

# 基于多任务学习的邮件过滤系统的研究

许棣华<sup>1,2</sup>, 王志坚<sup>1</sup>

(1. 河海大学 计算机信息工程学院, 江苏 南京 210098;

2. 南京邮电大学 计算机学院, 江苏 南京 210003)

**摘要:**随着电子邮件的广泛使用,如何有效地避免和防范垃圾邮件的侵扰已成为一个亟待解决的问题。受机器学习在邮件过滤中研究和应用的启发,利用多任务学习(multitask learning)的特性,将判断一个用户的邮件是否为垃圾邮件看作一个任务(task),利用多任务学习中任务相关性假设,提出一种基于多任务学习的邮件过滤系统。实验表明,该系统对中英文邮件语料都是可靠和有效的,尤其对于同一邮件列表(mail list)中的用户的邮件。

**关键词:**多任务学习;任务相关性;邮件过滤;分类

**中图分类号:**TP393

**文献标识码:**A

**文章编号:**1673-629X(2010)10-0137-04

## Research of Spam Filter System Based on Multitask Learning

XU Di-hua<sup>1,2</sup>, WANG Zhi-jian<sup>1</sup>

(1. College of Computer and Information Engineering, Hohai University, Nanjing 210098, China;

2. College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

**Abstract:** With the widespread use of e-mail, how to effectively avoid and prevent junk e-mail has become very urgent. Inspired by the research and application of machine learning in spam filter, a spam filter based on multitask learning is proposed, considering whether a user's e-mail is spam or legitimate as a task. Using tasks relevance coefficient, the system classifies emails, with the assumption of task relevance in multitask learning. Experiments show that the system is reliable and effective for both English and Chinese corpus, especially for the mails in a mail list.

**Key words:** multitask learning; task relevance; spam filter; classification

## 0 引言

作为网络中使用最频繁的应用之一,电子邮件(E-mail)作为一种快捷而经济的通信手段,越来越在人们生活中起着重要作用。但随之而来的是垃圾邮件(spam或junk email)的泛滥,严重地干扰了人们的工作、生活。各种邮件过滤系统(spam filter)应运而生。

通常意义上的垃圾邮件是指未经主动请求的、带有商业性或政治性目的的大批量的电子邮件。从过滤技术上划分,目前主要有基于安全列表的过滤、基于规则的过滤、基于内容的过滤,将前二者称为“显示规则”,后一种称为“隐式规则”<sup>[1]</sup>。由于邮件过滤器的使用,垃圾邮件发送者(Spammer)也不断开发出各种工

具和方法(如,对邮件进行伪装)来降低过滤器对垃圾邮件的识别率,使得具有“显示规则”的过滤器对于这样的邮件往往失效。因此具有“隐式规则”的过滤器往往决定了邮件服务器的服务质量。这种隐式规则表现为对垃圾邮件的自适应性和自我学习能力,它能够针对变化的邮件的特点,实时地、自动地更新过滤规则。基于机器学习(Machine Learning, ML)<sup>[2]</sup>的邮件过滤器就具备这样的特点,因此被广泛地研究与应用。研究者们普遍认为,通过机器学习,可以为垃圾邮件的区分提供自动的、自适应的方法,以应对垃圾邮件的不断变化,提高邮件分类及过滤的质量<sup>[3,4]</sup>。

## 1 基于机器学习的邮件过滤系统

机器学习是人工智能的重要分支,近年来成为解决许多现实问题的热点方法。机器学习使用实例数据或过去的经验来训练计算机,以优化性能标准。当人们不能直接编写计算机程序解决给定问题,而是需要借助于实例数据或经验时,就需要学习。

收稿日期:2010-02-27;修回日期:2010-05-14

基金项目:国家自然科学基金(60805022);国家高技术研究发展计划(863)(2007AA01Z178);南京邮电大学青兰计划(NY206034)

作者简介:许棣华(1974-),女,江苏如皋人,讲师,博士研究生,CCF会员,研究方向为机器学习、软件测试与复用技术;王志坚,博士,教授,博士生导师,研究方向为软件自动化、软件构件复用。

对邮件过滤系统而言,机器学习方法能够从所提供的信息(Message)中提取知识(knowledge),并使用这些知识对新接收到的邮件进行分类<sup>[4]</sup>,过滤掉垃圾邮件,保留合法邮件。

