文章编号:1005-0523(2010)06-0075-05

# 基于非负矩阵分解特征提取的垃圾邮件过滤

陈 俊,刘遵雄

(华东交通大学 信息工程学院,江西 南昌 330013)

**摘要**:随着垃圾邮件的不断增多,它的危害性越来越严重,为了消除这种危害性,垃圾邮件的过滤技术就显得异常重要。由于垃圾邮件数据具有稀疏性、高特征维数和多重相关性等,所以直接对它进行分类的话会造成运算量很大和错误分类的问题。本文针对这些问题,先用非负矩阵分解的方法对实验数据进行特征提取,然后再用分类方法对它进行分类。在实验中,比较之后发现经过分解之后的数据比没有经过分解的数据有更高的分类准确率。

**关 键 词**:垃圾邮件;非负矩阵分解;特征

**中图分类号**:TP181 **文献标识码**:A

## 1 研究概况

垃圾邮件一般是指没有经过用户许可,却被强塞到用户邮箱的电子邮件。随着电子邮件成为人民生活中交流的一个主要方式,垃圾邮件也不断的泛滥起来,它对我们的危害也越来越大。其危害主要表现在以下几个方面:垃圾邮件的大量的存在占用了很大的网络带宽,甚至可能堵塞网络带宽,使网络传输效率降低,并且还占用了存储空间;有的垃圾邮件还会侵害别人的隐私权,给了黑客有可乘之机;垃圾邮件给世界的经济造成了巨大的损失;有些垃圾邮件还会蛊惑人心,骗人钱财,传播色情等不健康内容的垃圾邮件,从而会对现实社会造成危害,尤其是对青少年的危害很大。因此,垃圾邮件过滤技术的研究成为一个急需解决的问题。

早期的垃圾邮件过滤技术主要是基于规则的技术,并且这些规则都是人工预先制定的,对邮件进行模式匹配过滤技术。若在邮件中找到属于垃圾邮件关键字的个数达到一定阀值则判定此邮件为垃圾邮件。基于规则的过滤技术有以下几个缺点:首先,其自身的工作原理决定了这种过滤技术总是滞后于垃圾邮件的出现。例如,新的垃圾词汇出现,规则库中没有就不能准确过滤。其次,它是一种生硬二值判断,缺少可信的知识和模糊的判断。再次,系统需要用户定制自己的规则库,这就对用户有较高的要求,同时需要用户花大量时间更新规则,如果用户兴趣发生变化,这些规则也要进行很大变动,因此这种过滤技术不利于在普通用户中推广。还有,为避免邮件中出现垃圾词汇而被过滤掉,垃圾邮件发送者常常对某些关键词做个变形,就能轻易骗过过滤器。

随着模式识别和机器学习的不断发展,现在大多数的垃圾邮件过滤技术都是自动构建过滤模型和对垃圾邮件进行分类。目前大多数的垃圾邮件分类方法都是原先通用的文本分类方法,分为基于规则的方法和基于内容的分类方法两大类。主要有贝叶斯方法、支持向量机、KNN 方法、逻辑回归方法、决策树方法等。

由于垃圾邮件是高维的数据,直接对高维数据进行分类的话很容易发生维数灾难和错分的问题。因此,对高维数据进行降维显得很重要。现在主要的降维方法有特征提取和特征选择两种方法。特征选择是通过一些标准的统计方法选择出对分类贡献最大的若干特征,常用的有文档频率、$\chi^2$ 统计和信息增益等方法;而特征抽取是将原始的特征空间投影到低维特征空间,投影后的二次特征是原始特征的线性或者非线性组合。常用的方法有主要成分分析、Fisher 线性判别分析、非负矩阵分解等。

**收稿日期**:2010-10-11

**基金项目**:江西省教育厅科技研究项目(GJJ10446)

**作者简介**:陈 俊(1987—),硕士研究生,主要研究方向为模式识别,机器学习。

本文使用的降维方法是非负矩阵分解,非负矩阵分解方法首先把原始的实验数据分解成 $\boldsymbol{B}$ 和 $\boldsymbol{C}$ 数据矩阵。而 B * $\boldsymbol{C}$ 正是原始数据的另外一种相似的表达。B 代表了原始数据的一个新的基坐标,并且这个基坐标的维数远小于原始数据的维数。而 C 是原始数据在新的基坐标 B 上的投影数据,也可以理解为在 B 上的权值系数。因此我们可以把对原始的高维的垃圾邮件分类的问题转化为对 C 进行分类的问题。而 C 是一个低维的数据,所以在对它进行分类的时候有很多的优点,并且也很容易对它进行准确的分类。然后再用支持向量机,逻辑回归和核逻辑回归方法对分解之后的数据进行分类。

