文章编号:1001 - 9081(2009)10 - 2755 - 03

利用遗传算法优化的支持向量机垃圾邮件分类

张艳秋,王 蔚
(南京师范大学 教育科学学院,南京 210097)
(ningning605@126.com)

摘 要:提出一种利用遗传算法优化支持向量机来进行垃圾邮件的分类方法。首先对邮件进行预处理,然后利用遗传算法优化支持向量机的惩罚因子和核函数参数的组合,最后利用优化后的支持向量机对邮件进行分类。在由5800 篇邮件构成的数据集上进行实验的结果表明,该方法能达到89.67% 的准确率,提高了对中文垃圾电子邮件过滤的准确性。

关键词:支持向量机;遗传算法;垃圾邮件;参数优化;模式识别
中图分类号: TP391.4 **文献标志码**:A

E-mail classification by SVM optimized with genetic algorithm

ZHANG Yan-qiu, WANG Wei
(*School of Education Science, Nanjing Normal University, Nanjing Jiangsu* 210097, *China*)

Abstract: A method of classifying E-mail by Support Vector Machines (SVM) and Genetic Algorithm (GA) was proposed. In the first step, the mails were preprocessed, and then the combination of support vector machine parameters of C and kernel function parameters was optimized by genetic algorithm. Finally, the E-mail was classified by the optimized SVM. The experiments on a data set composed of 5800 mails show that the precision is 89.67%, which indicates that this method indeed improves the accuracy of filtering Chinese spam.

Key words: Support Vector Machine (SVM); Genetic Algorithm (GA); spam; parameter optimization; pattern recognition

随着信息时代的发展,大量的信息以指数形式增长,为了能够从海量信息中迅速找到我们所需要的信息,就需要对信息进行分类,因此自动文本分类技术应运而生。文本分类是模式识别在文本信息领域的一种应用,其任务是将自然语言文本根据其内容分为预先定义的两类或者多类。文本分类的应用领域极为广泛,垃圾邮件分类就是其中一个很重要的应用。

1 基于支持向量机的垃圾邮件分类

1.1 垃圾邮件分类

垃圾邮件分类又称为垃圾邮件的过滤问题,该问题可以看作一个两类文本分类问题,所要解决的就是将垃圾邮件与非垃圾邮件区分开来。对垃圾邮件的过滤主要有黑白名单、规则过滤和基于概率统计分类等方法。目前的垃圾邮件主要是依据电子邮件的主题和正文中的文本内容进行分类,所以很多文本分类的方法被引进到垃圾邮件分类领域中,并取得了很好的效果。例如朴素贝叶斯分类器、决策树、支持向量机等方法,其中支持向量机方法能够较好地克服样本分布、冗余特征以及过拟合等因素的影响,成为文本分类中公认的较好的方法。垃圾邮件的分类过程如图 1 所示。

在该过程中将所有的样本分为训练样本和测试样本。邮件的预处理包括邮件内容抽取、分词、去除停用词等,然后将分词好的文档用合适的方法表示成分类器可以处理的形式。目前用得比较多的是用空间向量模型对文本进行表示,即将一篇文档转化成向量的形式,每个向量表示该文档的一个特征。假设向量为 $\boldsymbol{x}$,第 i 维的值为 x_i,对应的词为 w_i,本实验采用 TF-IDF(Term Frequency-Inverse Document Frequency) 方法来表示每个特征。$TF(x_i)$ 表示单词 w_i 出现的次数,D 为训练集中所有文档的总数,D_w 为 w_i 曾经出现过的文档数:

$$x_i = \frac{TF(w_i) \times \log\left(\frac{D}{D_w} + a\right)}{\sqrt{\sum_{k=1}^{n} TF^2(w_i) \times \log\left(\frac{D}{D_w} + a\right)}}$$

其中 a 为常数。

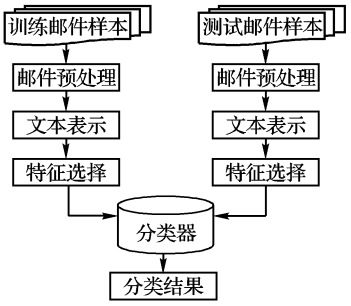

图 1 垃圾邮件分类过程

1.2 支持向量机(SVM)

SVM 是建立在统计学习理论的结构风险最小化原则基础之上的,主要是针对两类分类问题,在高维空间寻找一个超平面作为两类的分割,以保证最小的分类错误率。

支持向量机可以解决三种问题:两类问题、多类问题和回

收稿日期:2009 - 04 - 01。 **基金项目**:全国教育科学"十五"规划教育部重点基金项目(DCA050056);江苏省教育科学"十一五"规划项目(D/2006/01/096);教育部归国留学基金委项目。

