基于贝叶斯神经网络的垃圾邮件过滤方法

李惠娟 高峰 管晓宏 黄亮
(西安交通大学系统工程研究所，陕西 西安 710049)

摘 要： 垃圾邮件过滤是当前互联网应用中急需解决的一个重要课题，日益受到人们的关注。本文提出了一种基于贝叶斯神经网络BNN(Bayesian Neural Network)的垃圾邮件过滤方法，利用贝叶斯推理和神经网络相结合的贝叶斯神经网络算法对用户给定的正常/垃圾邮件集合进行训练，得到邮件过滤模型。并且提出了一种有效的特征选择方法，采用信息增益准则，有效降低了特征维数。经过实验测试，本文提出的方法可以实现对垃圾邮件的有效过滤。

关键词： 贝叶斯神经网络，垃圾邮件，特征选择，信息增益，分类器
中图法分类号： TP181 **文献标识码：** A **文章编号：** 1000-7180(2005)04-107

A Spam Filtering Method Based on Bayesian Neural Network

LI Hui-juan, GAO Feng, GUAN Xiao-hong, HUANG Liang
(Systems Engineering Institute, Xi'an Jiaotong University, Xi'an 710049 China)

Abstract: Spam filtering is an important task in the application of Internet, which receives increasing emphasis. In this paper a method of spam filtering based on the Bayesian Neural Network (BNN) algorithm is presented. The Bayesian approach is used for neural network to learn from the user given training spam/normal e-mail set. And the number of features needed as the input of the BNN model is reduced effectively through the proposed feature selection approach, where information gain is chosen as a criterion. The result of the experiment shows that the method in this paper can filter spam effectively.

Key word: Bayesian neural network, Spam, Feature selection, Information gain, Classification

1 引言

随着互联网的迅速发展，作为互联网应用成功典范的电子邮件越来越受到人们的青睐。但是垃圾邮件的出现和不断泛滥，严重干扰了人们正常的网络活动。据中国互联网网络中心(CNNIC)2004年1月公布的《中国互联网络发展状况统计报告(2004/1)》显示，中国网民平均每周收到13.7封电子邮件，其中垃圾邮件占据了7.9封，占收到邮件总量的57.67%，这一比例已较半年前提高了两个百分点。垃圾邮件已成为继电脑病毒之后的又一问题，因此垃圾邮件过滤成为了全球所关注的热点[1]。

垃圾邮件过滤本质上属于文本分类问题，目前垃圾邮件过滤方法有很多：基于Signature的过滤通过将接收的邮件和已知的垃圾邮件进行比较，判断它们的Signature是否相同来过滤垃圾邮件。基于规则的过滤通过事先建立的一系列规则来过滤垃圾邮件。朴素贝叶斯，支持向量机，神经网络和遗传算法等方法广泛地应用于文本分类问题，也经常被用于进行垃圾邮件过滤。但这些方法还存在一些不足，一定程度上限制了它们的应用，如：基于规则的过滤是静态的，不能实时快速的更新，神经网络难以确定模型的复杂度，容易出现过拟合现象等。

文中提出的贝叶斯神经网络方法，通过在神经网络中引入贝叶斯推理，能有效地控制模型的复杂度和克服过拟合的问题。目前贝叶斯神经网络受到越来越多的重视，广泛地应用于网络安全、自动控制、人工智能、生物医学等各个领域。

2 贝叶斯神经网络算法

传统的神经网络由于存在难以控制模型复杂度和克服数据过拟合问题，阻碍了神经网络的泛化。而基于贝叶斯推理的贝叶斯神经网络，很好地解决了这些问题。

对于2类$C_i(i=1,2)$的分类问题，对第k个样本有n维的输入特征向量x_k，1维的输出y_k，表示属于C_1的概率，而属于C_2的概率为$(1-y_k)$。假定有网络模型H，此网络相应的权$w=(w_1,w_2,...w_W)$，W为权

收稿日期：2004-09-08
基金项目：国家自然科学基金(60243001,60274054)
国家863计划项目(2003AA142060)

值的总个数,则对于数据集 $D=\{(x_k,y_k),k=1,2,\ldots,K\}$ (K 为样本总数)和网络模型 H 来说,传统神经网络方法通过最小化误差函数 $E_D=-\ln(p(D/w))$,调整权值 w,找到合适的网络模型。这种方法的缺点是当训练数据有限时,难以找到适合该数据的模型。在贝叶斯神经网络方法中,同时考虑权值的先验分布和数据的似然分布,通过引入超参数控制权值的分布,最小化误差,实现权值的优化。

