

文章编号:1005-0523(2014)04-0130-06

基于套索 (Lasso) 的中文垃圾邮件过滤

徐征¹, 刘遵雄², 张贤龙²

(华东交通大学1.电气与电子工程学院;2.信息工程学院,江西 南昌 330013)

摘要:使用向量空间模型表示的文本邮件数据高维而稀疏,不利于邮件过滤分类模型的建立,通常需在分类器训练前进行维数约减。Lasso回归是一种基于 l_1 正则化的多元线性模型,其在模型参数估计的同时实现了变量选择。提出使用Lasso回归进行垃圾邮件过滤,建立Lasso回归邮件分类模型、Lasso回归词条选择结合逻辑回归的分类模型,结合中文文本垃圾邮件数据集TREC06C进行垃圾邮件过滤实验。实验结果表明Lasso回归词条选择结合逻辑回归的邮件分类模型性能更佳。

关键词:中文文本邮件;垃圾邮件;过滤;Lasso;逻辑回归

中图分类号:TP391

文献标志码:A

基于内容的垃圾邮件过滤技术逐步成为垃圾邮件过滤研究重点^[1-2],研究方法多为模式识别、数据挖掘等领域的分类或回归技术,常用的模型算法有决策树、支持向量机^[3]、贝叶斯分类和逻辑回归(Logistic Regression)^[4-5]等。建立垃圾邮件过滤模型首先需要获得一定数目的正常邮件和垃圾邮件,使用向量空间模型(vector space model, VSM)表示的邮件数据特征维度高而且稀疏,基于这些数据建立分类模型时计算量庞大,分类器模型容易陷于“过学习”,其泛化能力不好。解决该问题通常的办法是先于分类器训练进行特征维数约减,以提高算法运算效率、改善分类性能。

特征选择是常用的维数约减方法,其采用一些统计和信息论方法,如文档频率、 χ^2 统计和信息增益等^[6],选择出对分类贡献最大的特征子集。最小绝对缩减和变量选择算子(least absolute shrinkage and selection operator, Lasso)回归是一种回归系数绝对值和受限制的多元线性回归方法,也称为“套索”回归,其可以同时实现模型参数估计和变量选择^[7-9]。Lasso回归研究发展很快,相应的改进模型(弹性网、组套索等)和算法相继被提出,而最小角度回归(least angle regression, LAR)算法能很好解决Lasso回归的计算问题^[10]。

提出基于套索的垃圾邮件过滤算法,使用Lasso回归选择特征词条,建立逻辑回归分类模型。结合中文垃圾邮件的数据集TREC(Text retrieval conference)^[11]进行垃圾邮件过滤模拟实验,并对求得的邮件过滤性能评价指标值给以分析说明。

1 基于VSM(向量空间模型)的邮件表示

研究文本邮件数据,垃圾邮件过滤等效于二元文本分类问题。进行垃圾邮件过滤,首先需使用向量空间模型VSM将邮件数据表示成易于计算机处理的形式,即使用中文分词对邮件进行预处理(得到邮件正文和主题的词条),选择一定词条特征进行统计分析,将每封邮件表示为一长度等于词条数的向量。从而邮件数据集表示成词条文档矩阵 $X_{n \times m} = (x_1, x_2, \dots, x_m)$ 和邮件目标向量 $y_{n \times 1}$,其中 m 为词条数, n 为文档数, y 的

收稿日期:2014-05-10

基金项目:国家自然科学基金项目(71361009,61065003);教育部人文社会科学研究项目(13YJC630192);华东交通大学校立科研课题(09DQ04)

作者简介:徐征(1978—),女,讲师,主要研究方向为数理统计、机器学习、非线性系统分析与建模及其在网络行为分析中的应用。

分量取-1或1表示对应邮件的类别(正常或垃圾邮件)。

X 每行成分值为对应词条关于不同邮件的权值,表示词条在邮件中的重要程度。常用的几种权值计算方法为布尔权重,词频权重, $tf-idf$ 权重、 tf 权重、 lfc 权重、 lrc 权重。在本文中选用 lrc 权重,使用词频的对数,减少词频差异带来的干扰。 lrc 权重的计算公式如下

$$x_{ki} = \frac{\log(f_{ki} + 1.0) \times \log(\frac{n}{n_i})}{\sqrt{\sum_{i=1}^m [\log(f_{ki} + 1.0) \times \log(\frac{n}{n_i})]^2}} \quad (1)$$

其中: x_{ki} 表示词条 i 在文档 k 中的权重; f_{ki} 表示词条 i 在文档 k 中的词频数; n 是邮件中文档的总数; n_i 表示词条 i 的文档频数。

