文章编号:1001 - 9081(2006)08 - 1980 - 03

一种有效的垃圾邮件过滤新方法

林 琛,李弼程
(信息工程大学 信息工程学院,河南 郑州 450002)
(linchen_ai@163.com)

摘 要:受到信息粒度原理的启发,给出了一种有效的垃圾邮件过滤新方法。该方法训练过程是将训练样本集合中合法邮件类和垃圾邮件类拆分成四个小类,得到四个小类的类中心向量,从粒度原理角度来看,就是采用更细的粒度来描述训练样本的先验知识。过滤过程则将新进来的邮件分别与四个小类的类中心向量进行相似度比较,最终来判定所属类别。在公共垃圾邮件语料库上测试新方法,同时与目前过滤性能较高的 KNN 方法进行比较,结果显示新方法具有过滤精度高,过滤速度快等优点。

关键词:垃圾邮件过滤;粒度;KNN
中图分类号: TP393.098 **文献标识码**:A

New effective method for spam filtering

LIN Chen, LI Bi-cheng
(*College of Information Engineering, Information Engineering University, Zhengzhou Henan* 450002, *China*)

Abstract: A new effective method for spam filtering according to the principle of granularity was presented. First, this method divided spam class and legit class in train corpus into four small classes, and four center vectors were obtained. In the view of the principle of granularity, smaller granularity is used to describe knowledge in train corpus. When filtering, the new E-mail was compared with four center vectors respectively to decide which class it belonged to. This method was tested on spam corpus and compared with KNN. The results show that the new method has some advantages including high accuracy, high speed of filtering and so on.

Key words: spam filtering; granularity; KNN

0 引言

当前垃圾邮件过滤技术发展的主要趋势是将文本分类方法应用于垃圾邮件过滤中。文本分类方法大体可以分为三种类型:判别分析法,机器学习方法和人工神经网络方法。如判别分析中的贝叶斯判别法[1,2],支持向量机[3],Winnow 方法[4],最近邻判别法[1,5]和 Rocchio 方法[6]等;机器学习方法中的决策树算法[7],基于 Rough Set 理论的机器学习方法[8]和 Boosting 方法[9]等。目前公共垃圾邮件语料库包含合法邮件和垃圾邮件两个大类,为了算法性能具有比较性,绝大部分算法的测试与算法间的比较都是在这些公共垃圾邮件语料库进行的,所以垃圾邮件过滤一般都按二分类问题进行处理。

垃圾邮件过滤的结果包含四种情况,即合法邮件判为合法邮件,垃圾邮件判为垃圾邮件,合法邮件判为垃圾邮件和垃圾邮件判为合法邮件。据此特点和信息粒度原理思想的启发,通过一种称为伪聚类的方法对训练样本集合进行重新划分,把合法邮件类和垃圾邮件类划分为四个小类:强合法邮件类,弱合法邮件类,强垃圾邮件类和弱垃圾邮件类,分别得到四个小类的类中心向量。当过滤新邮件时,分别计算新邮件与四个类中心向量的相似度,如果与之相似度最大的类为弱合法邮件类或强合法邮件类时,则该邮件属于合法邮件,反之为垃圾邮件。方法实质是将垃圾邮件过滤转化为四分类问题进行处理。

1 信息粒度相关知识

信息粒度是对信息和知识细化不同层次的度量。信息粒度的形式化描述为:一个问题可以用一个三元组 (X,F,τ) 来描述,X 表示问题的论域,即考虑的所有基本元素的集合;F 表示属性函数,定义为 $F:X\rightarrow Y$;Y 表示基本元素的属性集合;τ 表示论域的结构,定义为论域中各个基本元素之间的关系。

从一个较“粗”的角度来看,实际上就是对 X 进行简化,把性质相近的元素归为一类,整体作为一个新的元素,这样就形成了一个粒度较大的论域 $[X]$,从而把原问题转化为新层次上的问题。同样从一个较“细”的角度看,就是对 X 进行一种划分。因为粗粒度世界比原来的问题要简单,所以一般能够缩小问题求解的范围,加快求解的速度。当然,经过简化的粗粒度世界会造成信息的损失,这时把粒度适当的减小,使得论域个体之间的可区别性增大。

文献[10]认为对于一些元素,由于在大粒度世界不容易看出它们之间的区别,所以容易造成和其他类元素的混淆,只有适当的采用较小的一些粒度,即将它们转换到另一较细的粒度空间,才能够准确的对它们加以区分。

2 伪聚类

从信息粒度原理角度,我们分析造成邮件错分的主要原

收稿日期:2006 - 02 - 13;**修订日期**:2006 - 05 - 09 **基金项目**:河南省教育厅基金资助项目(sp200303099)

作者简介:林琛(1981 -),女,山东威海人,助理工程师,硕士研究生,主要研究方向:海量信息检索; 李弼程(1970 -),男,湖南衡阳人,教授,博士,主要研究方向:数据融合、海量信息处理.

