

改进 KNN 算法在垃圾邮件过滤中的应用*

张俊丽 张 帆

(华中师范大学信息管理系 武汉 430079)

【摘要】 提出一种改进的 KNN 算法,并将其用于垃圾邮件的过滤问题。经实验证明,改进的算法能够降低 K 值和训练文本的分布对过滤效果的影响,减少垃圾邮件的误判和漏判,具有较好的过滤性能。

【关键词】 KNN 垃圾邮件过滤 文本分类

【分类号】 TP391

Application of Improved KNN Algorithm in Spam E-mail Filtering

Zhang Junli Zhang Fan

(Department of Information Management, Huazhong Normal University, Wuhan 430079, China)

【Abstract】 In this paper, an improved K-Nearest Neighbor (KNN) is proposed and is applied to filter spam email. It's proved that the improved algorithm is less sensitive to the parameter K and the distribution of the training set, helps reducing the misclassification, and performances well in experiments.

【Keywords】 KNN Anti-spam email Text classification

1 引言

目前,常用的垃圾邮件过滤算法主要有三类:黑白名单过滤法、基于规则的方法和基于统计的方法。其中,黑白名单法是将黑名单地址发出的邮件进行拦阻和过滤,白名单地址发出的邮件判为合法,但在实际应用中,动态变化的邮件地址会导致这种方法失效^[1];基于规则的过滤方法是通过训练得到显式规则,再利用规则来进行过滤,如 Ripper、Decision tree、Boosting 等方法,此类算法的过滤正确率和召回率都在 80% 以上,其缺点是在规律性不明显的邮件中过滤效果比较差^[2];因此更多学者倾向于基于统计算法的研究。KNN(K-Nearest Neighbor)是一种简单的基于统计的过滤算法,Joachims T 和 Li Baoli 指出 KNN 算法是一种很好的分类算法,在不同的数据集上进行实验,都取得了很好的分类效果^[3,4]。Androutsopoulos I 等人将 KNN 应用于邮件过滤中,并与 Bayesian 及基于关键词的过滤算法进行比较,发现前两者过滤效果相当,而基于关键词的过滤算法效果较差^[5],因此,对 KNN 算法进行改进并运用到邮件过滤系统中是很有研究价值的。

经典 KNN 是一种简单的分类算法,由 Cover 和 Hart

提出^[6]。应用于邮件过滤中就是将训练文本分为两类,一类为合法邮件,一类为非法邮件,在训练文本集合中,待测文本找出与其最相似的 K 个文本,然后将其中的多数文本所属的类别赋给待测文本,从而判断出待测邮件是否合法。在经典 KNN 算法中,K 值的选择对分类的结果影响很大,如果 K 值过大,则将会使结果偏向于文本数较多的一类,如果 K 值过小(如 K=1),则会降低过滤效果。本文提出对 KNN 算法进行改进,降低 K 值和训练文本的分布对结果的影响,实验证明,改进后的算法能够提高邮件过滤系统的稳定性。

2 系统设计

邮件过滤系统的框架如图 1 所示。

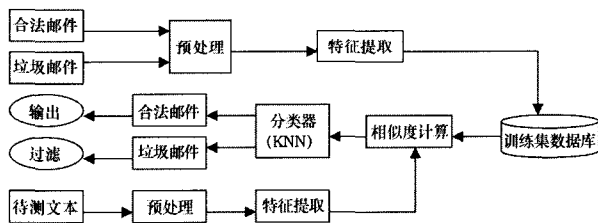


图 1 邮件过滤系统框架图

首先,将训练文本分为合法邮件和垃圾邮件,进行预处理,并提取特征词,将处理结果存入训练集数据库;待测文本经过特征提取后,与训练集数据库中的全部训练文本进行相似性计算,利用 KNN 分类器,将待测文本进

收稿日期: 2007-03-05

收修改稿日期: 2007-03-22

* 本文系 2006 年国家社科基金项目“网络信息过滤研究”(项目编号: 06BTQ024)的研究成果之一。

行分类。若待测文本被判为合法邮件,则系统输出该邮件,否则,系统予以过滤。

3 文本预处理

3.1 文本表示

用向量空间模型表示文本,即文本表示为 (x_1, x_2, K, x_n) ,特征词表示为 (t_1, t_2, K, t_n) ,特征词的权重表示为 (w_1, w_2, K, w_n) 。然后排除停用词,合并数字和人名等词汇,并统计词频,本文采用常见的 TF-IDF^[7] 公式统计词频。

$$W(t, x) = \frac{tf(t, x) \times \log(N/n_t + 0.01)}{\sqrt{\sum_{t \in x} [tf(t, x) \times \log(N/n_t + 0.01)]^2}} \quad (1)$$