2 Lasso 回归及 LAR 算法

提出使用 Lasso 回归进行词条特征选择。Lasso 回归是一种回归系数受限制的最小二乘(即惩罚最小二乘)问题,能较好地处理模型学习中的“过学习”问题,在进行模型参数估计的同时实现变量选择。Lasso 回归优化问题为

$$\arg \min_{\beta} \|y - X\beta\|_2^2 \quad \text{s.t.} \|\beta\| \leq s \quad (2)$$

其拉格朗日表达式为

$$\arg \min_{\beta} \|y - X\beta\|_2^2 + \lambda \|\beta\| \quad (3)$$

其中: λ 为惩罚参数, λ 与 s 存在某种对应关系。Lasso 回归求解算法很多,有射击算法(shooting algorithm)、同伦(homotopy)算法和最小角度回归(least angle regression, LAR)算法。LAR 算法是对传统的 stagewise forward selection^[8]方法加以改进而得到的有效精确方法,并且在计算上也比 stagewise 方法简单,它最多只需要通过 m 步(m 为变量个数),就能得到拟合解。LAR 与最小二乘计算复杂度相当,能很好的解决 Lasso 回归的计算问题。LAR 算法如下:

首先需要对数据进行预处理,使其去均值和标准化。

$$\sum_{i=1}^n x_{ij} = 0, \sum_{i=1}^n x_{ij}^2 = 1 \quad j = 1, 2, \dots, m \quad (4)$$

定义 $\hat{\beta} = (\beta_1, \beta_2, \dots, \beta_m)^T$ 为当前拟合向量 μ 的系数

$$\mu = X\hat{\beta} = \sum_{j=1}^m x_j \beta_j \quad (5)$$

则 x_i 跟残差 $y - \mu$ 的相关系数为

$$c_i = x_i^T (y - \mu) \quad (6)$$

逐步计算拟合值 μ : 刚开始时,相关系数都为0,然后找出跟残差(此时为 y)相关系数最大的变量,假设是 x_{j_1} , 将其加入到活动集,这时在 x_{j_1} 的方向上找到一个最长的步长 $\hat{r}_1$, 直到出现下一个变量(假设是 x_{j_2})跟残差的相关系数跟 x_{j_1} 与残差的相关系数相等,此时把 x_{j_2} 加入活动集里, LAR 继续在这2个变量等角度的方向进行拟合,找到第3个变量 x_{j_3} 使得该变量、活动集中变量跟残差的相关系数相等,随后 LAR 继续找寻下一个变量,直至使残差减小到一定范围内。

LAR 求解 Lasso 算法的步骤如下:

$$\mu = 0. \text{ while } (|y - \mu(k)| < \varepsilon \quad \text{and} (|A| \leq m))$$

$$1) \ c = X^T(y - \mu) = (c_1, c_2, \dots, c_m)^T, \ A = \{j: |c_j| = \max(c) = C\}, \ X_A = [\dots x_j \dots j \in A]$$

$$2) \ u(k) = X_A w_A \quad (w_A = a_A G_A^{-1} 1_A \quad a_A = (1_A^T G_A^{-1} 1_A)^{-\frac{1}{2}} \quad G_A = X_A^T X_A, 1_A = (1, \dots, 1)^T)$$

$$3) \ a_A = X^T u = (a_1, a_2, \dots, a_m)$$

$$\begin{aligned}
& 4) \tilde{r} = \min_{j \in A}^+ \left\{ -\frac{\beta_j}{w_j} \right\}, \text{ if } |A| < m \quad \hat{r} = \min_{j \in A^c}^+ \left\{ \frac{C - c_j}{a_A - a_j}, \frac{C + c_j}{a_A + a_j} \right\} \quad \text{else } \hat{r} = \frac{C}{A_A} \\
& 5) \text{ if } \hat{r} < \tilde{r} \quad \mu(k) = \mu(k-1) + \tilde{r}u(k) \quad \beta_A = \beta_A + \tilde{r}w_A \quad A = A - \{\hat{j}\} \\
& \quad \text{else } \mu(k) = \mu(k-1) + \hat{r}u(k) \quad \beta_A = \beta_A + \hat{r}w_A \quad A = A \cup \{\hat{j}\} \\
& \text{return } \beta_A
\end{aligned}$$

其中: u 为当前最小角度方向, 即角平分线方向; μ 为当前拟合的 y 值; c 为残差与变量的相关系数; $\hat{r}, \tilde{r}$ 为当前的最长步长; A 为变量集合; X_A 为搜索矩阵; W_A, a_A, G_A 为中间变量。