## 2 特征降维方法

### 2.1 特征提取

(1) 非负矩阵分解(NMF)。NMF 是一种常常用来分解和降维的方法,该方法是通过构造 $\boldsymbol{B}$ 和 $\boldsymbol{C}$ 来实现维度的降低。该算法的数学表达式是 $V \approx \boldsymbol{BC}$,它的基本思想是对非负的原始数据进行非负的分解。并且在功能上它基本上实现了对大脑的基于部分感知功能的模拟分析;在它的算法中它采用简单有效的迭代规则很好的保证了分解后数据的非负性;在应用上,非负性的数据大量的存在,且非负分解的结果具有明确的物理含义,作为一种低秩逼近的算法,它能有效的节约存储空间和计算资源。

为了实现矩阵的非负性的分解,首先需要定义一个差异函数来描述分解前后的逼近程度,然后才在非负性约束条件下求解.。最早提出的非负矩阵分解方法采用传统的梯度下降算法与加性迭代规则。有时也采用乘性迭代规则,更适合非负分解的特点,也就是在非负性初始化的基础上,在迭代过程中能简单地保持非负性,而加性迭代规则中就需要一个强制性的将负值变为零的步骤。

将矩阵分解看成如下含加性噪声的线性混合体模型

$$\boldsymbol{X}_{n\times m} = \boldsymbol{B}_{n\times r}\boldsymbol{C}_{r\times m} + E_{n\times m} \tag{1}$$

其中 $\boldsymbol{X}$ 代表是原始的数据,$\boldsymbol{B}$ 和 $\boldsymbol{C}$ 代表是分解之后的结果,而 $n$ 和 $m$ 代表原始矩阵有 $n$ 行 $m$ 列,$r$ 代表分解之后的行数,$\boldsymbol{E}$ 代表的是原始的数据和分解后的结果之间的差距,也就是损失数据。一般情况下 $r$ 的选择要满足 $(n+m) * r < n * m$,因为只有这样才可以保证分解之后的 $\boldsymbol{B}$ 和 $\boldsymbol{C}$ 的维数小于原始矩阵的维数。从上面的公式可以看出可以把 B 看成是对原始矩阵 X 的一组线性的逼近基向量,而 C 就可以看成是在基向量 B 上面的投影系数或者投影权值。这样的话就实现了用少量的基向量来表示大量的高维数据,也就是发现了原始数据之间的潜在关系。并且可以取得很好的逼近效果。

于是原先的问题就转化成了如何求解 $\boldsymbol{B}$ 和 $\boldsymbol{C}$。下面是求解 B 和 C 的过程。为了求解出 B 和 C 的数值,一般是通过最大似然法来求解。它的过程如下

$$\left.\begin{aligned}\{\boldsymbol{B},\boldsymbol{C}\} &= \arg\max_{B,C} p(\boldsymbol{X}\mid \boldsymbol{B},\boldsymbol{C})\\ &= \arg\max_{\boldsymbol{B},\boldsymbol{C}}[-\log p(\boldsymbol{X}\mid \boldsymbol{B},\boldsymbol{C})]\end{aligned}\right\} \tag{2}$$

假设噪声服从的是不同的概率分布,就可以得到不同类型的目标函数.考虑噪声是高斯噪声,也就是

$$\boldsymbol{P}\left(X_{ij}\mid \boldsymbol{B},\boldsymbol{C}\right) = \exp\left\{-\frac{1}{2}\left[\frac{\boldsymbol{X}_{ij}-\left(\boldsymbol{BC}\right)_{ij}}{\sigma_{ij}}\right]^2\right\} / \left(\sqrt{2\pi\sigma_{ij}}\right) \tag{3}$$

从而最大似然解就是下面的损失函数

$$\boldsymbol{L}\left(\boldsymbol{B},\boldsymbol{C}\right) = \frac{1}{2}\sum_{ij}\left[\boldsymbol{X}_{ij}-\left(\boldsymbol{BC}\right)_{ij}\right]^2/\sigma_{ij}^2 + \sum_{ij}\log\left(\sqrt{2\pi\sigma_{ij}}\right) \tag{4}$$