其中, $W(t, x)$ 为词 t 在文本 x 中的权重, $tf(t, x)$ 为词 t 在文本 x 中的词频, N 为训练文本的总数, n_t 为文本集中出现 t 的文本数。

3.2 特征提取

对词进行特征项选择,可以降低向量空间的维数,提高程序运行效率。考虑到垃圾邮件所出现的词特征突出(如“赚钱”、“成人”等),过滤时只需考虑这些特征词即可,故在电子邮件过滤系统中,采用互信息进行特征提取效果比较好^[8]。互信息描述的是特征词 t 与类别 c 之间的关联程度。互信息量越大,名词和类别同时出现的概率就越大,因此应该选择互信息大的词作为特征词。其计算公式^[9]如下:

$$MI(t_k) = \sum_{j=1}^2 P(c_j) \log \frac{P(t_k, c_j)}{P(t_k)P(c_j)} \quad (j=1, 2) \quad (2)$$

其中 $P(t_k)$ 为训练集中特征词 t_k 出现的概率, $P(c_j)$ 表示训练集中 c_j 类文本出现的概率, $P(t_k, c_j)$ 表示特征词 t_k 出现在 c_j 类中的概率,设 c_1 为合法邮件, c_2 为垃圾邮件。

3.3 相似度计算

在 KNN 算法中,相似度的选择也很重要,算法的关键就在于找出与其最相似的 K 个文本,本文利用夹角余弦^[10]计算相似度。

$$\text{Sim}(x_i, x_j) = \frac{\sum_{k=1}^m w_{ik} \times w_{jk}}{\sqrt{(\sum_{k=1}^m w_{ik}^2)(\sum_{k=1}^m w_{jk}^2)}} \quad (3)$$

其中, m 为特征向量的维数 k , w_{ik} 表示第 i 个文本的第 k 个特征词的权重值。

4 经典 KNN 算法

该算法的基本思路是:在训练文本集中找出与待测文本距离最近(最相似)的 K 个文本,然后计算新文本属于每类的权重,最后将其分到权重最大的一类中,算法如下:

(1) 在训练文本集中选出与待测文本最相似的 K 个文本。

(2) 依次计算新文本属于每类的权重,根据文献[6],计算公式如下:

$$P(x, c_j) = \sum_{x_i \in KNN} \text{Sim}(x, x_i) y(x_i, c_j) \quad (4)$$

其中, x 为新文本的特征向量, $y(x_i, c_j)$ 为类别属性函数,如果文本 x_i 属于类 c_j ,那么函数值为 1,否则为 0。

(3) 比较权重值,将新文本划分到权重最大的那个类别中。

5 改进的 KNN 算法

5.1 算法描述

在经典 KNN 算法中,一般先设定一个初始 K 值,然后根据实验测试的结果调整 K 的大小,然而,在电子邮件过滤系统中, K 值不能自动调整,而且 K 取值不当或训练文本分布不均会降低过滤性能,影响过滤效果。因此,笔者对 KNN 的权重算法进行改进,使其能适应动态变化的电子邮件过滤系统。改进的 KNN 算法如下:

(1) 利用特征项集合描述训练文本向量。

(2) 用向量表示待测文本,删除停用词汇、数字和其它字符(如 * & 等),利用公式(1)计算词频。

(3) 利用公式(2)提取能表现文本特征的关键词。

(4) 利用公式(3),在训练文本集中选出与待测文本最相似的 K 个文本。

(5) 依次计算新文本属于每类的权重平均值,计算公式如下:

$$P(x, c_1) = \frac{\sum_{x_i \in KNN} \text{Sim}(x, x_i) y(x_i, c_1)}{k_1} \quad (5)$$

$$P(x, c_2) = \frac{\sum_{x_i \in KNN} \text{Sim}(x, x_i) y(x_i, c_2)}{k_2} \quad (6)$$

其中 k_1 为 K 个文档中属于 c_1 类的文本数, k_2 为 K 个文档中属于 c_2 类的文本数, $K = k_1 + k_2$, $P(x, c_1)$ 为文本 x 属于 c_1 类的权重平均值, $P(x, c_2)$ 为文本 x 属于 c_2 类的权重平均值,此算法与向量空间距离分类法有所相似但不相同,后者是利用算术平均为每类文本生成一个代表该类的中心向量,然后计算待测文本与类中心向量间的距离,最后判定新文本属于与其距离最小(最相似)的类。笔者提出的算法是与待测文本最相似的 K 篇文本在每个类中的平均相似度,即删去了一些相似度很低的文本对分类的影响,这样可避免训练集中属于类 c_1 或 c_2 文本过多,而造成前 K 个文本更靠近某一类的倾向,使结果更客观。