贝叶斯神经网络算法流程如图1所示[5]。

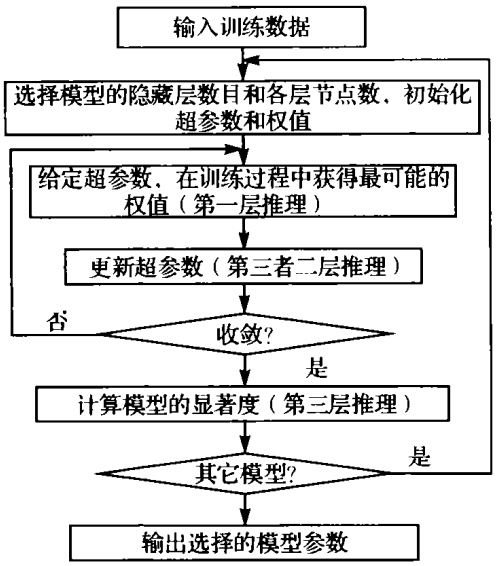


图1 贝叶斯神经网络训练流程图

根据贝叶斯规则和神经网络原理,整个过程要经过以下三层推理[6]:

Level 1:在给定网络模型 H,观察数据 D,及超参数 α 和 β 的初始值的条件下,计算权值 w 的后验概率,求得满足后验概率最大的权值 w_{mp}。

$$p(w/D,\alpha,\beta,H)=\frac{p(D/w,\alpha,\beta,H)p(w/\alpha,\beta,H)}{p(w/\alpha,\beta,H)}$$

Level 2:结合观察数据 D,计算超参数的后验概率,更新超参数 α 和 β

$$p(\alpha,\beta/D,H)=\frac{p(D/\alpha,\beta,H)p(\alpha,\beta/H)}{p(D/H)}$$

Level 3:通过比较各个模型的显著度,获得具有最大后验概率的模型,从而确定最优网络

$$p(H/D)=\frac{p(D/H)p(H)}{p(D)}$$

利用上面的推理,在假定先验分布和似然均服从指数分布的情况下,得到网络权值的后验分布为

$$p(w/D,\alpha,\beta)=\frac{1}{z_M(\alpha,\beta)}\exp(-\beta E_D-\alpha E_w)$$

$$=\frac{1}{z_M(\alpha,\beta)}\exp(-M(\omega))$$

要获得最优的权值 w_{mp},则要获得最大的后验概率,即最小的总误差 $M(w)$。目前用于权值更新的方法有很多,主要有共轭梯度法,拟牛顿法,基于马尔可夫链的蒙特卡罗数值积分方法(MCMC)等。

同样,利用贝叶斯推理,通过对 $p(\alpha,\beta/D,H)$ 和 $p(H/D)$ 求最大,也能实现超参数的更新和取得最优的网络模型。

从前面的分析可得:贝叶斯神经网络算法在训练的过程中,通过不断迭代训练数据来更新超参数,优化权值,选择网络模型。在训练中,利用超参数的更新,控制模型的复杂度,防止过拟合。

3 基于贝叶斯神经网络的垃圾邮件过滤方法

垃圾邮件过滤问题实际上是文本分类问题。所谓文本分类,即是根据一定的分类算法和预定义的类别标号,确定待分类文本的类别。根据电子邮件的自身特点,垃圾邮件过滤问题如图2所示。一般要经过文本表示、特征选择、模型训练和邮件分类几个步骤。

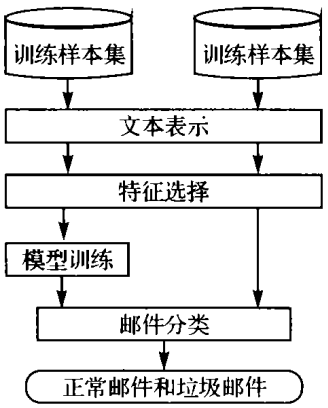


图2 垃圾邮件过滤过程框图

3.1 文本表示

因为计算机并不具有人类的智能,所以要将邮件表示成计算机能理解的形式,即进行文本表示。目前主要应用向量空间模型(Vector Space Model, VSM)进行文本表示,以特征向量来表示邮件,1个特征通常代表1个词或词组。本文采用词频表示法进行文本表示,分别计算邮件中每个特征出现的次数,即为特征向量空间中每维特征的值。

