文章编号:1009-2552(2015)08-0174-05 DOI:10.13274/j.cnki.hdzj.2015.08.046

改进的神经网络算法在垃圾邮件过滤中的应用

王明璐

(天津大学管理与经济学部,天津 300072)

摘 要:针对垃圾邮件威胁信息安全而又屡禁不止的现状,如何从技术上增加对垃圾邮件的控制,维护网络安全成为一个研究的热点问题。人工神经网络具有自适应的特点,在处理变化多端的垃圾邮件问题上有显著优势,但传统算法存在效率低下的问题。现结合模糊理论遗传算法提出了一种改进的BP神经网络算法,在一定程度上提高了算法的效率。通过对中文邮件分类的实验分析,结果表明,本算法的效率优于传统算法,并有较高的识别准确率。

关键词:人工神经网络;垃圾邮件过滤;遗传算法

中图分类号:TP393.098 **文献标识码**:A

Application of an improved neural network algorithm in spam filtering

WANG Ming-lu

(**Department of Management and Economics, Tianjin University, Tianjin** 300072, **China**)

Abstract: Since spam has been increasingly threatening our information security, how to increase the spam control technically so as to maintain network security has become a hot issue in study. With the adaptive feature, artificial neural networks have a significant advantage in dealing with the spam problem which is changing all the time. But the traditional algorithms have the problem of being inefficient. This paper puts forward an improved BP neural network algorithm with the combination of genetic algorithm and fuzzy theory. The efficiency of the algorithm is improved to a certain extent. Through the experimental analysis of Chinese e-mail classification, the results indicate that the efficiency of the proposed algorithm is superior to the traditional algorithm, and has high recognition accuracy.

Key words: artificial neural networks; spam filtering; genetic algorithm

0 引言

随着信息技术的快速发展,在互联网得到普及的同时,大量的病毒、入侵、欺诈、垃圾也随之而来。小到杀毒软件、大到国家机密,信息安全问题正得到越来越多的关注。仅以垃圾邮件为例,据联合国贸易和发展会议报告称,截至2007年底,在世界范围,约57%的邮件都是不请自来的,由此可能造成价值435亿美元的损失。在我国,中国互联网协会反垃圾邮件中心2008年第四季度数据显示,中国互联网用户收到的垃圾邮件数已经占邮件总数的57.89%,已高于世界平均水平。

在这样的背景下,亟需通过最新的技术对不良信息进行提前预防,维护信息安全。数据挖掘(Data Mining)技术是从海量数据中通过一定的处理和方法找寻知识与规律的过程。该技术非常适合处理互联网络中与大量数据有关的问题,数据挖掘算法当中又以神经网络(Neural Network)算法尤为突出。Taeho(2010)[1]从神经网络方法的角度对文本分类问题进行了讨论,通过改进神经网络的算法和使用方式,使文本输入向量的维度大幅下降,维度容忍度超过SVM。Xu和Yu(2010)[2]在垃圾邮件识别问题中介绍了最传统的前馈神经网络(BP网络),在肯定其优势的基础上充分分析了其解释性差、模拟

收稿日期:2014-08-18

作者简介:王明璐(1990-),女,在读硕士,研究方向为数据挖掘、推荐系统。

语言方面有其极限等缺点，并指出通过对垃圾邮件特征词库的调整以及引入解释性好的改进算法可以解决这一问题。Zhang 和 Wang(2009)[3]探讨了能够优化传统分类算法的遗传算法(Genetic Algorithm)，针对中文垃圾邮件的大量输入维度以及算法本身参数的设定都可以使用 GA 方法进行处理。许哲万等(2011)[4]针对 T-S 模糊推理的模糊神经网络进行了改进，并介绍了常见的模糊推理方法：Mamdani 模糊推理、Larsen 模糊推理、Takagi-Sugeno(T-S)模糊推理等。对于模糊逻辑、神经网络以及遗传优化三者的结合，邱兴兴(2007)[5]对其详细的实现过程作出了阐述，其中提到的模糊化处理方式是采用参数尽量少的隶属度函数进行处理，尽可能的提高运行效率。但是，该方法最终采用距离来分类文本，这种做法实则是靠近 KNN 方法，与数据权值关系减弱。熊志斌(2010)[6]克服了这一问题，而且在许多技术细节上做出了自己的改进。最大的缺点就是其设计的模型只能支持 6 维以下的数据输入，不适合垃圾邮件识别。