使用 LAR 算法, 求出 Lasso 回归的解向量 β , 而对应 β 系数不为 0 的变量, 即为选出的关键词条。然后基于这些词条建立逻辑回归分类模型。

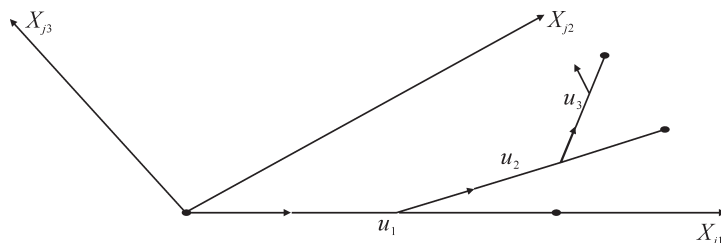


图 1 LAR 选择示意图

Fig. 1 The schematic diagram of LAR selection

3 实验和结果分析

3.1 数据准备及评价标准

本文使用的中文垃圾邮件数据集 TREC06c 是 TREC (text retrieval conference) 2006 年给出的中文垃圾邮件评测公共数据, 数据及选择概况见表 1。

表 1 中文垃圾邮件数据集

Tab. 1 Dataset of Chinese spam

封

数据集	Ham		Spam		总计	
	原始	选取	原始	选取	原始	选取
TREC06C	21 766	9 845	42 854	20 155	64 620	30 000

选取 3 万封原始邮件进行预处理, 从邮件数据中提取出邮件的内容, 并用 VSM 模型将邮件数据表示成文档词条矩阵。预处理过程用 C++ 语言实现, 提取邮件主题、正文等重要信息, 包括邮件内容的解码、中文的文本分词、构建邮件内容词语的词库和 ltc 权重计算等步骤。考虑到邮件主题可能包含更多邮件信息, 将出现在邮件主题中的词条权重进行加倍。

使用中科院的分词组件 ICTCLAS 进行中文分词, 对分词后的邮件数据进行统计得到 49 452 个词条, 考虑到邮件中许多词条的出现频数很少, 去除文档频率小于 0.2% 的词条, 以及空格、特殊的标点符号等干扰字符, 最终得到 8 879 个词条。实验中发现, 有一些邮件数据没有正文, 从而生成全 0 数据, 删除这样的数据。考虑到在实际生活中, 垃圾邮件的数目往往比正常邮件数据多, 实验中以垃圾邮件跟正常邮件 2:1 的比例随机抽取了 5 000 封垃圾邮件, 2 500 封正常邮件构成训练样本, 在其余数据中随机抽取 4 000 封垃圾邮件, 2 000 封正常邮件作为测试数据。建立 Lasso 回归邮件分类模型 (简称“Lasso 方法”), 再者基于 Lasso 回归选取的特征词条建立逻辑回归分类模型 (简称“逻辑回归方法”) 进行垃圾邮件过滤实验, 使用如下指标进行性能评价:

$lHm\%$ 表示正常邮件的误判率; $lsm\%$ 表示垃圾邮件的误判率; $l-ROCA\%^{[10]}$: ROC 曲线上方的部分, 曲线下面部分的面积反映了分类器阈值在取得; 所有可能值上, 分类器效率 (effectiveness) 的一个累计度量, 从而避免用单一的 $hm\%$; 或 $sm\%$ 进行衡量的局限性。1-ROCA% 是众多指标中最重要的一项; $llam\%$: 平均

误分率对数。定义如下

$$lam\% = \log it^{-1} \left(\frac{\log it(hm\%) + \log it(sm\%)}{2} \right) \quad (7)$$

其中: $\log it(x) = \log\left(\frac{x}{1-x}\right)$, 而 $\log it^{-1}(x) = \frac{e^x}{1+e^x}$ 。

$hm:1$ 为当 $hm\%=0.1$ 时, 对应的 $sm\%$ 的值。因为在实际中, 将一封正常邮件错分为垃圾邮件的代价系数要比将一封垃圾错分为正常邮件的代价系数高的多。

3.2 实验及结果分析

使用获得的数据进行垃圾邮件过滤实验, 训练样本是一个维度为 $7\,500 \times 8\,879$ 的文档词条矩阵和对应样本的类别目标值向量。首先进行 Lasso 回归模型训练, 使用 LAR 算法求解 Lasso 问题的解 β , 其中对应系数不为 0 的词条就是选中的词条。式(3)中的 λ 控制着回归系数 l_1 范数的惩罚, 使用不同的 λ 值进行垃圾邮件过滤实验, λ 的值分别取 12、10、8 和 7, 分别求解得出不同数目的词条, 分别为 467、672、1 106 和 1 481。实验中将使用不同 λ 值训练获得的 Lasso 回归模型, 并将其用于 6 000 封测试邮件的分类, 计算相应评价指标值见表 2, 对应的 ROC 曲线见图 2。根据表 2, 随着词条特征维度的增加, 邮件过滤的各项指标大体上有所提高, 但是正常邮件误判率($hm\%$)有所增大。