作者简介:张艳秋(1984 -),女,山东潍坊人,硕士研究生,主要研究方向:机器学习、模式识别; 王蔚(1966 -),女,安徽凤阳人,教授,博士,主要研究方向:机器学习、模式识别。

归问题。一般用 X 表示输入空间，Y 表示输出域，对于两类问题，$Y = \{-1,1\}$；对于多类问题，$Y = \{1,2,\cdots,m\}$，m 即为 m 个类别；对于回归问题，$Y \in \mathbf{R}$。训练集也称为训练数据，通常表示为：$S = [(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)] \in (X \times Y)^l$。其中 l 为样本数目，x_i 为样例的特征向量值，y_i 为 x_i 的类别标记。

1.2.1 线性支持向量机

以两类文本为例，定义函数 $f(x) = w \cdot x + b$，x 为输入样本特征值，w 和 b 为控制函数的参数，$f(x) = 0$ 定义的超平面将输入空间 x 分为两部分，这两部分对应输入的两类。如图2所示。

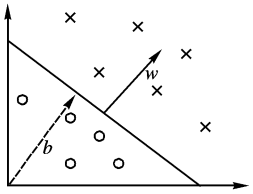

图2 线性支持向量机

1.2.2 非线性支持向量机

线性学习器的计算能力非常有限，而现实世界复杂的应用的目标概念通常不能由给定属性的简单线性组合来产生。支持向量机使用核函数方法将数据映射到高维的特征空间来增加线性学习器的计算能力。该方法首先使用一个非线性映射将数据变换到一个特征空间 F，然后在这个特征空间中使用线性学习器进行分类。定义核函数 K，对所有的 $x,z \in X$，满足：

$K(x,z) = \langle \phi(x) \cdot \phi(z) \rangle$，这里 ϕ 是 X 到(内积)特征空间 F 的映射。有了这个核函数，决策规则就可以对核的 l 次计算来得到：

$$f(x) = \sum_{i=1}^{i=l} a_i y_i K(x_i,x) + b$$

常用的核函数有以下几种：

1) 线性核函数：$K(x,x') = (x \cdot x')$；

2) 径向基核函数：$K(x,x') = \exp(-\gamma \| x - x' \|^2)$；

3) 多项式核函数：$K(x,x') = [(x \cdot x') + c]^d$，其中 $c \geqslant 0$。

由于径向基核函数(Radial Basis Function，RBF)对非线性和高维数据有很好的适应性，所以本文采用径向基核函数。

1.3 支持向量机参数的选择对分类精确度的影响

在本实验中我们选择的支持向量机核函数为径向基核函数(RBF)。支持向量机参数选择对分类的精度影响是很大的，一个是惩罚因子 C，另一个便是 r。为了证明两个参数的重要性，我们进行了如下实验，训练样本为103个，其中垃圾与非垃圾邮件分别为69和34个；测试样本为60个，其中垃圾与非垃圾邮件分别30个，以分类正确率作为评测标准，一组实验中RBF参数不变，惩罚因子变化；第二组中惩罚因子不变，RBF参数进行变化。实验结果如表1所示。

从表1中可以看出，惩罚因子和RBF参数对数据的分类正确率有着很大的影响，且两个参数的变化能使得分类正确率取得一个最优的值。但是当我们取每组实验中最优的惩罚因子和最优的径向基核函数参数作为组合进行实验时，却仅得到61.67%的精度，可见它们并不是最优的组合。基于以上实验分析和讨论，本文试图针对垃圾邮件数据集寻求最优的惩罚因子和核函数的支持向量机参数组合，对垃圾邮件进行分类。以期得到良好的分类正确率。由于遗传算法对优化问题具有全局寻优能力，被广泛用于解决优化问题，因此本文选取遗传算法对支持向量机进行优化。

表1 SVM参数对正确率的影响

惩罚因子	r	数据正确率/%
0.01	1	48.55
0.1	1	50.00
1	1	56.67
10	1	60.00
100	1	66.00
1000	1	58.33
10000	1	56.67
1	0.01	51.23
1	0.1	53.00
1	1	53.66
1	10	60.00
1	100	56.67
1	1000	50.00
100	10	61.67