因为要得到的解是一个局部的最优解,所以就需要对上式进行求导。分别对 $\boldsymbol{B}$ 和 $\boldsymbol{C}$ 进行求导就可以得到

$$\boldsymbol{C}_{kj} \leftarrow \boldsymbol{C}_{kj}\frac{\sum_i \boldsymbol{B}_{ik}\boldsymbol{X}_{ij}/\left(\boldsymbol{BC}\right)_{ij}}{\sum_i \boldsymbol{B}_{ik}} \tag{5}$$

$$\boldsymbol{B}_{ik} \leftarrow \boldsymbol{B}_{ik}\frac{\sum_i \boldsymbol{C}_{kj}\boldsymbol{X}_{ij}/\left(\boldsymbol{BC}\right)_{ij}}{\sum_j \boldsymbol{C}_{kj}} \tag{6}$$

而这也就是在优化过程中不断的迭代公式,经过一定次数的迭代之后就会发现它的结果会最终趋于某一个值,而这个值也就是所要求的局部的最优解。

(2) 主要成分分析 (PCA)。PCA 是一种常用的特征提取的算法,也称为 K2L 变换。PCA 是将多个变量通过线性变换以选出较少个数重要变量的一种多元统计分析方法。它在人脸识别领域得到了广泛的应用,并且提出了特征脸的概念。它首先通过求解原始数据矩阵 $\boldsymbol{X}$ 减去它平均值之后的协方差矩阵,然后求解协方差矩阵的特征值和特征向量。最后按照特征值绝对值的大小对它所对应的特征向量进行排序,选择出前面几个最大特征值所对应的特征向量作为主要成分。因为对应最大特征值的特征向量能够反映数据的最大相异性。

(3) 偏最小二乘 (PLS)。PLS 是 1983 年被提出来的一种综合考虑自变量对因变量的一种回归建模的方法。它的基本思想是,分别在 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 中提取他们各自的潜在变量 $\boldsymbol{t}$ 和 $\boldsymbol{u}$,即分别对 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 进行线性化。并且要求 t 和 u 应尽可能的携带它们各自数据表中的变异信息,并且 t 和 u 之间的相关程度应该最大。这样依次对 X 和 Y 进行最大差异潜在变量的提取,直到达到满意的精度为止。

### 2.2　特征选择

特征选择也是一种常见的特征降维方法,它是从原始众多的属性和特征之中根据具体的方法选择出一些对分类或者表现差异贡献大的特征,并且选择出来的特征是原始特征的真子集。在刚开始的文本预处理阶段都会用到特征选择的方法。常见的特征选择的方法有:度量词条 $w$ 和文档类别 $c$ 之间的相关程度的 CHI 方法;选择有效的特征进行相关反馈的几率比方法;计算特征权值的方法;构造树结构模型的方法和信息增益等方法。

## 3　垃圾邮件分类算法和技术

目前存在着很多垃圾邮件的分类方法,其中包括支持向量机方法,贝叶斯方法,KNN,Ripper 方法,决策树和逻辑回归等方法。这里主要采用的方法是支持向量机,逻辑回归和核逻辑回归方法。

### 3.1　支持向量机 (Support Vector Machine,SVM)

SVM 是一种在机器学习领域比较流行的分类算法,它不但有很好的性能比并且可以通过它特有的核函数来处理高维的数据。所以它在处理二元分类的问题上有很大的优点,该算法是通过构造一个最大间隔的最优分类超平面来实现分类的目标。由于 SVM 的求解最后转化成二次规划问题的求解,因此 SVM 的解是全局唯一的最优解。

### 3.2　逻辑回归(Logistic Regression,LR)

逻辑回归模型是 1920 美国学者柏尔和利德(Robert B·Pearl and Lowell J·Reed)提出的,并开始用在人口估计和预测中推广应用,并引起了广泛的注意。逻辑回归是研究因变量为二分类或多分类观察结果与影响因素(自变量)之间关系的一种多变量分析方法,属概率型非线性回归。逻辑回归根据因变量的取值类型不同,又可分为二项分类逻辑回归、有序分类逻辑回归和无序多项分类逻辑回归,其中二项分类逻辑回归是其他逻辑回归的基础。LR 的建模过程本身就具有挑选变量的功能,即只有对因变量贡献率达到一定程度的特征变量才能进入回归模型中,对因变量没有贡献或者贡献很小的特征变量最终会被剔除。由于 Logistic 回归是非线性模型,因此最大似然估计法经常用于模型估计。