最初提出将 Naïve Bayes 分类方法应用在邮件过滤中的是 M. Sahami 等人,他们考虑了在决策理论框架中的信息分类问题<sup>[5]</sup>。Bayes 框架最吸引人的的是对于来源不同的集成数据的适应性。Drucker 等人将支持向量机(Support Vector Machine)应用于邮件过滤系统<sup>[6]</sup>,他们使用词袋(Bag of Word)表示法,根据信息增益(Information Gain)来选择属性,并且发现对于不同的数据集和预处理过程,SVM 方法具有很好的鲁棒性。J. Clark 等人利用神经网络中的多层感知机对邮件进行分类,提出了 LINGER<sup>[7]</sup>。他们使用信息增益,使 LINGER 获得了很好的效果。并且断定,神经网络在邮件过滤与分类上是比较成功的,但在过滤器的可移植性还需要作出努力。J. Goodman 和 W. Yih 用的 logistic 回归模型对邮件进行过滤<sup>[8]</sup>。这种方法简单并且易于修正,并在很多语料中取得了较好的效果。G. Sakakis 研究了懒惰学习(lazy learning)算法,提出基于记忆(Memory-based)的邮件过滤系统<sup>[9]</sup>,并且用代价因子(cost-sensitive)来评价系统的性能。T. Oda 和 T. White 提出用人工免疫系统(Artificial Immune System)方法来对邮件进行过滤<sup>[10]</sup>,检测器表示为正则表达式,在信息被分析时用来作模式匹配。实验表明,在检测器较少时,这种方法是很有有效的。X. Carreras 和 L. Marquez 将 AdaBoost 算法的变体应用到邮件过滤中<sup>[11]</sup>,用决策树作为基本分类器。实验表明,如果有足够的训练次数,AdaBoost 优于 Naïve Bayes 和决策树。V. Zorkadis 等人从通信理论的概念中得到启发,用信息理论的度量对多分类器进行集成(ressemble)<sup>[12]</sup>,并表明这种过滤系统比投票式分类器更具有优势。

研究者们尝试用不同的机器学习算法对邮件进行过滤,取得了不小的成绩。但无论用何种算法,邮件过滤器的处理模式基本相同,处理过程如图 1 所示。在信息被用于分类以前,都需要进行合适的预处理过程,如:提取词汇、分词处理、表示等,然后用分类器对信息进行分类,过滤出合法邮件。

## 2 多任务学习

虽然目前大多数机器学习技术主要面对的是单个任务的学习场景,但在真实世界中多个相关(related)的学习任务往往是同时出现的。事实上,人类在学习如何完成一项任务时,往往都需要对一些与该任务相

关的任务进行学习并从中汲取经验。研究表明,机器学习的过程可以看作是对与问题相关的经验数据进行分析,从中归纳出反映问题本质的模型的过程。归纳偏置的作用就是用于指导学习算法如何在模型空间中进行搜索,搜索所得模型的性能优劣将直接受到归纳偏置的影响,而任何一个缺乏归纳偏置的学习系统都不可能进行有效的学习<sup>[13]</sup>。多任务学习恰恰为上述思想的实现提供了一条可行途径,即利用相关任务中所包含的有用信息,为所关注任务的学习提供更强的归纳偏置。能够利用多个相关任务之间蕴涵的有价值的信息来提高学习系统的性能,正是多任务学习(multitask learning)成为机器学习重要分支之一的主要原因。

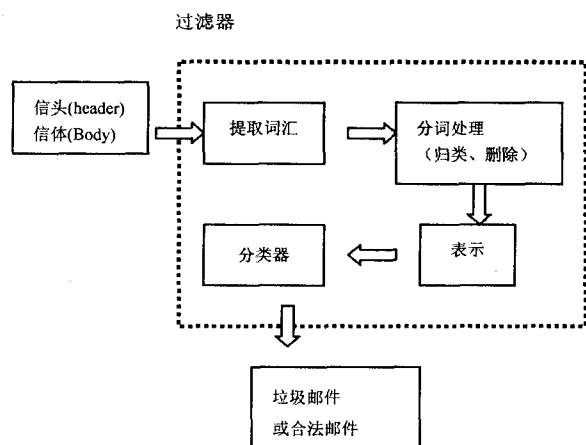


图1 邮件过滤器处理模型

### 2.1 任务相关性(task relevance)

任何多任务学习模式都是基于这样一个事实:这些任务在某种程度上是相关的。一般地,认为与相关性较大的任务一起学习,能获得更好的学习效果,而很多研究已经证实了这种猜想<sup>[14-16]</sup>。实验也同时表明,利用任务相关性进行学习,对样例较少的情况是非常有效的。文中基于这种多任务之间相关性假设,利用 EM 算法<sup>[17]</sup>,估计出任务之间相关系数,用这个相关系数决定各任务的参与程度。