因是:当将垃圾邮件过滤视为二分类时,相当于在一个较粗的粒度空间下描述训练样本集合,在这个粒度空间下,训练样本集合中的一些邮件与其相反类别之间的区别不是很明显,容易造成和相反类邮件的混淆,当一封新邮件为区别不明显邮件时,可能就被错分到相反类中。我们称那些与相反类区别不明显的邮件为弱邮件,相对应的具有很强的类区别性的邮件称为强邮件,可见训练样本集合中实际上是隐藏着弱合法邮件类,弱垃圾邮件类,强合法邮件类和强垃圾邮件类四个小类(如图 1 所示)。

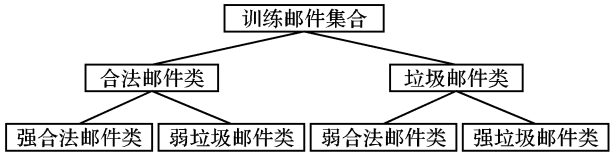

图 1

为了减少错分邮件的数目,需要在更细的粒度空间下描述训练样本集合。本文方法是将弱邮件单独提取出来,形成两种弱邮件类,两个大类被拆分成四个小类,用四个小类来描述训练样本集合的先验知识。下面采用了一种类似于 k 均值聚类的方法对训练样本集合进行拆分,这个过程不同于真正意义上的聚类,我们称之为伪聚类。

对训练样本集合进行伪聚类之前,首先标注每个训练样本的类别标号:合法邮件类和垃圾邮件类,然后使用伪聚类方法对训练样本集合进行划分。伪聚类的过程也是训练过滤器的过程,方法如下:

(1)分别从训练样本集合的合法邮件样本和垃圾邮件样本中随机选择一个作为初始样本,设为四个小类中的两个类;

(2)从剩余样本中选择一个样本分别与两个初始样本进行相似度比较,会出现以下四种情况:

如果该样本与合法邮件初始样本相似度大于与垃圾邮件初始样本的相似度,且该样本类别标号为合法邮件类,则把它归到强合法邮件类;

如果该样本与合法邮件初始样本相似度小于与垃圾邮件初始样本的相似度,且该样本类别标号为垃圾邮件类,则把它归到强垃圾邮件类;

如果该样本与合法邮件初始样本相似度大于与垃圾邮件初始样本的相似度,但该样本类别标号为垃圾邮件类,则把它归到弱垃圾邮;

如果该样本与合法邮件初始样本相似度小于与垃圾邮件初始样本的相似度,但该样本类别标号为合法邮件类,则把它归到弱合法邮件类。

(3)计算每个小类的类中心向量作为新的一轮相似度比较的初始样本向量;

(4)从剩余样本中再选择一个样本分别与四个小类的中心向量进行相似度比较,把它归到与其相似度最大的那个类,转到(3)直至所有训练样本全部聚成四个小类。其中,样本之间的相似度测量采用的是夹角余弦相似度函数。

在经过伪聚类后,原始的训练样本集合被重新划分,实现了在一个更细的粒度空间描述训练样本集合。

3　实验部分

3.1　实验相关参数

实验采用信息增益[1,4,5]来进行特征选择,选择特征集合中信息增益值最大的前 k 个特征形成最优特征集合,采用经典的 TF * IDF 加权方法对特征进行加权。为了表现不同特征数目对过滤算法的性能的影响,实验中抽取不同大小的最优特征集合,大小分别为 400 到 1000 间隔 200,1500 到 3000 间隔 500,4000 到 10000 间隔 1000。

3.2　实验方法及评价指标

实验中采用由希腊 Androutsopoulo[1] 提供的 Ling_Spam 公共测试语料库,它包含了 2 389 封邮件,481 封垃圾邮件,2412封合法邮件,共分为 10 份。采用十次交叉验证方法对过滤算法测试,即将 10 份语料子集中的 9 份作为训练集,另一份作为测试集,结果取 10 次测试结果的平均值。一般常用下列三个指标来衡量垃圾邮件过滤算法的性能,其定义如下:

$$\text{召回率:}SR = \frac{S \to S}{S \to S + S \to L}$$

$$\text{正确率:}SP = \frac{S \to S}{S \to S + L \to S}$$

$$\text{精确率:}Acc = \frac{L \to L + S \to S}{N_L + N_S}$$

召回率反映过滤垃圾邮件的能力;正确率又称为垃圾邮件的检对率,值越大说明合法邮件误判为垃圾邮件越少;精确率是指所有邮件的检对率,它的值越大说明邮件误判率越小,也就是被错分的邮件数目越少。公式中,$S \to S$ 是被正确地分到垃圾邮件数目;$L \to S$,$S \to L$ 是被误分进垃圾邮件和合法邮件的邮件数目;N_L 是指所有的合法邮件类的个数;N_S 是指所有垃圾邮件的个数。