2 应用遗传算法优化SVM的参数

2.1 遗传算法

由于遗传算法有隐含的并行性和强大的全局搜索能力，可以在短时间内搜索到全局最优点，因此本文采用遗传算法(Genetic Algorithm，GA)来求解。

基本的遗传算法由以下几步组成。

1)编码：通过某种编码机制把对象抽象成由特定符号按照一定的顺序排列成串。

2)适应度函数：遗传算法对一个个体(解)的好坏用适应度函数值来评价，适应度函数值越大，解的质量越好。它的设计应结合求解问题本身的要求而定。

3)遗传算子：包括选择算子、交叉算子、变异算子。遗传算法使用选择运算来实现对群体中的个体进行优胜劣汰操作：适应度高的个体被遗传到下一代群体中的概率大；适应度低的个体，被遗传到下一代群体中的概率小。例如轮盘赌选择方法。交叉算子是指对两个相互配对的染色体依据交叉概率 P_c 按某种方式相互交换其部分基因，从而形成两个新的个体，它在遗传算法中起关键作用，是产生新个体的主要方法。例如单点交叉算子。变异算子是指依据变异概率 P_m 将个体编码串中的某些基因值用其他基因值来替换，从而形成一个新的个体。遗传算法中的变异运算是产生新个体的辅助方法，它决定了遗传算法的局部搜索能力。

4)运行的参数：种群规模(M)、遗传算法终止进化代数(T)、交叉概率(P_c)和变异概率 P_m。

2.2 遗传算法优化SVM参数的过程

由于遗传算法具有隐含的并行性和强大的全局搜索能力，可以在很短的时间内搜索到全局最优点。因此本实验使用GA对SVM进行参数的优化，寻找最优的惩罚因子和RBF参数的组合。结合RBF参数 r 和惩罚因子 C，可以得到需要优化的参数组合：$P = \{C,r\}$，其中 $C > 0$；$r > 0$。我们以SVM的分类识别精度 RA 作为遗传算法的适应度函数：

$$RA = \frac{\text{测试数据集中分类正确的样本数}}{\text{测试数据集样本总数}}$$

对于参数 r 和 C，采用浮点数编码方式，设置终止进化代数为25，交叉概率和变异概率分别为0.1和0.5。

将遗传算法应用到SVM优化中时，算法的基本步骤如下：

1) $T = 0$；

2) 初始化种群 $P(T)$；

3) 计算个体适应度函数的值 $F(T)$;

4) 如果种群的适应度函数足够大,或者 T 达到我们预设的终止代数值(25 代) 则转到 8);

5) $T = T + 1$;

6) 应用选择算子从 $P(T-1)$ 中选择新的 $P(T)$;

7) 对 $P(T)$ 进行交叉变异操作之后转到 3);

8) 得出最佳的惩罚因子 C 与核函数参数的组合,并利用优化后的支持向量机对训练样本进行训练,得到全局最优分类面。

3　实验结果与结论

本实验的实验数据来自 CERNET 共享的 2005 年 6 月份的电子邮件(http://www.ccert.edu.cn/spam/sa/datasets.htm),其中垃圾邮件 3 290 份,非垃圾邮件 2 510 份,将数据集分为五个近似相等的子集,四个用作训练集,一个作为测试集,以分类正确率作为评测标准。为方便比较,我们用台湾林智仁等人开发的 LIBSVM 对相同的数据集进行了实验,LIBSVM 包中有一自带的优化程序 grid.py,同样得到优化的参数组合。但是分类结果与本文提出的 GASVM 相比不甚理想。实验结果如表 2 所示。

表 2　两种分类结果对比

使用的分类器	C	r	分类正确率/%
LIBSVM	2048	0.5	81.67
GASVM	114	0.5	89.67

从表 2 可以看出,利用遗传算法优化的支持向量机取得了较好的分类效果,说明本文提出的 GASVM 算法是可行的。

4　结语

本文利用支持向量机对垃圾邮件进行分类,并且利用遗传算法对支持向量机的参数组合进行了优化,获取了最优的参数组合,从而取得了较好的分类正确率。今后的主要工作集中在:综合优化特征提取部分,使得分类的正确率能够获得更大的提高;考虑平衡数据与非平衡数据对分类的影响,并寻求合适的方法来解决这个问题。

参考文献:

[1] CERVANTES J, LI XIAO-OU, YU WEN. SVM classification for large data sets by considering models of classes distribution[C]// Proceedings of the 2007 Sixth Mexican International Conference on Artificial Intelligence, Special Session. Washington, DC: IEEE Computer Society, 2007: 51 - 60.

