基于改进平衡 Winnow 算法的短信过滤系统

闫红静[1],邸书灵[2]

(1.石家庄铁道大学 研究生学院,河北 石家庄 050043;
2.石家庄铁道大学 信息科学与技术学院,河北 石家庄 050043)

摘 要:将黑白名单技术与 Balanced Winnow 算法相结合,实现对垃圾短信的过滤。采用 CHI 特征提取算法并对权重计算方法进行改进,同时提出了去除训练样本中野点的想法,通过判定去除野点,减缓在训练过程中出现的抖动现象。实验表明这种改进对于提高训练速度及提高短信过滤的性能均有很好的作用。

关键词:balanced Winnow;短信过滤;CHI;野点

中图分类号:TP393 **文献标识码**:A **文章编号**:1674-7720(2010)19-0010-03

Short message filtering system based on improved balanced Winnow

YAN Hong Jing[1], DI Shu Ling[2]

(1.Department of Graduate, Shijiazhuang Railway University, Shijiazhuang 050043, China;
2.Department of Computer Informatics, Shijiazhuang Railway University, Shijiazhuang 050043, China)

Abstract: Spam short message (SMS) had been a big problem that disturbed cell-phone users and mobile operators recently. Combined black-white list technology and balanced Winnow algorithms to realize the aim of filtering spam short message, adopts CHI statistics feature selection algorithms in addition improve the calculating methods of feature's weight, meanwhile, put forward view of casting aside the outliers in training set, estimate and take out the outliers to reduce the jittery action at the training. Finally, experiments show that this amelioration improves the speed of training and profits performance of filtering.

Key words: balanced Winnow; short message filter; CHI; outliers

手机短信以其短小、迅速、简便、价格低廉等优点成为一种重要的通信和交流方式,受到众多人士的青睐。然而,手机短信与邮件一样存在着垃圾信息问题。

目前,垃圾短信过滤主要有黑名单过滤、关键词过滤和基于文本分类的内容过滤等方式。黑名单过滤和关键词过滤方式能快速过滤垃圾短信,但这两种过滤方式实质是基于规则的过滤,虽然在一定程度上阻挡了一些垃圾短信,但规则的方法需要更多的用户自定义设置,很容易被反过滤。基于文本分类的短信过滤采用常见的分类算法,如朴素贝叶斯、SVM、神经网络等。黎路[1]等人将贝叶斯分类应用到 J2ME 模拟环境中成功地过滤了中奖短信和祝福短信。浙江大学的金展、范晶等[2]将朴素贝叶斯和支持向量机结合,解决了传统垃圾短信过滤系统短信特征和内容未能得到及时更新而导致过滤性能降低的问题。王忠军[3]将基于朴素贝叶斯短信过滤算法与基于最小风险贝叶斯算法进行了实验分析和比较,结论是基于最小风险的短信过滤算法具有较好的性能。然而,短信过滤的准确率依赖于其训练样本的数量及质量,这些分类算法需要经过训练学习建立分类器模型,因此在速度上不能很好地满足短信过滤实时性的要求。从现有技术上来说,垃圾短信的过滤在准确率和效率方面仍然不能满足现实需要。

本文针对现有短信过滤技术的不足,设计了在手机终端的短信过滤系统,根据垃圾短信的特点将黑白名单和基于内容过滤相结合。这种过滤方式要求能够快速地对短信进行分类,并且能够实现用户对短信过滤的个性化要求,使垃圾短信过滤系统具有更好的过滤性能。

Winnow 算法是在 1987 年由 Nick Littlestone 提出并对可行性做了严格证明的线性分类算法[4]。当时的目标是想找到一种时空复杂度仅仅与分类对象相关属性相

关的数量呈线性相关的算法。平衡 Winnow 算法是对基本 Winnow 算法的一种改进，该算法具有过滤速度快、性能好、支持反馈更新的优点，在信息过滤领域有很好的应用前景，尤其适合于对实时性要求较高的短信过滤系统。

本文设计并实现了一个基于平衡 Winnow 算法的短信内容过滤系统，对该算法在短信过滤系统上的应用进行了详细分析。分类器的训练过程分成预处理、训练、分类和反馈四个部分。

