selection using neural—network models[J]. Neurocomputing, 2000, 31(1):113.

[5] Reed R. Pruning algorithms—A survey[J]. IEEE Trans on Neural Network, 1993, 4(3):740—662.

[6] Sung A H, Mukkamala S. Identifying important features for intrusion detection using support vector machines and neural networks[C]// Proc. of the 2003 Int Sym On Applications and the Internet Technology. Washington, D C, USA: IEEE Computer Society Press, 2003:209—216.

[7] The apache spamassassin project[EB/OL]. [2010—09—02]. <http://spamassassin.apache.org/>.

作者简介:

蒙海涛 男,(1971—),硕士,讲师.研究方向为计算机应用技术.