[2] NHUNG N P, PHUONG T M. An efficient method for filtering image-based spam[C]// Proceedings of the 2007 IEEE International Conference on Research, Innovation and Vision for the Future. [S.l.]: IEEE Press, 2007: 96 - 102.

[3] KIM D S, NGUYEN H-N. Genetic algorithm to improve SVM based network intrusion detection system[C]// Proceedings of the 19th International Conference on Advanced Information Networking and Applications. Washington, DC: IEEE Computer Society, 2005: 155 - 158.

[4] DRUCKER H, WU DONG-HUI, VAONICK V N. Support vector machines for spam categorization [J]. IEEE Transactions on Neural Networks, 1999, 10(5): 1048 - 54.

[5] VAPNIK V N. An overview of statistical learning theory [J]. IEEE Transactions on Neural Network, 1999, 10(5): 988 - 999.

[6] 刘伍颖, 王挺. 一种多过滤器集成学习垃圾邮件过滤方法[C]// 全国信息检索与内容安全学术会议论文集. 苏州: [出版者不详], 2007.

[7] 李钢, 王蔚, 张胜. 支持向量机在脑电信号分类中的应用[J]. 计算机应用, 2006, 26(6): 1431 - 1433.

[8] 樊兴华, 孙茂松. 一种高性能的两类中文文本分类方法[J]. 计算机学报, 2006, 29(1): 124 - 131.

[9] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报, 2000, 26(1): 32 - 42.

[10] 王清祥, 广凯, 潘金贵. 基于支持向量机的邮件过滤[J]. 计算机科学, 2007, 34(9): 93 - 95.

(上接第 2754 页)

[3] SCHOLKOPF B, BURGER C, VAPNIK V N. Extracting support data for a given task[C]// Proceedings of First International Conference on Knowledge Discovery and DataMining, [S. l.]: AAAI Press, 1995: 262 - 267.

[4] OSUNA E, FREUND R, GIROSI F. An improved training algorithm for support vector machines[C]// Proceedings of the 1997 IEEE Workshop on Neural Networks and Signal Processing. Amelia Island: IEEE Press, 1997: 276 - 285.

[5] PLATT J C. Fast training of support vector machines using sequential minimal optimization[C]// Advances in Kernel Methods: Support Vector Learning. Cambridge: MIT Press, 1999: 185 - 208.

[6] SUYKENS J A K, VANDEWALLE J. Least squares support vector machine classifiers[J]. Neural Processing Letter, 1999, 9(3): 293 - 300.

[7] SUYKENS J A K, LUKAS L, WANDEWALLE J. Sparse approximation using least squares support vector machines[C]// ISCAS: Proceeding of the IEEE International Symposium on Circuits and Systems. [S.l.]: IEEE Press, 2000: 757 - 760.

[8] LI YONG-MIN, GONG SHAO-GANG, SHERRAH J, *et al.* Support vector machine based multi-view face detection and recognition [J]. Image and Vision Computing, 2004, 22(5): 413 - 427.

[9] ROMDHANI S, TORN P, SCHOLKOPF B, *et al.* Efficient face detection by a cascaded support-vector machine expansion[J]. Royal Society of London Series A-Mathematical Physical and Engineering Sciences, 2004, 13(4): 3283 - 3297.

[10] SHIH P C, LIU C J. Face detection using discriminating feature analysis and support vector machine[J]. Pattern Recognition, 2006, 39(2): 260 - 276.

[11] ESPINOZA M, SUYKENS J A K, MOOR B D. Load forecasting using fixed-size least squares support vector machines[C]// Computational Intelligence and Bioinspired Systems, LNCS 3512. Berlin: Springer, 2005: 1018 - 1026.

[12] ESPINOZA M, SUYKENS J A K, MOOR B D. Fixed-size least squares support vector machines: A large scale application in electrical load forecasting[J]. Computational Management Science, 2006, 3(2): 113 - 129.

[13] SUYKENS J A K, LUKAS L, WANDEWALLE J. Sparse approximation using least squares support vector machines [C]// SCAS'2000: Proceeding of the IEEE International Symposium on Circuits and Systems. [S.l.]: IEEE Press, 2000: 757 - 760.

[14] VINGAA S, JONAS S, ALMEIDA J A. Renyi continuous entropy of DNA sequences[J]. Journal of Theoretical Biology, 2004, 231(3): 377 - 388.

[15] SHANNON C E. A mathematical heory of communication[J]. The Bell System Technical Journal, 1948, 27(3): 379 - 423; 623 - 656.