在垃圾邮件过滤系统中,任务为判断某用户的邮件是否是垃圾邮件或合法邮件。由于用户的不同偏好和意图,即用户存在个性化差异,有理由认为这些任务是不同但彼此相关的。对于  $N$  个用户,假设有  $N$  个相关任务,设  $D_i = \{X_i^{\mu}, Y_i^{\mu}\}$  为第  $i$  个任务的样例,则整个数据集为  $D = \{D_i\}$ 。其中  $i = 1 \cdots N, \mu = 1 \cdots N_i, N_i$  为每个任务的样例数,  $X_i = \{x_{i1}, x_{i2}, \cdots, x_{id}\}$ ,  $d$  为  $X_i$  的属性个数,  $Y_i = \{-1, 1\}$ , 其中,  $-1$  表示垃圾邮件,  $1$  表示合法邮件。

用  $\alpha_{i,j}$  表示任务  $i$  与任务  $j$  的相关度(或任务  $j$  对任务  $i$  的参与度),  $\alpha_i = \{\alpha_{i,j}\}, j = 1 \cdots N$ , 有

$$\begin{cases} 0 \leq \alpha_{i,j} < 1, & \text{if } j \neq i \\ \alpha_{i,j} = 1, & \text{if } j = i \end{cases} \quad (1)$$

由文献[18],得到以下推导过程:设  $A$  为与  $N$  个任务的参数矩阵,  $A_i$  为第  $i$  个任务的参数向量,则有

$$P(D | A, \alpha_i) \propto \prod_{j=1}^K P(D_j | A_i)^{\alpha_{i,j}} \quad (2)$$

(2) 式的最大值问题转化为求  $\alpha_{i,j}$  的最大化问题。在求  $\alpha_{i,j}$  的过程中,使用 EM 算法。在 E-step, 计算  $E(\alpha_{i,j}^{(t+1)} | \alpha_i^{(t)})$ , 即由  $\alpha_i^{(t)}$  求得  $\alpha_{i,j}^{(t+1)}$  的数学期望。

$$E(\alpha_{i,j}^{(t+1)} | \alpha_i^{(t)}) = \int P(A_i | \alpha_i^{(t)}, D) * \log(\alpha_{i,j}^{(t+1)} | D, A_i) dA_i \propto E(A_i | \alpha_i^{(t)}, D) * [\log(\alpha_{i,j}^{(t+1)}) + \sum_{j=1}^n \alpha_{i,j}^{(t+1)} \log P(D_j | A_i)] \quad (3)$$

由(1)  $\alpha_{i,j}$  的特性,取

$$P(\alpha_i) = \prod_{j=1}^k \text{Gamma}(\alpha_{i,j} | a, b) \propto \alpha_{i,j}^a * e^{-b\alpha_{i,j}} \quad (4)$$

( $b \geq a > 0$ )

将(4)代入(3)可得:

$$E(\alpha_{i,j}^{(t+1)} | \alpha_i^{(t)}) \propto \sum a \log \alpha_{i,j}^{(t+1)} - b \alpha_{i,j}^{(t+1)} + \alpha_{i,j}^{(t+1)} E(A_i | \alpha_i^{(t)}, D) * \log P(D_j | A_i) \quad (5)$$

在 M-step, 求  $E(\alpha_{i,j}^{(t+1)} | \alpha_i^{(t)})$  的最大值, 即(5)式对  $\alpha_{i,j}$  的一阶导数为 0, 得到

$$\alpha_{i,j}^{(t+1)} = \frac{a}{b - E(A_i | \alpha_i^{(t)}, D) * \log P(D_j | A_i)} \quad (6)$$

由于  $P(A_i | \alpha_i^{(t)}, D)$  为高斯分布, 可将其简化为 Dirac delta 函数, 即  $\delta(A_i - A_i^{\text{MAP}})$ , 其中  $A_i^{\text{MAP}}$  为  $A_i$  的最大后验概率。最后得到

$$\alpha_{i,j} = \frac{a}{b - \log P(D_j | A_i^{\text{MAP}})} \quad (7)$$

从(7)式分析来看, 当  $\log P(D_j | A_i^{\text{MAP}})$  的值越大,  $\alpha_{i,j}$  越大, 即任务  $i$  与任务  $j$  的相关性越大, 这与我们的直觉也是一致的。

## 2.2 邮件过滤系统算法

总结以上推导, 得到邮件过滤系统算法:

- 1) 给定数据集  $D$ , 初始化  $\alpha_i^{(0)}$
- 2)  $A_i^{\text{MAP}} = \arg \max_{A_i} (\log(A_i | D, \alpha_i^{(t)})), t = 0, 1, 2, \dots$

...