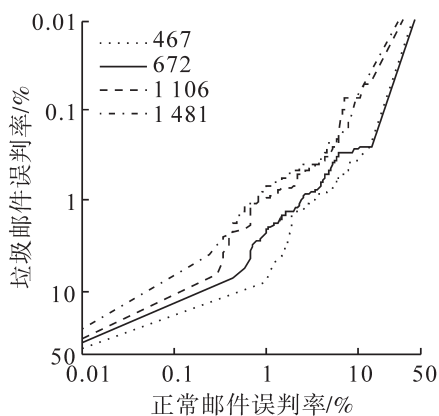


图2 逻辑回归分类 ROC 曲线图

Fig. 2 ROC curve of ogistic regression and classification

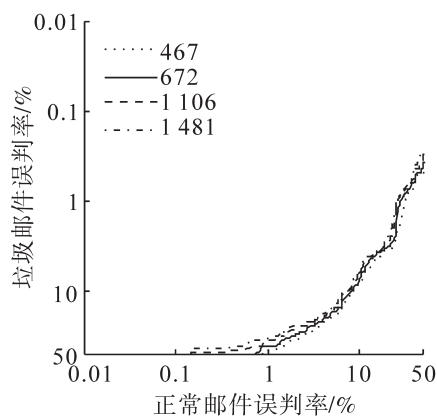


图3 Lasso 分类 ROC 曲线图

Fig. 3 ROC curve of Lasso Classification

表2 用 Lasso 回归实验结果

Tab. 2 Experimental results of the Lasso regression

λ	Terms	1-ROCA%	$hm:1$	$lam\%$	$hm\%$	$sm\%$	Time/ms
12	467	0.034 363	0.091	0.105 903	0.064	0.170 25	4.961
10	672	0.030 658	0.077 25	0.092 351	0.067	0.126	4.212
8	1 106	0.027 174	0.065 25	0.08 776	0.069 5	0.110 25	4.321
7	1 481	0.025 315	0.064 25	0.085 761	0.071	0.103 25	4.181

表3 Lasso 词条选择结合逻辑回归的实验结果

Tab. 3 The experimental results with the logistic regression and Lasso lexical selection

λ	Terms	1-ROCA%	$hm:1$	$lam\%$	$hm\%$	$sm\%$	Time/ms
12	467	0.007 371	0.003 5	0.018 337	0.037	0.009	23.135
10	672	0.004 316	0.002 75	0.016 319	0.033	0.008	37.553
8	1 106	0.002 247	0.000 75	0.011 65	0.022 5	0.006	83.873
7	1 481	0.001 436	0.000 75	0.009 394	0.016	0.005 5	143.062

基于 Lasso 回归实验选出的词条,建立相应的逻辑回归分类模型,并进行测试,相应评价指标值见表3。对比表2和表3,可以看出,在性能方面等方面,使用逻辑回归的 $1-ROCA\%$ 确实明显优于 Lasso 回归,而且 Lasso 过滤器的误判率也高于逻辑回归,但是逻辑回归方法的时耗要比直接使用 Lasso 做分类判别的时耗高,因为前者比后者多出了逻辑回归模型建立的时间。所以在时间效率上,逻辑回归不如 Lasso 过滤算法。

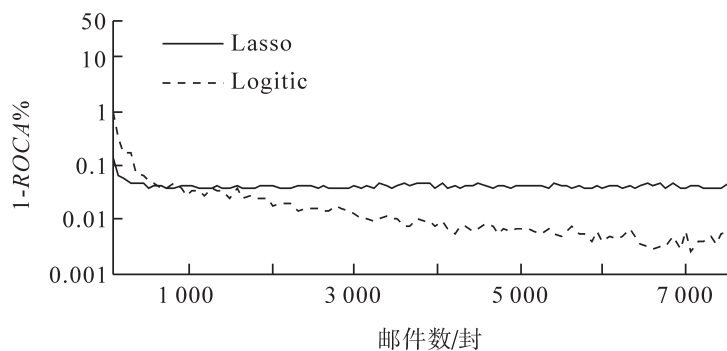


图4 过滤器的 ROC 学习曲线

Fig. 4 ROC learning curve of Filter

图4是两种方法的 ROC 学习曲线,可看出,当训练样本数目很少的情况下,Lasso 的效果比较好,随着训练样本数目的增加,虽然两种过滤算法对于垃圾邮件的预测性能都有所提高,但是,当训练样本达到一定数目时,使用 Lasso 过滤算法的 $1-ROCA\%$ 值基本趋于稳定了,而逻辑回归方法的性能有比较显著的提升。当训练样本足够多的时候,选用逻辑回归分类器的效果会更好。