3.2 特征选择

邮件由大量的词汇组成,且构成不同邮件的词汇也各不相同,这常常使得表示邮件样本集的特征向量空间的维数很大,有的甚至达到几万维。但有些特征对邮件分类是不需要的,因此,有必要进行特征选择。

目前比较常用的特征选择方法有词频方法(Doc-

ument Frequency, DF), 信息增益方法(Information Gain,IG),互信息方法(Mutual Information, MI)等[8]。

本文采用信息增益方法进行特征选择。信息增益是信息论中的重要概念,在邮件过滤问题中信息增益表示的是考虑词汇 t 前、后确定邮件类别所需的信息量之差。词汇 t 的信息增益越大,则说明以 t 作为特征时确定邮件类别所需的信息量越小。所以信息增益定义如下:假定有2类 $C_i(i=1,2)$ 的邮件。那么词汇 t 的信息增益为[8]

$$G(t)=-\sum_{i=1}^{2}p_r(C_i)\log p_r(C_i)+p_r(t)\sum_{i=1}^{2}p_r(C_i/t)\log p_r(C_i/t)$$
$$+p_r(\bar{t})\sum_{i=1}^{2}p_r(C_i/\bar{t})\log p_r(C_i/\bar{t})$$

其中 log 是以2为底的对数,$P_r(C_i)$ 表示任意邮件属于类 C_i 的概率,$P_r(t)$ 表示包含词汇 t 的邮件数目占邮件总数的概率,$p_r(\bar{t})$ 表示不包含词汇 t 的邮件数目占邮件总数的概率,$P_r(C_i/t)$ 表示包含词汇 t 的邮件属于类 C_i 的概率,$P_r(C_i/\bar{t})$ 表示不包含词汇 t 的邮件属于类 C_i 的概率。

根据上面的定义,在只有垃圾邮件和正常邮件的两类分类问题中,信息增益又可表示如下:

$$G(t)=+P_r(t)[P_r(s/t)\log P_r(s/t)-P_r(s)\log P_r(s)]$$
$$+P_r(t)[P_r(s/n)\log P_r(s/n)-P_r(n)\log P_r(n)]$$
$$+P_r(\bar{t})[P_r(s/\bar{t})\log P_r(s/\bar{t})-P_r(s)\log P_r(s)]$$
$$+P_r(\bar{t})[P_r(n/\bar{t})\log P_r(n/\bar{t})-P_r(n)\log P_r(n)]$$
$$=PST+PNT+PSTB+PNTB$$

从上式看出,信息增益由四部分表示,*PST* 表示邮件中出现词汇 t 时为垃圾邮件的信息增益,*PNT* 表示出现词汇 t 时为正常邮件的信息增益,*PSTB* 表示不出现词汇 t 时为垃圾邮件的信息增益,*PNTB* 表示不出现词汇 t 时为正常邮件的信息增益。

在后面介绍的垃圾邮件过滤测试的例子中,取正常邮件和垃圾邮件各1000封,进行特征表示,计算各个特征的信息增益并按大小排序,得到图3。其中横轴表示特征,纵轴表示信息增益大小。

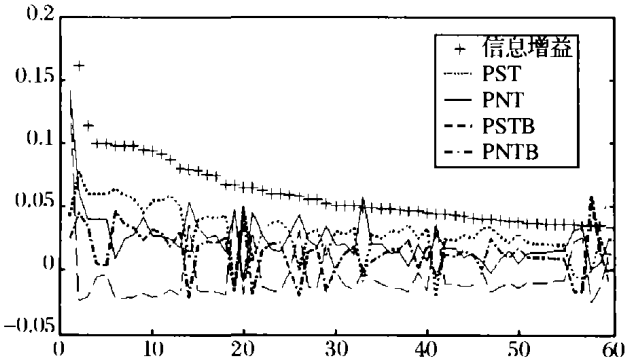
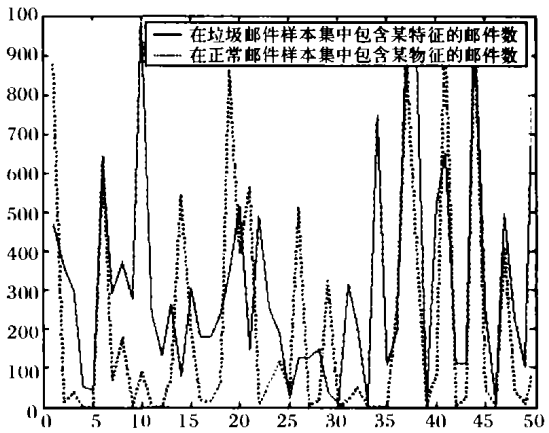