研究中最突出的不足之处就在于对三层多分类器的实现并未给出明确的方式，而两层的多分类器中又存在着不同细节的技术问题，寻找一种有效的算法组合并作出适当的改进是十分有必要的。本文在数据挖掘技术的基础上，研究人工神经网络算法，加以模糊化，再通过遗传算法优化，从而消除单纯的神经网络算法存在的弊端，克服遗传算法收敛的问题，模仿人脑识别，提高了神经网络算法的效率，最终实现提高邮件系统处理垃圾邮件的效率和正确率。

1 神经网络算法的改进

传统的神经网络在进行训练时处于随机赋予初始权值的状态，因此就很有可能陷入局部极小值状态。遗传算法具有良好的全局搜索性，能够克服神经网络所存在的问题，因此本文采用遗传算法来为神经网络赋予初值。

然而，遗传算法又可能陷入“早熟”问题，即由于缺乏多样性而过早结束优化。所以本文采用多子群遗传算法来保持种群多样性。同时遗传算法需要合适的适应度函数来计算，本文选取自身构建的模糊神经网络作为适应度函数，使用训练样本与期望输出 MSE 的倒数作为输出适应度，在保证误差小的基因适应度大的前提下，将遗传算法与目标神经网络紧密结合。

在整体数据处理时，针对邮件中某些词语既出现在垃圾邮件类，也出现在正常邮件中，因此引入模糊隶属度函数进行二分，每一个权值分为偏上值和偏下值，目标输出为 0 和 1，分别代表垃圾邮件和正常邮件。从而增强了神经网络的解释性，解决了一定程度上的语义问题。

模糊理论也有其本身的问题，基于模糊规则的模糊神经网络是无法处理海量数据的，这受限于模糊规则的数量，其数量越多，占用资源越多，处理效率越低。因此本文采用具有模糊产生器和模糊消除器的模糊系统，即 Mamdani 型模糊系统。此外，传统的加乘型模糊系统会面临“维数灾难”的问题，无法处理高维数据。其中一部分原因在于计算高维输入连乘时很可能导致系统无法计算或溢出而崩溃。本文针对这一问题保持了 BP 神经网络原本的计算方式，取消了连乘的存在，同时不影响函数的逼近，这样就较好地解决了“维数灾难”的问题。