1 预处理模块

预处理模块包括中文分词、特征提取以及短信的向量表示子模块。

1.1 中文分词

中文分词是汉语所特有的研究课题。英语、法语等印欧语种词与词之间存在着自然的分割，一般不存在分词的问题。本系统采用了目前国内较多使用的中科院计算所开发的汉语词法分析系统 ICTCLAS[5](Institute of Computing Technology, Chinese Lexical Analysis System)。ICTCLAS 3.0 分词速度单机 996 Kb/s，分词精度 98.45%，API 不超过 200 KB，各种词典数据压缩后不到 3 MB，是当前相对较好的汉语词法分析器。

1.2 特征提取

特征提取的方法目前也有很多，常用的特征选取方法有[5]：文档频率 DF(Document Frequency)、信息增益 IG(Information Gain)、互信息 MI(Mutual Information)、χ^2 统计等。

本文将分词后的词作为候选特征，然后使用特征提取算法从中提取出对分类最有用的一些特征，去除对分类贡献不大的候选特征，以降低特征的维数。其中 χ^2 的主要思想是认为词条与类别之间符合 χ^2 分布。χ^2 统计量的值越高，特征项和类别之间的独立性越小、相关性越强，即特征项对此类别的贡献越大。χ^2 是一个归一化的值，该方法比其他方法能减少 50%左右的词汇，具有分类效果好的优点[6]。本文中采用 χ^2 统计进行特征提取。但不是简单地令特征项的权重 $x_i=1$ 或 0，而是令 $x_i=f(\chi^2)$ 或 0，这里 χ^2 特指特征对应的 χ^2 统计值，对应关系 f 根据实际情况而定。实验中令 $x_i=\sqrt[n]{\chi^2}$（n 是一个正整数，取 $n=4$）。实验表明比用布尔权重表示效果要好。

1.3 文本向量表示

目前应用较多的是向量空间模型 VSM(Vector Space Model)，文中用 VSM 将一条短信表示为 $(W_1, W_2, \cdots, W_k, \cdots, W_n)$ 的向量形式。其中：$W_k(k=1, 2, \cdots, n)$ 为第 k 个特征的权重，n 为选定的特征数。

2 构造分类器

训练分类器是研究的重点，采用 Balanced Winnow 算法并对其进行改进。

2.1 Winnow 分类算法

Winnow 算法是二值属性数据集上的线性分类算法。线性分类问题中表示分类界限的超平面等式如下：$w_0\alpha_0+w_1\alpha_1+w_2\alpha_2+\cdots+w_k\alpha_k=0$，其中：$\alpha_0, \alpha_1, \cdots, \alpha_k$ 分别是属性的值；$w_0, w_1, \cdots, w_k$ 是超平面的权值。如果其值大于 0，则预测为第一类否则为第二类。

Winnow 算法是错误驱动型的分类算法，即当出现错分的实例时才更新权值向量。设定两个学习系数 α 和 β（其中 $\alpha>1$，$\beta<1$），通过将权值乘以参数 α（或 β）来分别修改权值。

2.2 Balanced Winnow 分类算法

标准的 Winnow 算法不允许有负的权值，于是就有了另一个称为平衡的 Winnow 版本，允许使用负的权值。对 Winnow 算法的基本形式，权重向量的每一维都是正数。Balanced Winnow 是用 w^+-w^- 代替 w，当 $\sum_{j=1}^{d}(w_j^+-w_j^-)x_j>\theta$，则将实例归为该类。Balanced Winnow 的权重更新策略为：

(1)如果 $\sum_{j=1}^{d}(w_j^+-w_j^-)x_j>\theta$，但文本不属于该类，则要降低权重：对 $j=1, 2, \cdots, d$ 如果 $x_j\neq 0$，则 $w_j^+=\beta w_j^+$，$w_j^-=\alpha w_j^-$，$\alpha>1$，$0<\beta<1$。

(2)如果 $\sum_{j=1}^{d}(w_j^+-w_j^-)x_j<\theta$，但文本应属于该类，则要提高权重：对 $j=1, 2, \cdots, d$，如果 $x_j\neq 0$，则 $w_j^+=\alpha w_j^+$，$w_j^-=\beta w_j^-$，$\alpha>1$，$0<\beta<1$。