图3 信息增益曲线图

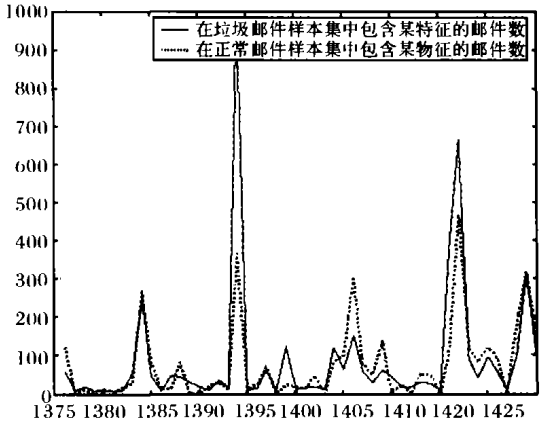
对图3进行分析得到:信息增益开始以较大的速度递减,前五个具有最大信息增益的特征分别是 In, email, receive, Bulk, Reject。在第50个特征附近信息增益的变化趋于稳定;信息增益主要由 *PST* 和 *PNT* 构成,说明信息增益主要体现的是某个特征出现时的信息量;曲线图中 *PST* 和 *PNT* 的影响刚好相对,*PST* 处于峰尖则 *PNT* 处于谷底, 即 *PST* 大说明在垃圾邮件样本集中包含该特征的邮件数目较在正常邮件样本集中多,该特征属于垃圾邮件类特征的可能性比较大,反之亦反。

同时计算在两类邮件中,包含各个特征的邮件数目,得到下面两图,其中横轴表示特征,纵轴表示包含特征的邮件数目。

图4(a)表示按信息增益从大到小排序后,在两类邮件中包含前50个特征的邮件数目的比较;图4(b)表示在两类邮件中包含最后50个特征的邮件数目的比较。通过比较发现:对于具有大的信息增益的特征,一般在两类邮件或至少在一类邮件中包含这些特征的邮件数目比较多,且在两类邮件中包含同一特征的邮件数目相差的比较大;而对于信息增益小的特征则刚好相反,一般在两类邮件中包含这些特征的邮件数目较少,且数目相差不大。这与信



(a)具有最大信息增益的50个特征



(b)具有最小信息增益的50个特征

图4 特征-包含特征的邮件数目关系图

息增益定义公式中,相关概率的计算是吻合的。

根据前面的分析,选取一定数量的具有大的信息增益的特征即可满足特征选择的要求,所以确定特征选择步骤如下:

(1) 稀有词汇过滤。删除在所有邮件中出现的次数均小于 n 次的特征(词汇)。其中次数 n 依次取 1,2,...n 等整数进行实验,通过分类结果的好坏来确定。

(2) 通用词汇过滤。删除一些通用的、在所有样本中都普遍存在的特征。以公式

$$P_p=\frac{p(s/t)}{p(n/t)+p(s/t)}$$

来评价,其中,$p(s/t)$,$p(n/t)$分别表示已知特征(词汇)t,邮件属于垃圾邮件的条件概率和属于正常邮件的条件概率,$P_p\in(a,b)$,$0<a<b<1$。

(3) 单字符过滤。删除单个字符表示的特征。

(4) 信息增益过滤。根据信息增益公式计算剩下各个特征的信息增益,并按照信息增益的大小,从大到小排列特征。然后根据信息增益的变化趋势,确定选择特征的个数。

3.3 邮件训练和分类

训练样本集经过文本表示和特征选择后,得到特征向量作为贝叶斯神经网络的输入,通过贝叶斯神经网络的训练,得到最优的网络权值和网络模型。

以同样的方法对测试样本集进行文本表示和特征选择,并利用训练好的网络进行分类。对每封测试邮件(每个测试样本)能得到一个输出 y,表示该邮件为垃圾邮件的概率,则该邮件为正常邮件的概率就为$(1-y)$。