综合以上算法改进的考虑，发现这一多分类器的构建弥补了不同算法本身的缺陷，结合不同算法的流程，构建了如图 1 所示的算法结构，下面对每一层的输入输出进行表述。

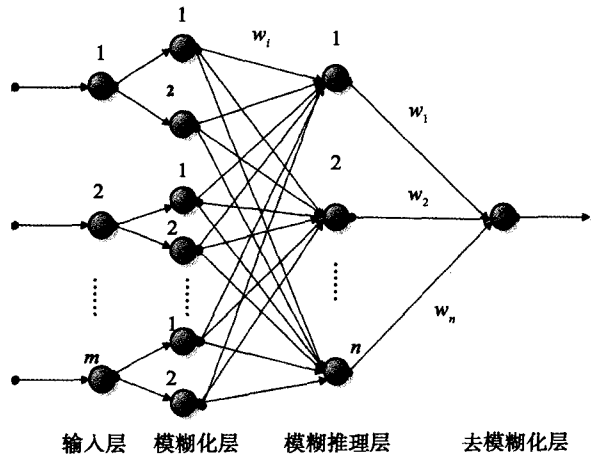

图 1 改进神经网络结构

①第一层(输入层)：这一层有 m 个节点，仅仅是将所有输入值传递到第二层。

②第二层(模糊化层)：这一层接收第一层的数据，并根据模糊隶属度函数 π 函数进行二分，得到输出为 2m 维的向量。

③第三层(模糊推理层)：模糊推理层是以模糊产生器的结果为输入，同时相当于传统 BP 神经网络层次中的隐层，需要对数据首先进行如式(1)处理：

$$O = \sum \omega I + \theta \tag{1}$$

之后将阶段输出值 O 输入正切 sigmoid 函数，得到这一过程的总输出。其中正切 sigmoid 函数是

输入范围为全体实数、输出范围映射在[-1,1]的激活函数。权值和偏置范围均为[-1,1],初始权值由遗传算法输出提供。

此外,这一层的节点数为 n,取值为模糊化层和去模糊化层节点数目乘积平方根在向上取整。公式如下:

$$n = \sqrt{2m \times 1} \tag{2}$$

④第四层(去模糊化层)重复第三层的输出过程,其中激活函数用对数 Sigmoid 函数代替,确保函数输出值为[0,1],符合分类的最终要求。权值和偏置范围均为[-1,1],初始权值由遗传算法输出提供。

以上就是算法计算的流程,综合来看算法的实施克服了神经网络解释性差、遗传算法"早熟"问题以及模糊理论存在的"维数灾难"问题,在数据预处理并赋予初始权值偏置之后,算法会显现出明显的效率提高、准确性增强等特点。

2 邮件过滤模型设计

基于已将改进的算法,本文构建了如图 2 所示的邮件过滤模型。该邮件系统可以分为训练网络阶段和过滤系统测试阶段两部分。下面对这两部分的流程加以介绍。

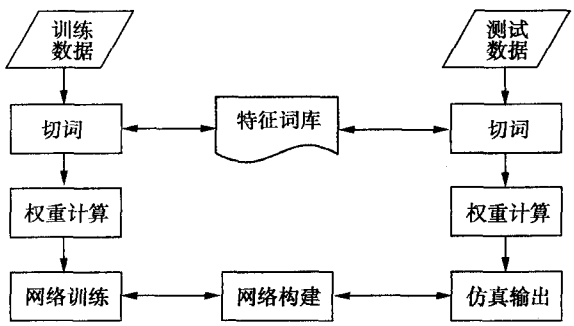

图 2 邮件过滤模型

训练网络阶段流程如下:

①将邮件样本库中的训练数据依据内容进行分词处理,得到训练样本的特征项。

②将分词得到的特征项按照互信息计算和人工审核相结合的方式进行特征提取,选取 400 个词作为特征项。存入特征库中。

③对特征库中的特征项进行权重计算。

④把训练样本集合表示成 400 维的高维向量集合。

⑤对 400 维高维向量进行模糊化处理,每个维度向量获得接近正常邮件和垃圾邮件的不同隶属度值,再合成 800 维的向量。

⑥将部分向量输入到已经建立起来的神经网络中,同时启动遗传算法赋予初始权值,用训练数据测试出最优的初始权值。

⑦在初始权值的基础上对神经网络进行训练,得到全局最小值的训练网络。

过滤系统测试阶段流程如下:

①对测试样本邮件按照内容进行分词处理。

②按照特征库对分词结果进行特征提取。

③按照训练阶段步骤③的方法计算特征项的权重。

④把测试样本集合表示成 400 维的高维向量集合。经过以模糊化处理后,就可以利用挖掘算法过滤垃圾邮件,得到仿真输出。

⑤评价输出结果。

综合来看本文设计的邮件过滤系统与一般的过滤系统存在着一定的不同,其中大部分存在于技术细节。比如对于分词系统的处理,本文对互信息方法的计算结果作出调整。互信息方法只能计算出样本频数相差较大的特征词,却缺乏对频数接近词汇分类的能力,这使得某些对分类有帮助的词被排除在外。同样道理,对于稀有词汇会对分类产生较大作用,本身稀有词汇就可能判断出邮件类别,而当无法判断邮件类别时又会被已知的类别做上标记,之后测试文本中一旦出现,将会导致分类错误。本文针对这样的技术细节加入了人工审核部分,对于词汇出现次数小于 10 的特征词以及词汇频数接近相等的特征词进行了人工判断,这一步骤在输入维数较少时是行之有效的。

还有,特征词在计算权值时本文采取了 TFIDF 法,但是该方法在特征词无法覆盖测试样本时权值无法计算,这就导致了系统存在可能崩溃的隐患。原本在训练阶段计算 TFIDF,不可能存在特征词不出现的情况。然而在测试阶段特征词阶段这种情况确实存在,为了使计算得以继续,必须对这种情况的取值进行补充。取值思路包括 0、1、0.5 以及可以消弭该特征词权值影响的特殊值。由于本文权值计算之后要进行模糊化处理,实际上是二分处理,因此补充权值取 0.5 看似合理,但是这使得原本的函数出现间断点,函数不连续。基于以上考虑本文特征词不出现时权值一律补充 0 值,这也是针对本文算法系统所进行的选择。

3 邮件过滤模型的实现

3.1 数据选取与预处理

本文的数据均为中文邮件,其中共计 402 封邮件,201 封正常邮件,201 封垃圾邮件,编码方式相

同。在数据的分配中，选择 100 封正常邮件、100 封垃圾邮件进行网络训练。再选择剩下的 202 封邮件作为模型的测试文档。

接下来对文档进行分词处理。本文用 ROST 中文词频统计软件采用最大匹配法进行分词匹配，得到了 200 行、801 列的训练样本和 202 行、801 列的测试文本，其中最后一列为目标值列。将这些值输入算法即可实现对垃圾邮件的识别。

3.2 算法实现与仿真输出

本文通过 MATLAB 工具箱中的 GA 工具箱进行了遗传算法的实现，其中适应度函数为所建模型对应的神经网络。根据计算可以判断出神经网络的隐层数目为 28，所需要的权值和偏置值共计 22457 个，本文采用实数编码，因此每个基因有 22457 个变量，依此可以构建完整的遗传算法并进行计算。具体过程中使用了 GA 工具箱，因此最重要的是参数选择，在 workplace 中输入 optimtool 即可进行操作，表 1 为遗传算法中所选择的参数，构建了多子群的遗传算法，剩余参数采用工具箱默认值。

表 1 遗传算法参数设置

初始子种群数	初始子种群规模	基因每点的取值范围	间隔迁移代数	适应度尺度变换函数	变异概率	选择算法	交叉方式	精确值	总代数
4 个	20 个	[-1,1]	20 代	Rank	0.05	Remainder	单点	1e-10	100

构建完遗传算法，将训练数据输入，点击开始获得 22457 维的最优初始输出值。将初始值输入所构建的神经网络，训练神经网络并最终保存训练好的神经网络，该网络不以数值形式存在，只可阅读基本信息，可重复使用。其中神经网络训练的初始参数如表 2 所示。

表 2 神经网络参数设置

学习速度	最大训练次数	目标训练精度
0.05	1000	1e-10

将仿真输出的结果与目标列值进行对比，将 0.5 以下的值确定为垃圾邮件，0.5 及以上的值确定为正常邮件，最终得出结论并评价算法及过滤系统。

3.3 算法改进效果与分析

遗传算法的改进效果可以从图 3 和图 4 中看出。在适应度变化趋势中可以看出，最佳染色体呈现平滑收敛，说明遗传算法搜索的鲁棒性很好。与此同时，在原先设定好的 100 代算法中，仅用了 50 代就达到了稳定的收敛，说明算法的效率也不错。当然，过早收敛可能是“早熟”问题，不过通过遗传算法多样性变化图可以看出，虽然多样性在逐渐降低，但是在 50 代结束的时候依然保持着很高的多样性，这也说明达到收敛时，遗传算法有效地避免了这一问题。此外，系统较早地获得了最优输出值却仍进行着进化，进一步说明算法没有因为“早熟”而停止。