3.3　实验结果及分析

图 2 是两种算法在不同特征个数下的过滤精确度的比较曲线,图 3 是二者在不同特征个数下的垃圾邮件召回率和正确率的比较曲线(info 表示本文算法)。

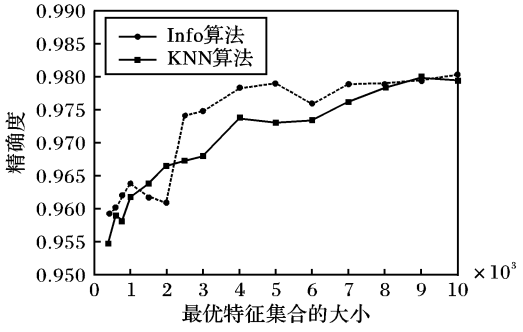

图 2　不同特征个数下的过滤精确度比较

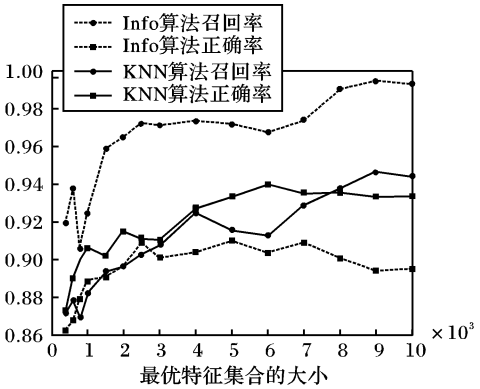

图 3　不同特征个数下的垃圾邮件召回率和正确率比较

图 2 中过滤精度和图 3 中的垃圾邮件正确率明显优于 KNN 方法,验证了 info 算法能够减少邮件错分个数,理论与实验结果相统一。尽管 info 算法中垃圾邮件的召回率略低于 KNN,但是对于实际的垃圾邮件过滤而言是可以忍受的。从算法的复杂度方面来看,KNN 没有训练过程,过滤时直接将过滤邮件与训练集合中的每个邮件进行比较,然后根据前 k

封相似邮件得到新邮件的类别,原理简单,但是过滤的速度很慢,不适用于过滤速度要求较高的场合。而 info 算法相对于 KNN 来说虽然需要训练时间,但是训练过程时间比较短,而且训练的过滤器(只包含了四个小类的中心向量)所需存储空间小,过滤时新邮件只与四个中心向量进行相似度比较,过滤速度大大的快于 KNN。

4 结语

提出了一种新的垃圾邮件过滤方法,比较实验结果显示它具有过滤精度高和过滤速度快等特点,可以应用在对这两方面要求较高的过滤场合。根据方法特点下一步工作主要有:在算法改进上,由于伪聚类方法中形成的类中心向量相当程度上模糊了类别特征,所以需要寻找一种能更好地表现类特征的向量,比如:重心向量等;此外,还要测试文本预处理,特征选择方法等方面对 info 算法性能的影响等。

参考文献:

[1] ANDROUTSOPOULOS I, PALIOURAS G, KARKALETSIS V. Learning to filter spam E-mail: A comparison of a naive bayesian and a memory-based approach[A]. Proceedings of the workshop: Machine Learning and Textual Information Access[C]. 2000. 1 - 13.

[2] SAHAMI M, DUMAIS S, HECKEMAN D, *et al*. A bayesian approach to filtering junk E-mail[A]. Learning for Text Categorization-Papers from the AAAI Workshop[C]. 1998. 56 - 62.

[3] COHEN WW. Learning rules that classify e-mail[A]. Proceedings of AAAI Spring Symposium on Machine Learning in Information Access [C]. 1996. 18 - 25.

[4] 潘文峰. 基于内容的垃圾邮件过滤研究[D]. 中国科学院计算技术研究所硕士毕业论文, 2004.

[5] SAKKIS G, ANDROUTSDOPOULOS I, PALIOURAS G, *et al*. A memory-based approach to anti-spam filtering for mailing list[J]. Kluwer Academic Publishers, Information Retrieval, 2003, 6(1): 49 - 73.

[6] DRUCKER H, WU D, VAPNIK VN. Support vector machines for spam categorization[J]. IEEI Transactions on Neural Networks, 1999, 20(5): 1048 - 1054.

[7] CARRERAS X, MÁRQUEZ L. Boosting trees for anti-spam E-mail filtering[A]. Proceedings of 4th Int' l Conference on Recent Advances in Natural Language Processing[C]. 2001. 58 - 64.

[8] 刘洋, 杜孝平, 罗平等. 垃圾邮件的智能分析, 过滤及 Rough 集讨论[A]. 第十二届中国计算机学会网络与数据通信学术会议[C]. 武汉, 2002.

[9] NICHOLAS T. Using AdaBoost and Decision Stumps to Identify Spam E-mail[EB/OL]. Stanford University Course Project (Spring 2002/2003) Report, http :// nlp. stanford. edu/courses/cs224n/2003. fp, 2003.

[10] 卜东波. 聚类/分类理论研究及其在文本挖掘中的应用[D]. 中国科学院计算所博士学位论文, 2000.