4 结论

Lasso 回归是一种基于回归系数 l_1 正则化的统计建模技术,已被广泛用于社会、经济和科学等领域的数据分析与判决策上。利用 Lasso 回归对中文垃圾邮件数据进行建模分析,建立 Lasso 回归分类模型,基于 Lasso 回归选取的特征词条建立逻辑回归分类模型,结合中文垃圾邮件数据进行模拟,实验表明基于 Lasso 回归词条选择的逻辑回归分类模型以时间效率损失为代价获得了较好的分类性能。

参考文献:

- [1] 赵俊生,苏依拉,马志强.基于内容过滤的反垃圾邮件系统模型研究[J]. 内蒙古农业大学学报:自然科学版,2013,34(3):163-167.
- [2] 陈俊,刘遵雄. 基于非负矩阵分解特征提取的垃圾邮件过滤[J]. 华东交通大学学报,2010,27(6):75-80.
- [3] 沈跃伍. 基于在线学习的垃圾邮件过滤技术研究[D]. 哈尔滨理工大学,2012.
- [4] WEI DENGPIG. A logistic regression model for semantic web service matchmaking[J]. Science China(Information Sciences), 2012,07:1715-1720.
- [5] 张恒,秦宾,许金凤. 上市公司财务预警的正则化逻辑回归模型[J]. 华东交通大学学报,2011,28(6):42-47.
- [6] 张文良,黄亚楼,倪维健. 基于差分贡献的垃圾邮件过滤特征选择方法[J]. 计算机工程,2007(8):80-82.
- [7] TIBSHIRANI R. Regression shrinkage and selection via the Lasso[J]. Journal of the Royal Statistical Society, 1994,58(B):267-288.
- [8] 龚建朝. Lasso 及其相关方法在广义线性模型模型选择中的应用[D]. 中南大学,2008.
- [9] 孙燕. 随机效应 Logit 计量模型的自适应 Lasso 变量选择方法研究[J]. 数量经济技术经济研究, 2012(12):147-157.
- [10] EFRON B, HASTIE T, JOHNSTONE L, et al. Least angle regression[J]. Annals of Statistics, 2004,32:407-451.
- [11] 方慧. TREC 发展历程及现状分析[J]. 新世纪图书馆,2010(1):57-62.

Spam Filtering of Chinese Text Email Via Lasso

Xu Zheng¹, Liu Zunxiong², Zhang Xianlong²

(1. School of Electrical and Electronic Engineering, East China Jiaotong University, Nanchang 330013, China; 2. School of Information Engineering, East China Jiaotong University, Nanchang 330013, China)

Abstract: Text email data depicted with vector space model are of high dimensionality and sparsity, which are not suitable for establishing email filtering classification model. Generally, such data should be reduced before classifier training. Lasso regression is a multivariate linear model based on l_1 regularization, which can estimate model parameters while selecting the variables simultaneously. In this paper, the approaches to email classification based on Lasso are proposed. Also, the Lasso classification model and the logistical model with the selected term are established. Besides, simulation experiments with TREC06C are carried out, and the results show that logistic regression model plus the term selected with Lasso achieves better performances.

Key words: Chinese text email; spam; filtering; Lasso; logistic regression

(上接第 122 页)

Treatment Experiments of Coordination Extraction-Resin Absorption for Caprolactam Wastewater

Lu Xiuguo, Lin Pan, Deng Yue, Huang Yanmei

(School of Civil Engineering and Architecture, East China Jiaotong University, Nanchang 330013, China)

Abstract: The treatment of caprolactam wastewater by coordination extraction-resin absorption is discussed in this paper. The experimental study showed that for caprolactam wastewater of $\text{COD}_{\text{Cr}} 58000 \text{ mg/L}$ with 3500 times of chromaticity, about 99% of COD_{Cr} was removed and chroma reached 100% through coordination extraction-resin absorption. Besides, the factors of the experiment are analyzed and the further research problems are illustrated in this paper.

Key words: coordination extraction; resin absorption; caprolactam; removal rate