- 3) 更新相关系数  $\alpha_{i,j}^{(t+1)} = \frac{a}{b - \log P(D_j | A_i^{\text{MAP}})}, j = 1 \dots N$

- 4) 重复步骤 2 和 3, 直到  $\alpha_{i,j}^{(t+1)}$  收敛;

- 5)  $F_i = \sum_{j \neq i} \alpha_{i,j} f_j$ ,  $f_i$  为单任务学习的得到的学习公式,  $i = 1 \dots N$ ;

6) 根据  $F_i$  的值判断用户  $i$  的某封邮件是否为垃圾邮件, 若  $F_i > 0$  则被判为合法邮件, 否则被判为垃圾邮件。

## 3 实验结果与分析

考虑到本实验的基本假设是任务之间具有较显著的相关性, 文中使用的英文公开语料为 Lingspam。Lingspam 由提供者收到的垃圾邮件和来自于语言学家列表 (Linguist list) 的非垃圾邮件组成, 它提供了四种方式的语料, 分别为: bare、lemm、lemm-stop、stop。每种方式的语料均由 2893 封邮件组成, 其中垃圾邮件 481 封, 合法邮件 2412 封。为减少无用信息的干扰, 实验中采用 lemm-stop 形式的语料库, 语料由 10 部分组成, 每部分约 290 封邮件, 9 个部分作为训练集, 剩下的 1 个部分作为测试集, 如此交叉做 10 次取平均值。特征选择采用信息增益 (Information Gain) 方法: 将训练集中所有词按照信息增益计算值的大小排序, 选取排在前面 80% 的词作为特征集。算法性能主要由 Recall、Precision 和 Accuracy 这三个参数来衡量。其中 Recall 为召回率, 即垃圾邮件检出率; Precision 为精确率, 即垃圾邮件检对率; Accuracy 为正确率, 即所有邮件的检对率。实验结果如表 1 所示。

表 1 Lingspam 语料邮件过滤结果数据表

$a$	$b$	Recall (%)	Precision (%)	Accuracy (%)
1.0	1.0	79.21	94.32	93.12
0.5	1.0	82.24	98.30	94.87
0.1	1.0	80.34	97.78	92.45

文中对中文邮件语料也进行了实验, 中文邮件语料来源于本实验室人员 (12 人) 的私人邮件 564 封, 其中合法邮件 397 封, 垃圾邮件 167 封。对邮件样本进行的预处理 (包括去停用词、词汇还原) 是通过 Stanford NLP 的 stanford parser<sup>[19]</sup> 进行的。利用信息增益对排在前 80% 的词作为特征集。算法性能仍然由 Recall、Precision 和 Accuracy 这三个参数来衡量。实验结果如表 2 所示。

表 2 中文邮件语料过滤结果数据表

$a$	$b$	Recall (%)	Precision (%)	Accuracy (%)
1.0	1.0	79.78	95.24	90.34
0.5	1.0	79.56	98.78	90.54
0.1	1.0	78.34	97.58	89.92

从实验结果看, 邮件过滤器对中文、英文语料同样是有有效的。过滤器的好坏直接与  $a$ 、 $b$  的取值及单个

任务学习的算法有关。这是由于实验中的  $\alpha_i$  的先验概率用 Gamma 函数近似, 由于 Gamma 分布是单峰的, 且峰值是  $a/b$ , 当  $b > a$  时, 峰值在 0 和 1 之间。虽然从实验结果来看, 似乎  $a/b$  的在接近 0.5 时更为理想, 但并不意味着, 对所有数据集或语料都如此。 $a$ 、 $b$  的值应该在交叉验证时确定。另外, 本实验中, 对单个任务的学习是用 Naïve Bayes, 如果使用决策树算法 C4.5, 则能够得到更好的效果。结果如表 3 所示(取  $b = 1.0$ ,  $a = 0.5$ , 取 10 次交叉验证的平均值, 使用的语料如前所提及)。

表 3 单任务学习采用 C4.5 的结果数据表

	Recall(%)	Precision(%)	Accuracy(%)
Lingspam	81.24	98.76	94.37
中文语料	79.56	99.12	93.21