(6) 比较权重均值,将文本划分到值最大的那个类别中。

5.2 改进 KNN 算法的主要流程

(1) 找出待测文本的 k 个近邻文本,并存入集合 E 中

初始化:输入训练文本集 $X = \{x_1, x_2, K, x_n\}$,选择常数 k ,输入待测文本 x ,初始化集合 E ,令 $i = 1$ 。

Do until ($i \leq n$)

计算待测文本与训练文本的相似度 $\text{Sim}(x, x_i)$

If $i \leq k$

```
Then 将  $x_i$  放入集合 E
Else if (Sim( $x, x_i$ ) 小于集合 E 中某训练文本与待测文本的距
离)
Then 删去集合 E 中训练文本与待测文本距离最大的文
本, 将  $x_i$  放入集合 E
End If
i = i + 1
End Do Until
```

(2) 对待测文本进行分类

```
For i = 1 to K
    计算  $P(x, c_1)$  和  $P(x, c_2)$ 
End For
If  $P(x, c_1) > P(x, c_2)$ 
    则  $x$  属于  $c_1$  类
Else  $x$  属于  $c_2$  类
End If
```

6 算法评价

垃圾邮件过滤的性能评价通常借用文本分类的相关指标。评估映射准确程度的参照物是人工分类结果(假设人工分类完全正确且排除个人思维差异的因素),测试结果与人工分类结果越相近,分类的准确程度就越高。

假设待测集邮件集合中共有 N 封邮件,垃圾邮件的判定结果如表 1 所示。

表 1 垃圾邮件系统判定情况分布

	实际为垃圾邮件	实际为合法邮件
系统判为垃圾邮件	A	B
系统判为合法邮件	C	D

其中 $N = A + B + C + D$, 则有:

(1) 正确率^[11] (Precision): $P = \frac{A}{A + B}$, 即垃圾邮件检出率。它反应了过滤系统“找对”垃圾邮件的能力, 正确率越大, 将合法邮件误判为垃圾邮件的可能性越小。

(2) 召回率^[12] (Recall): $R = \frac{A}{A + C}$, 即垃圾邮件检出率。这个指标反映了过滤系统发现垃圾邮件的能力, 召回率越高, “漏网”的垃圾邮件就越少。

(3) 代价因子^[13] (Total Cost Ratio): $TCR = \frac{A + C}{\mu B + C}$, 其中 μ 为参数, 表示将合法邮件误判为垃圾邮件的损失为垃圾邮件判为合法邮件的 μ 倍。TCR 越高, 则垃圾邮件过滤系统的损失越低。

以上是垃圾邮件过滤系统性能评价中的三个最重要的指标, 在实际中, 正确率比召回率更重要。

7 实验

本实验的测试语料来源于 Ling - Spam^[14], 它是由希腊学者 Androutsopoulos 等人提供, 由提供者收到的垃圾邮件和来自于语言学家邮件列表 (Linguist List) 的合法邮件构成, 其公用的合法邮件没有加密。语料中含合法邮件 2 412 篇, 垃圾邮件 481 篇, 为了验证训练文本集集的分布对过滤性能的影响, 本文选取合法邮件 50 篇, 垃圾邮件 480 篇, 取其中 4/5 作为训练集, 1/5 为测试集, 改变 K 值进行实验, 并计算其平均值, 实验中 $\mu = 9$, 实验结果如表 2 所示。

表 2 实验结果

K 值	经典 KNN				改进 KNN			
	正确率 (%)	召回率 (%)	代价因子	时间(s)	正确率 (%)	召回率 (%)	代价因子	时间(s)
5	88.13	82.45	2.25	138.5	89.18	82.13	3.92	148.2
8	88.25	83.27	2.12	142.7	92.46	81.96	3.54	149.1
10	85.46	83.78	2.02	145.1	93.83	82.72	4.07	147.6
20	81.81	85.16	1.85	139.2	92.95	82.48	3.83	147.3
25	77.02	86.54	1.88	138.8	93.72	81.56	3.71	148.4
平均	83.13	84.24	2.02	140.8	92.43	82.17	3.82	148.1