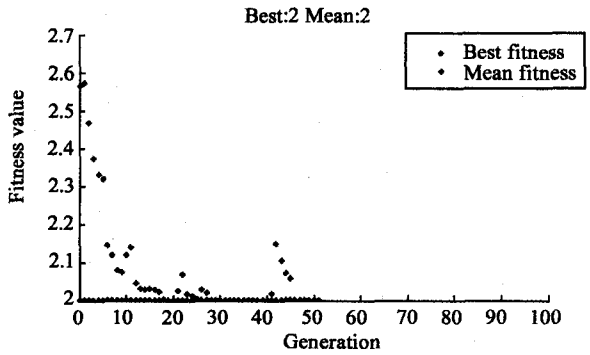

图 3 遗传算法适应度变化图

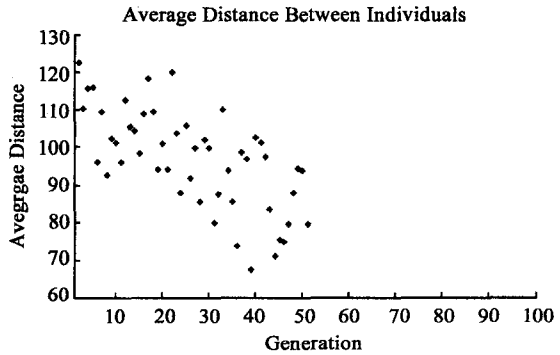

图 4 遗传算法多样性变化图

针对神经网络算法，实际训练效果可以从图 5 和图 6 中看出。神经网络在第 42 次训练时达到了目标精确度 1e-10。在此之前，对同样的数据设定不同的目标精确度得到结果为：当精度为 1e-3，需要训练 7 次；当精度为 1e-7 时，需要训练 35 次。对比最高目标值所需要的训练次数，神经网络训练的效率非常之高，其梯度值的变化则进一步表明这一过程的细节，在训练的后半段梯度平滑且快速下

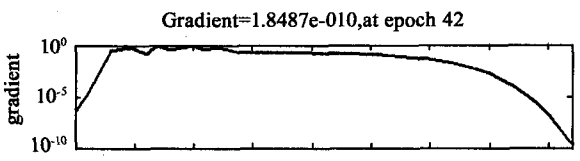

图 5 神经网络梯度值

降。造成这种现象的主要原因是遗传算法已经优化过神经网络的搜索范围，极大的减少了神经网络的训练时间。

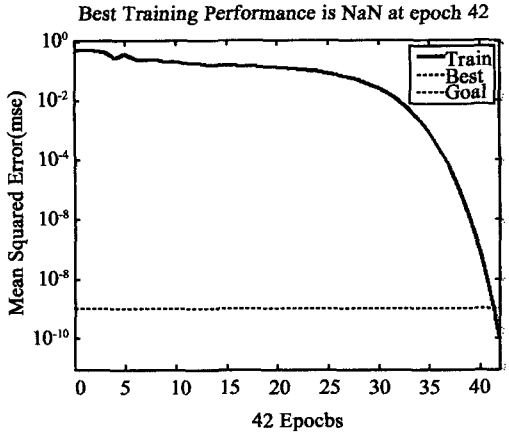

图6 神经网络最佳训练过程

因此，无论是遗传算法还是神经网络都得到了改进。

3.4 分类器结果与分析

多分类器的构建基本已经实现，可以说这种算法在理论上是可用的，在实际上是可实现的。但是是否是可行的还需要对邮件分类结果加以分析。再通过测试数据输入与神经网络仿真输出后，将输出结果目标值进行对比得到如表3所示的结果。