过滤垃圾邮件时,要在保证不过滤掉正常邮件的情况下,尽可能地减少漏报垃圾邮件的数量,所以认为把正常邮件错判成垃圾邮件的代价是把垃圾邮件判为正常邮件代价的 λ 倍,只有当时 $\frac{p(C=s/\text{邮件 }k)}{p(C=n/\text{邮件 }k)}>\lambda$,才判定邮件 k 为垃圾邮件。其中 $p(C=s/\text{邮件 }k)$表示邮件 k 为垃圾邮件的概率,$p(C=n/\text{邮件 }k)$ 表示邮件 k 为正常邮件的概率,λ 为设定的阈值。采用这种标准,能降低正常邮件被错判为垃圾邮件的概率。

4 应用结果及评析

文中用到的样本集由两部分构成:从互联网新闻组数据集中提取出共 1000 封邮件作为正常邮件;从网上收集的用户认定的 1000 封垃圾邮件作为垃圾邮件。

4.1 算法实现

样本集中的邮件全部是英文邮件,所以首先以空格符和换行符作为分隔符,对邮件进行分词,每个词汇代表 1 个特征。然后根据前面制定的预处理规则进行文本表示和特征选择。考虑实验样本集较大,并且邮件正文部分内容也较长,经过多次实验,确定删除在所有邮件中出现的次数均小于 4 次的特征,再把 p_p 属于 $0.45<p_p<0.55$ 的特征删除,然后删除单个字符表示的特征,剩下的特征按照信息增益的大小进行排列,从图 3 可以看出,在第 50 个特征时信息增益的值是 0.0375,且后面特征的信息增益变化已经趋向平稳,所以取前 50 个特征作为特征向量 X。

从正常邮件和垃圾邮件中各取 900 封作为训练样本集,剩下的各 100 封作为测试样本集。把选取的训练集作为输入,训练贝叶斯神经网络。用训练好的网络对测试集进行分类,并分析分类效果。

4.2 分类结果评估

运用贝叶斯神经网络算法进行垃圾邮件过滤是一个文本分类问题,所以评估的主要标准是分类的精确度,同时,考虑到过滤垃圾邮件的实际应用,要求过滤垃圾邮件的错纠率要尽可能低。在实验中主要参考了以下指标:

错纠率:$R_{fa}=\frac{N_{n\to s}}{N_n}$

漏纠率:$R_m=\frac{N_{s\to n}}{N_n}$

精确度:$A_{cc}=\frac{N_{n\to n}+N_{s\to s}}{N_n+N_s}$

修正精确度:$WA_{cc}=\frac{\lambda N_{n\to n}+N_{s\to s}}{\lambda N_n+N_s}$

其中 $N_{n\to n}$ 表示将正常邮件归为正常邮件的数量,$N_{s\to s}$ 表示将垃圾邮件归为垃圾邮件的数量,$N_{s\to n}$ 表示将垃圾邮件归为正常邮件的数量,$N_{n\to s}$ 表示将正常邮件归为垃圾邮件的数量。错纠率 R_{fa} 表示将正常邮件错误归为垃圾邮件占所有正常邮件的比例,漏纠率 R_m 表示将垃圾邮件错误归为正常邮件占所有正常邮件的比例,A_{cc} 表示正确分类的邮件数目在所有分类邮件中所占的比例。在考虑了阈值 λ 后,有了新的计算精确度的公式 WA_{cc},对精确度进行修正。