### 3.3　核逻辑回归(Kernel Logistic Regression,KLR)

逻辑回归是一种经典的利用线性模型估计的算法,属于广义线性模型方法中的一种,它不需要任何关于样本分布的假设。由逻辑回归模型的方法是一种线性的方法,所以它在解决一些非线性问题上有很多缺陷,因此就诞生核逻辑回归。核逻辑回归 把支持向量机里面的核方法引入到逻辑方法里面。从而可以解决一些非线性的问题,核逻辑回归经常使用的核方法有线性核函数,多项式核函数,RBF 核函数和 sigmoid 核函数。

## 4　试验结果

本文的垃圾邮件的试验数据是从 UCI 机器学习知识库上下载的。这里总共有 4 601 封电子邮件,其中包括 1 813 封垃圾邮件和 2 788 封非垃圾邮件。并且一般的 Zero-rule 分类方法采用这个数据的准确率

只有60.6%左右。

从UCI机器学习知识库中下载的数据是以文本形式存储的，并且每篇邮件都有不同的长度。在这个数据的基础上排除一些不相关和一些无关紧要的属性之后，得到了转换之后的实验数据。

原始的邮件数据是不能被计算机识别的，因此需要选择邮件数据中的某些有代表性的特征来表示原始的邮件数据。这里用57个属性和一个是否为垃圾邮件的属性来表示原来的垃圾邮件数据，其中57个属性中的48个属性是邮件中常用词组，它们是make，address，all，3d，our，over，remove，internet，order，mail，receive，will，people，report，addresses，free，business，email，you，credit，your，font，000，money，hp，hpl，george，650，lab，labs，857，data，415，85，technology，1999，parts，pm，direct，cs，meeting，original，project，re，edu，table，and conference。除了这些词组以外还有6个常用的标点符号，它们是：；（，[，！，$，and ＃。最后还有最后3个属性是：平均每个句子有多少词组；在邮件中最长的句子有多少词组和邮件中总共包括多少词组。最后的那个属性中1代表垃圾邮件，0表示不是垃圾邮件。

本实验主要采用支持向量机，逻辑回归，核逻辑回归3种方法来对实验数据进行分类。第一部分是直接用上面的3种方法对垃圾邮件的数据进行分类，结果它们的分类准确率都只有61%左右。表1为直接用分类方法对垃圾邮件数据进行分类的结果。

**表1 原始数据的分类准确率和时间**

| 分类方法 | 分类正确率/% | 分类时间/s |
|---|---|---|
| SVM | 60.8 | 24.5 |
| LR | 61.3 | 21.9 |
| KLR | 61.2 | 43.2 |

从上面的实验结果可以看出，直接用分类方法对实验数据进行分类的话，准确率只有61%左右。从实验所用的时间来看，KLR所花的时间最长，因为它要把低维的数据投影到高维的空间再进行分类，所以它花的时间比其他的2个方法要多。

本实验的第二部分是先用非负矩阵分解的方法对实验数据进行分解之后，然后再用分类方法对分解之后的数据进行分类。为了更好的得出不同维度对原始数据的解释能力，本文特意把原始的57维的实验数据分解为5维，10维，20维，30维，40维。然后在上述分解之后的数据基础上，分别用支持向量机，逻辑回归和核逻辑回归三种方法对它们进行分类。表2和表3分别为经过特征提取之后的分类准确率和分类时间表。

**表2 特征提取之后的分类准确率**

| 分类方法 | 5维/% | 10维/% | 20维/% | 30维/% | 40维/% |
|---|---|---|---|---|---|
| NMF＋SVM | 80.1 | 80.7 | 81.3 | 81.2 | 81.0 |
| NMF＋LR | 63.2 | 64.1 | 64.3 | 63.7 | 62.0 |
| NMF＋KLR | 64.3 | 65.6 | 65.7 | 64.9 | 63.1 |

**表3 特征提取之后的分类时间**

| 分类方法 | 5维/s | 10维/s | 20维/s | 30维/s | 40维/s |
|---|---|---|---|---|---|
| NMF＋SVM | 12.1 | 10.7 | 9.5 | 20.3 | 21.2 |
| NMF＋LR | 17.4 | 15.0 | 14.8 | 15.2 | 16.7 |
| NMF＋KLR | 31.3 | 28.9 | 26.4 | 27.1 | 26.7 |