在实验中，采用文献[7]中统一 α 和 β 为一个参数的方法，令 $\beta=1/\alpha$，没有影响分类效果，但有效简化了参数的选择。可以为不同的类别确定不同的 θ 值，但实验表明：对于不同的类别选择同样的 θ 值，结果几乎是一样的，所以在每次独立的实验中都取相同的 θ 值，大小是训练文本所含的平均特征数，而初始的 w^+ 和 w^- 分别取全 2 和全 1 向量。

在平衡 Winnow 算法中，一旦参数 α、β 和阈值 θ 确定下来后，将在训练过程中不断更新权重向量 w^+ 和 w^- 至最适合这组参数。因此对参数的依赖较小，需要手工调整的参数不多。

2.3 去除野点

在短信过滤中，短信样本是由手动或自动方式收集的，收集的过程中难免会出错，因此短信样本集中可能存在一些被人为错分的样本点，即野点。这些野点在训练时，会使得分类器产生严重的抖动现象，降低分类器的性能。因此，好的分类器应具有识别野点的能力[11]。

对于 Winnow 算法，若样本中存在野点，则野点在训练时以较大的概率出现在两分类线之外，且分类错误。这些野点对分类器的训练过程产生很大的影响，可能会造成分类器的“过度学习”。因此引入损失函数，按照损失函数的定义，这些野点损失较大，因此可以通过给损失函数设置一个上界函数来处理线性分类器中的野点问题，如图 1 所示。

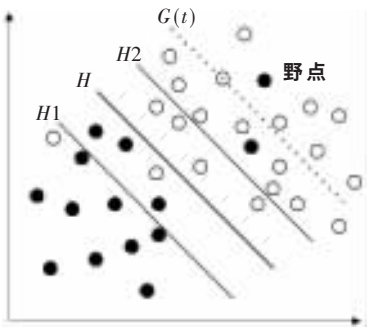

图 1 上界函数示意图

图 1 所示为两类线性可分情况，图中实心点和空心点分别表示两类训练样本，H 为两类样本没有被错误地分开的分类线，$H1$ 和 $H2$ 分别为平行于分类线 H 且与分类线 H 的距离为单位距离的两条直线。直线 $G(t)$为平衡 Winnow 算法中第 t 轮迭代后损失函数的上界线。该上界线是关于迭代次数 t 的函数，因此可以将该上界线 $G(t)$对应的上界函数记为 $g(t)$。从图 1 可知，在直线 $G(t)$ 左下侧误分样本的损失较少，可以认为这些误分样本是由于当前分类器的性能较低而误分的；在直线 $G(t)$右上侧误分的样本由于在第 t 轮迭代后损失仍较大，则可以认为这些误分的样本是野点。根据线性分类器和野点的性质可知，上界函数 $g(t)$具有以下性质：

(1)随着 Winnow 算法中迭代次数 t 的增加，上界函数 $g(t)$单调递减，并且递减的速率也随着 t 的增加而递减，即上界函数的导数 $g(t)$为单调递减函数；

(2)上界函数既不能太大，也不能太小。太大会降低判断野点的能力，太小则会误判正常样本为野点。

根据上界函数的这些特性，可以考虑一个平行于分类线 H 的线性函数作为损失函数的上界函数。即 $g(t)=\sum_{j=1}^{d}(w_j^{+}-w_j^{-})x_j-\theta+\varepsilon-\eta t$。其中，$\varepsilon$ 为常数值；直线 $G(t)$平行于分类线 H；η 为损失因子，也称为学习率，可以在训练分类器的时候指定其值。

在每一轮训练中，若该样本的 $G(t)$值大于分类线的值，并且超过一定的阈值，且不属于该类，则判定该样本具有野点的性质，应当在训练集中将该样本去除，以便提高下一轮训练的准确性。这样不仅有效削弱了分类器的抖动现象，而且提高了分类器的性能。

3 系统反馈

Winnow 是一种在线学习的、以错误为驱动的分类器，适于结合增量式学习来解决自适应问题，实现用户的个性化要求。平衡 Winnow 算法是基本 Winnow 算法的另外一种形式，同样具有在线更新能力。在分类器训练过程中，对错分的短信通过 α 和 β 更新类别权重向量，实现对分类器的更新，平衡 Winnow 算法中 w^{+}和 w^{-}的双向调节，使算法的训练速度更快，适合于对分类实时性要求较高的短信过滤系统。