为了更好地验证贝叶斯神经网络算法在过滤

垃圾邮件应用中的优良性能,实验中,把贝叶斯神经网络分类结果与朴素贝叶斯分类结果进行了比较,结果如表1和表2所示。

表1 贝叶斯神经网络邮件过滤结果

阈值	漏纠率	错纠率	精确度(%)	修正精确度(%)
1	0.03	0.01	98.0	98.0
9	0.06	0.00	97.0	99.4

表2 朴素贝叶斯邮件过滤结果

阈值	漏纠率	错纠率	精确度(%)	修正精确度(%)
1	0.10	0.05	92.5	92.5
9	0.12	0.04	92.0	95.2

从结果可以看出:运用贝叶斯神经网络,在 $\lambda=1$ 时,分类精确度为98.0%,此时 $R_{fa}=0.01$, $R_m=0.03$,说明有1封正常邮件被误判为垃圾邮件,3封垃圾邮件没有被过滤;当 $\lambda=9$ 时,分类精确度为99.4%,此时的 $R_{fa}=0.00$, $R_m=0.06$,说明所有的正常邮件都正确地归为正常邮件,有6封垃圾邮件被错误地归为正常邮件,上面的结果表明分类器能对垃圾邮件进行有效地过滤。同时过滤垃圾邮件的目标是保证低错纠率的情况下,尽量降低漏纠率。所以,认为 $\lambda=9$ 时能取得更好的过滤效果。

取同样的数据和特征,用朴素贝叶斯方法进行训练和测试,在 $\lambda=9$ 时,其分类精确度为95.2%, $R_{fa}=0.04$, $R_m=0.12$,这说明贝叶斯神经网络相对于朴素贝叶斯方法,无论是分类精确度还是错纠率、漏纠率都有了较大的改善。

5 结束语

垃圾邮件过滤是网络信息安全研究领域的重要组成部分,目前的反垃圾邮件技术还不够成熟,不能完全识别垃圾邮件,甚至会阻断正常邮件的交往。本文根据英文邮件的特点,用词频表示法对邮件进行文本表示。在进行特征选择时,针对垃圾邮件过滤的实际应用,扩充了信息增益的定义,详细分析信息增益与特征的选择之间的关系,在对稀有词汇,通用词汇及单字符词汇进行过滤的基础上,利用信息增益进行特征选择,极大地降低了特征维数。贝叶斯神经网络算法将神经网络与贝叶斯推理相结合,有效避免了过拟合问题。通过实验证明将贝叶斯神经网络算法应用于垃圾邮件过滤问题,在错纠率和精确度上都取得了理想的效果。

但实验中用到的数据集还仅限于英文邮件,今后的研究中,要对中文邮件实现过滤和如何更好地表示特征方面进行进一步的研究,这里首先遇到的将是中文语句的切词问题。同时,为了让垃圾邮件过滤能适应个性化的要求,还要在引入反馈,对训练集进行迭代训练以实现系统的自学习方面进行新的探索,以期取得更好的过滤效果。

参考文献

[1] Androutsopoulos, I., Koutsias, J., etc. An Evaluation of Naive Bayesian Anti-Spam Filtering, Proceedings of the workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning, Barcelona, Spain, 2000, 9~17.

[2] 阎平凡,张长水,人工神经网络与模拟进化计算,清华大学出版社,2000,303~314

[3] MacKay, D. J.C.Probable networks and plausible predictions-a review of practical Bayesian methods for supervised neural networks Network: Computation. In Neural Systems. 6 (August 1995) 469~505

[4] Androutsopoulos, I., Koutsias, I., etc. An Evaluation of Naive Bayesian Anti-Spam Filtering, Proceedings of the workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning, Barcelona, Spain, 2000, 9~17。

[5] Ma, Q. C., Wu, C.H., etc. Application of Bayesian Neural Networks to Biological Data Mining: A Case Study in DNA Sequence Classification, 4~6

[6] MacKay, D. J. C. ,Bayesian Methods for Neural Networks: Theory and Applications. Neural Computation, 4, 448-472

[7] Guyon, I., An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 2003,3: 1157~1182.

[8] Yang Y M., Pedersen J O. A Comparative Study on Feature Selection in Text Categorization. Proceedings of the Fourteenth International Conference on Machine Learning Pages: 412~420.

[9] Eyheramendy S, Lewis D., etc. On the Naive Bayes Model for Text Categorization. In Proceedings of Artificial Intelligence & Statistics 2003. Key West, FL.

[10] 范明, 孟小峰. 数据挖掘-概念与技术, 高等教育出版社 2001.5, 187~207.

李惠娟 女,(1980-),硕士研究生。

高 峰 男,(1967-),教授。研究方向为网络安全,机器学习,电力市场仿真与预测等。

管晓宏 男,(1955-),教授,博士生导师,长江学者特聘教授。研究方向为网络安全、大系统优化理论及其应用。

黄 亮 男,(1981-),硕士研究生。