从实验结果数据来看, 过滤器性能比用表 1、表 2 要好些。估计是因为 C4.5 算法在邮件分类问题上好于 Naïve Bayes 算法, 可见, 单任务的学习的优劣也决定着最终系统性能的好坏。

#### 4 结束语

在邮件过滤系统中运用机器学习的方法已经被广泛研究和应用, 如 Naïve Bayes、支持向量机、Logistic 回归、神经网络、人工免疫系统等等。多任务学习在提高学习系统泛化能力方面具有显著的能力, 而将多任务学习模式运用到邮件过滤系统还较少。

文中将不同邮件用户当作不同任务处理, 利用多任务的性质, 估计出任务之间的相关系数, 以决定在学习过程中任务之间的参与程度, 并以此作为判断垃圾邮件与否的标准。实验表明, 这种基于多任务的邮件过滤系统是可靠和有效的, 特别是对于同一个邮件列表的用户邮件。过滤器的好坏取决于相关系数的先验概率的参数取值, 以及单任务学习的学习算法。

进一步工作, 重点将放在研究以下问题: (1) 如何在用多任务学习中解决大规模、分散用户邮件问题; (2) 如何在多任务学习中体现邮件用户偏好问题。

#### 参考文献:

- [1] 王 斌, 潘文锋. 基于内容的垃圾邮件过滤技术综述[J]. 中文信息学报, 2005(5): 1-10.
- [2] Mitchell T M. Machine Learning[M]. New York: McGraw-Hill, 1997.
- [3] Guzella T S, Caminhas W M. A review of machine learning approaches to Spam filtering[J]. Expert Systems with Applications, 2009, 36(7): 10206-10222.

- [4] Caruana R. Multitask learning[J]. Machine Learning, 1997, 28(1): 41-75.
- [5] Sahami M, Dumains S, Heckerman D, et al. A Bayesian approach to filtering junk E-mail[R]. [s.l.]: AAI Press, 1998.
- [6] Drucker H, Wu D, Vapnik V N. Support vector machines for spam categorization[J]. IEEE Transactions on Neural Networks, 1999, 10(5): 1048-1054.
- [7] Clark J, Koprinska I, Poon J. A neural network based approach to automated e-mail classification[C]// In Proc of the IEEE/WIC int conf on web intell. [s.l.]: [s.n.], 2003.
- [8] Goodman J, Yih W. Online discriminative spam filter training [C]// In Proc of the third conf on email and anti-spam. [s.l.]: [s.n.], 2006.
- [9] Sakkis G, Androustopoulos I, Paliouras G, et al. A memory-based approach to anti-spam filtering for mailing lists[J]. Information Retrieval, 2003, 6(1): 49-73.
- [10] Oda T, White T. Developing an immunity to spam[J]. Lecture Notes in Computer Science, 2003, 2723: 231-242.
- [11] Carreras X, Marques L. Boosting trees for anti-spam email filtering[C]// In Proc of the fourth int conf on recent adv in nat lang proc. [s.l.]: [s.n.], 2001.
- [12] Zorkadis V, Karras D A, Panayoto M. Efficient information theoretic strategies for classifier combination, feature extraction and performance evaluation in improving false positives and false negatives for spam e-mail filtering[J]. Neural Networks, 2005, 18(5-6): 799-807.
- [13] Mitchell T M. The need for biases in learning generalizations [C]// In: Shavlik J W, Dietterich T G, eds. Readings in Machine Learning. San Mateo, CA: Morgan Kaufmann, 1990: 184-191.
- [14] Ben-David S, Schuller-Borbely R. A notion of task relatedness yielding provable multiple-task learning guarantees[J]. Machine Learning, 2008, 73: 273-287.
- [15] Bakker B, Heskes T. Task clustering and gating for Bayesian multitask learning[J]. Journal of Machine Learning Research, 2003(4): 83-99.
- [16] Juba B. Estimating relatedness via data compression[C]// In Proceedings of the 23rd International Conference on Machine Learning. [s.l.]: [s.n.], 2006.
- [17] Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society B, 1977, 39: 1-38.
- [18] Fang J, Ji S, Xue Y, et al. Multitask Classification by Learning the Task Relevance[J]. IEEE Signal Processing Letters, 2008, 15: 593-596.
- [19] The Stanford Natural Language Processing Group. stanford parser(OnLine)[EB/OL]. 2009. <http://nlp.stanford.edu:8080/parser/index.jsp>.