通过以上的实验, 可以看出: 随着 K 值的增大, 经典 KNN 正确率减小, 而召回率增大, 这是因为本文实验样本中, 合法邮件数量小于非法邮件, 系统将待测文本分到垃圾邮件类的概率增大, 所以此算法将合法邮件误判为垃圾邮件的可能性大, 而垃圾邮件漏网的少, 改进 KNN 受 K 值变化影响不大。通过比较改进 KNN 和经典 KNN 的值可以发现: 改进 KNN 的正确率、代价因子值均高于经典 KNN, 改进 KNN 能够提高电子邮件的过滤性能, 但是改进的 KNN 同时也增加了时间开销。

8 结 语

本文通过改进 KNN 算法, 从而降低 K 值和训练文本的分布对结果的影响, 实验证明, 改进后的算法能够提高垃圾邮件的过滤效果。如何提高算法运行速度, 并使其适应适时性要求较高的电子邮件过滤系统, 还需要进一步研究和探讨。

参考文献:

1 张帆. 信息组织学. 北京: 科学出版社, 2005: 411 - 412
2 王斌, 潘文锋. 基于内容的垃圾邮件过滤技术综述. 中文信息学报, 2005, 19(5): 4 - 5
3 Joachims T. Text Categorization with Support Vector Machines; Learning with Many Relevant Features. European Conference on Machine

- Learning, 1998
- 4 Li Baoli, Chen Yuzhong, Yu Shiwen. A Comparative Study on Automatic Categorization Methods for Chinese Search Engine. In: Proceedings of the Eighth Joint International Computer Conference, 2002; 117 - 120
 - 5 Androutsopoulos I, Koutsias J, Chandrinos K V, Spyropoulos C D. An Experimental Comparison of Naive Bayesian and Keyword - Based Anti - Spam Filtering with Encrypted Personal E - mail Messages. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2000; 160 - 167
 - 6 Cover T M, Hart P E. Nearest Neighbor Pattern Classification. IEEE Trans. Inform. Theory, 1967(13); 23
 - 7 Salton G, Wong A, Yang C S. A Vector Model for Automatic Indexing. Communication of ACM, 1975, 18(11); 613 - 620
 - 8 Sahami M, Dumais S, Heckerman D, Horvitz E. A Bayesian Approach to Filtering Junk E - Mail. AAAI Technical Report, 1998(5); 55 - 62
 - 9 Mitchell T M. Machine Learning. New York: McGraw - Hill, 1997
 - 10 Salton G, McGill M J. Introduction to Modern Information Retrieval. McGraw Hill, Computer Series, 1983
 - 11 徐洪伟, 方勇, 音春. 垃圾邮件过滤技术分析. 通信技术, 2003, 142(10); 127
 - 12 Georgios Sakkis, Ion Androutsopoulos. Stacking Classifiers for Anti - Spam Filtering of Email. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2001; 44 - 50
 - 13 Androutsopoulos I, Koutsias J, Chandrinos K V, Paliouras P, Spyropoulos C D. An Evaluation of Na? ve Bayesian Anti - Spam Filtering. In: Proceedings of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning. 2000; 9 - 17
 - 14 The Linguist List. <http://listserv.linguistlist.org/archives/linguist.html>. (Accessed Dec. 20, 2006)
- (作者 E - mail: elili62@126.com)



e - Learning 项目——BlendEd 成果预发布

BlendEd 项目是苏格兰基金管理委员会资助的一个合作式 e - Learning 改革项目, 将于 2007 年春完成。该项目由英国瑞德克学院发起, 项目组成员包括 6 所进修大学, 大学开放学习交换组 (Colleges Open Learning Exchange Group, COLEG) 和 JISC 苏格兰西南区域支持中心 (JISC Regional Support Centre Scotland S&W, JISC RSC Scotland S&W)。

BlendEd 的目标是在大学专科社会公益和商业管理两个专业中引进一个复合学习传输模块。这种方法的本质是为现在的高等教育机构提供基于主题的电子资源, 支持教育者以一种新的、灵活的方式将这些资源传递给学习者。该项目的主要成果包括:

(1) 一个完全测试好的复合学习模块和实施计划;

(2) 有高质量的资源去支持复合学习传输模块;

(3) 强有力的员工开发程序以支持部门采纳该模块;

(4) 有实施复合学习的指南;

(5) 教育者角色新转变——成为复合学习技术专家 (Blended Learning Technologist, BLT)。

该项目初期试验效果不错, 得到了学习者及教学人员的一致好评。很多合作伙伴都任命了新的复合学习技术专家支持该模块的运行, 目前, 该项目正在对 500 多个学习对象进行最后编辑。

(编译自: The BlendEd Project Nears Completion. <http://www.dlib.org/dlib/december06/12inbrief.html#DYET>. [2007 - 1 - 3])

(本刊讯)