表3 邮件分类结果表

	误判个数	测试总数	误判率	准确率
垃圾邮件	11	101	10.89%	89.11%
正常邮件	1	101	5.94%	99.01%

从邮件分类的结果可以看出，本文设计的邮件过滤系统还是有效地将垃圾邮件和正常邮件分开，准确率高达89.11%和99.01%，这已经可以称得上是有效的系统了，算法、系统均得以实现并且具有一定的现实意义。

4 结束语

本文结合前人在垃圾邮件识别、数据挖掘算法等方面的成果，深入研究了神经网络、遗传算法和模糊理论的结合算法，构造了多分类器并加以实现。通过在实际邮件分类问题上的实验验证，分析结果表明本文提出的算法提高了效率，并保证了分类准确率。

参 考 文 献：

[1] Taeho J. NTC (Neural Text Categorizer): Neural Network for Text Categorization[J]. International Journal of Information Studies, 2010, 2(2): 83-96.

[2] Xu H, Yu B. Automatic thesaurus construction for spam filtering using revised back propagation neural network [J]. Expert Systems with Applications, 2010, 37(1): 18-23.

[3] Zhang Y Q, Wang W. E-mail classification by SVM optimized with genetic algorithm[J]. Journal of Computer Applications, 2009, 29(10): 2755-2757.

[4] 徐哲万，李晶皎，王爱侠，等. 一种基于改进T-S模糊推理的模糊神经网络学习算法[J]. 计算机科学，2011，38(11)：196-219.

[5] 邱兴兴. 基于模糊逻辑和神经网络的文本分类方法[D]. 南昌：南昌大学，2007.

[6] 熊志斌. 基于遗传进化模型模糊神经网络的信用风险评估模型构建及应用[M]. 广州：华南理工大学出版社，2010:14-67.

责任编辑：张荣香

(上接第173页)

到轮廓清晰、受噪声影响小的分割结果。实验结果证明基于Canny边缘水平集分割算法对肝脏区域的分割效果很好，得到了轮廓线准确，信噪比高、信息丰富且不需要人为设置种子点或者阈值区间从而对结果产生影响的目标区域，还能够较好地处理医学图像中常见的解剖结构和拓扑变化。但是该方法也有其显著的缺点：一是在高一维空间上进行曲面演化，因此数据计算量很大，计算复杂性较高且计算时间较长；二是当两个物体存在嵌入时，不能分割出所有物体。因此，还需要更深入的学习和研究来克服这些困难。

总而言之，不存在一个分割算法对所有类型的医学图像都一样能奏效，最有效的方法就是把很多不同的方法组合起来使用，扬长避短，来求得较理想的分割结果。

参 考 文 献：

[1] 李俊. 基于曲线演化的图像分割方法及应用研究[D]. 上海：上海交通大学图书馆，2001.

[2] Ibanze L, Schrader W, Ng L, et al. The ITK Software Guide[M]. The Insight Software Consortium. USA: [s. n], 2005.

[3] 章毓晋. 图象分割[M]. 北京：科学出版社，2001.

[4] Sethian J A. Level Set Methods and Fast Marching Methods[M]. Cambridge University Press, 1996.

[5] 王安明. 基于ITK的医学图像分割[D]. 南昌：南昌大学图书馆，2007.

[6] 高琳，罗晓辉，何立新. 水平集方法在CT肝脏图像分割中的应用[J]. 计算机工程与应用，2005(36)：201-203.

[7] 彭微. 基于区域的肝脏病灶CT图像分割及实现[J]. 信息技术，2011(11)：132-133.

[8] 彭微. 连接门限阈值法在肝脏CT图像上的应用[J]. 咸宁学院学报，2011，31(6)：72-73.

责任编辑：肖滨