4 实验资源及分析与评价

本文在自建短信语料库的基础上完成对比实验，其中正常短信 1 892 条，垃圾短信 270 条，将短信语料库随机分成 5 等份，其中 4 份用于训练样本，1 份作为测试样本。

4.1 评价指标

分类系统评价指标如下，包括两类短信各自的准确率(precision)和召回率(recall)，由于系统目标是垃圾短信过滤，于是增加了针对垃圾短信的综合评价指标(F1)：F1=(2×准确率×召回率)/(准确率+召回率)。

4.2 实验结果分析

(1)实验 1：探讨改进的特征权重计算方法对实验结果的影响。实验结果如表 1 所示。

表 1 特征权重计算方法对实验结果的影响

短信	训练短信数/条	测试短信数/条	准确率/%	召回率/%	F1
垃圾短信	216	54	67.7	85.2	75.5
正常短信	1514	378	95.2	94.2	94.7

其中测试样本中正常短信被误分为垃圾短信条数为 22 条，正常短信召回率为 94.2%；垃圾短信被误分为正常短信 8 条，准确率仅为 67.7%。

(2)实验 2：统一参数和取固定的阈值 θ 之后对实验结果的影响。该实验中取：α=1.5、β=1/1.5、θ=15。实验结果如表 2 所示。

表 2 选定参数对实验结果的影响

短信	训练短信数/条	测试短信数/条	准确率/%	召回率/%	F1
垃圾短信	216	54	71.0	81.5	75.9
正常短信	1 514	378	97.3	95.2	96.1

其中测试样本中正常短信被误分为垃圾短信条数为 18 条，正常短信召回率为 96.1%；而测试用的垃圾短信正确识别了 44 条，准确率为 71.0%。由此可见，参数对实验结果的影响不大。

(3)实验 3：去除野点对实验结果的影响。实验结果如表 3 所示。

表 3 去除野点对实验结果的影响

短信	训练短信数/条	测试短信数/条	准确率/%	召回率/%	F1
垃圾短信	216	54	79.3	85.2	82.1
正常短信	1514	378	97.9	96.8	97.3

从实验结果分析，仅有 12 条正常短信和 8 条垃圾短信被错误分类。通过去除野点，发现不仅缓减了抖动现象，而且提高了分类器的分类性能及正常短信的召回率。

Balanced Winnow 在训练速度和分类速度上具有较大优势，所以具有更高的实用价值，非常适合短信过滤的要求。另外，Winnow 作为一种在线学习方法，在训练集合不断扩大的情况下能够快速对分类器进行更新。正是基于 Winnow 不断学习、不断调整的机制，使其非常适合用户自己定制需要的分类标准。随着用户不断地反馈调整，整个系统会表现出越来越好的效果。

参考文献

[1] 黎路，秦卫平.浅析贝叶斯分类方法在手机垃圾短信过滤系统中的应用[J].科技广场，2007(7):76-78.

[2] 金展，范晶.基于朴素贝叶斯和支持向量机的自适应垃圾短信过滤系[J].计算机应用，2008，28(3):714-718.

[3] 王忠军.文本分类在短信过滤中的应用[D].辽宁:大连理工大学，2006.

[4] LITTLESTONE N. Learning quickly when irrelevant attributes abound: a new linear threshold algorithm. Machine Learning，1988(2):285-318.

[5] YANG YI MING，PEDERSEN JAN. A comparative study on feature selection in text categorization proceedings of the 14th international conference on machine learning [C]. Nashville：Morgan Kaufmann，1997:412-420.

[6] 周志军.中文邮件分类系统的研究及其实现[D].苏州:苏州大学，2005.

[7] 潘文峰，王斌.Winnow 算法在垃圾邮件过滤中的应用[C].第一届全国信息检索与内容安全学术会议论文集，上海，2004.

(收稿日期：2010-06-10)

作者简介：

闫红静，女，1984 年生，硕士研究生，主要研究方向：自然语言处理、数据仓库与数据挖掘。

邱书灵，女，1962年生，教授、研究生导师，主要研究方向：自然语言处理、人工智能、数据仓库、数据挖掘。