从上面的实验结果可以看出，分类的正确率呈现一个先升后降和分类时间减少的发展过程。这是由于维度过低不能完好的表达原始的所有特征，所以随着维度的增高有一个分类准确率上升和时间缩短的发展阶段。但是实验所用的数据本身的维度不是很高，所以当达到20维左右就相似的表达了原始邮件的信息，因此正确率不再上升和时间不在缩短；并且随着维度的增加有一个很小的递减过程。比较表2和表1可以看出，经过非负矩阵分解特征提取之后的正确率比直接分类的方法有更高的分类准确率。其中以支持向量机提高的最多，提高了20%左右；比较表3和表1可以知道分类的时间缩短了一半左右。

表 4 是实验的查全率对比的一个表格,其中前面几列是经过 NMF 分解之后的结果,最后一列是没有经 NMF 分解的实验结果。

**表 4　几种方法查全率的比较**

| 分类方法 | 5 维/% | 10 维/% | 20 维/% | 30 维/% | 40 维/% | 57 维/% |
|---|---|---|---|---|---|---|
| SVM | 71.8 | 73.1 | 74.0 | 76.2 | 79.3 | 63.1 |
| LR | 81.2 | 82.2 | 84.5 | 87.1 | 88.3 | 62.5 |
| KLR | 74.3 | 75.1 | 76.4 | 78.2 | 79.6 | 64.2 |

从表 4 中可以看出,经过 NMF 之后的实验数据比没有经过 NMF 分解的数据有更高的查全率。

## 5　结论和今后的工作

通过上面的理论分析和试验的结果可以知道,经过非负矩阵分解之后的数据不但实现了降维,而且还很好的保留了原始数据潜在的特征。这点可以从支持向量机的试验中很清楚的看出,试验结果表明经过非负矩阵分解之后的数据比原始的实验数据有更高的分类正确率和查全率,并且缩短了分类的时间。因此它在今后的垃圾邮件的分类中会有很好的应用。

在今后的工作中将从以下几方面作进一步研究:第一,邮件之间的潜在联系好多时候都是非线性的,尝试寻找一些非线性的特征提取方法。比如核偏最小二乘,核主要成分分析等;第二,在邮件处理中,一般来说,将一封正常邮件错分为垃圾邮件的代价要远远大于将垃圾邮件分为正常邮件的代价,这种评价机制被称为代价敏感评价机制。可尝试在今后的工作中引入这种机制进行邮件过滤。

**参考文献:**

[1] 张禾.逻辑回归法在遥感数据特征选择和分类中的应用[J].Science & Technology Association,2007,5(3):17-18.
[2] 盛鹏.基于全文过滤的垃圾邮件防范机制[D].云南:西南大学,2006:40-45.
[3] 曹兆龙.基于支持向量机的多分类算法研究[D].上海:华东师范大学,2007:15-17.
[4] 陈治平,王雷.基于自学习 K 近邻的垃圾邮件过滤算法[J].计算机应用,2005,25(12):7-8.
[5] 赵向军,路梅.垃圾邮件过滤算法研究[J].徐州师范大学学报,2006,24(5):52-55.
[6] 刘维湘,郑南宁,游屈波.非负矩阵分解及其在模式识别中的应用[J].科学通报,2006,51(3):241-250.
[7] GUILLAMET D,VITRIà J,SCHIELE B.Introducing a weighted non-negative matrix factorization for image classification[J].Pattern Recognition Letters,2003,24(14):2447-2454.
[8] 李滔,王俊普,吴秀清.基于特征矢量集的核 Logistic 回归[J].小型微型计算机系统,2007,27(6):980-985.
[9] 喻军.几种典型特征抽取方法比较及其在人脸识别中的应用[J].江南大学学报,2009,8(5):543-546.
[10] 王鹏鸣,吴水秀,王明文,黄国斌.基于偏最小二乘特征抽取的垃圾邮件过滤[J].中文信息学报,2008,22(1):74-79.

# Spam Filtering Based on Non-negative Matrix Factorization

Chen Jun, Liu Zunxiong

(School of Information Engineering, East China Jiaotong University, Nanchang 330013, China)

**Abstract**: The growing spam is more and more harmful to people, so the technology of filtering spam becomes extremely important. With data sparseness, high dimensionalities and multiple correlation of spam, its direct classification will lead to large computation and false classification. Therefore, the paper firstly uses the non-negative matrix factorization method to extract the features of the experimental data, and then classifies it. In the experiment, comparing results show that data of decomposition has high accuracy than original data.

**Key words**: spam; non-negative matrix factorization; characteristic



(责任编辑　王建华)