

文章编号:1001-9081(2009)02-0374-03

基于 CART 算法的垃圾邮件过滤模型设计与实现

孔 颖¹, 裘彬强¹, 徐从富²

(1. 浙江科技学院 信息与电子工程学院, 杭州 310023; 2. 浙江大学 计算机学院, 杭州 310027)

(kongying - 888@163. com)

摘要: 介绍分类与回归树(CART)算法在垃圾邮件过滤中的应用。首先对样本邮件进行文本预处理, 并对正常邮件和垃圾邮件训练集进行训练, 用 CART 算法建立单分类器模型, 随后又采用 Boosting 思想组合 CART 算法建立多分类器模型。对比实验表明, 基于 CART 算法的多分类器模型效果更好。

关键词: 垃圾邮件过滤; 分类与回归树算法; Boosting 算法

中图分类号: TP391 **文献标志码:**A

Design and realization of spam filtering model based on CART algorithm

KONG Ying¹, QIU Bin-qiang¹, XU Cong-fu²

(1. School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou Zhejiang 310023, China;

2. College of Computer Science, Zhejiang University, Hangzhou Zhejiang 310027, China)

Abstract: The Classification and Regression Tree (CART) algorithm was introduced, and the application of the CART algorithm in spam filtering was discussed. Firstly, text messages of email samples were pre-processed, regular email and spam training sets were trained to establish the single classifier with the CART algorithm. Then a boosting based model which was combined with multiple CART classifiers was established. The comparison test results show that the multiple CART classifiers based on boosting has achieved better results.

Key words: spam filtering; CART algorithm; Boosting algorithm

0 引言

随着 Internet 网的广泛普及, 电子邮件已经成为人们联系沟通的重要手段, 人们在享受通信便捷的同时, 也深受垃圾邮件泛滥所带来的负面影响。如何有效地遏制垃圾邮件迫在眉睫。90 年代, 基于机器学习的自动邮件文本分类方法逐渐代替了知识工程的分类方法, 成为文本分类的主流技术。基于机器学习^[6]的自动分类方法有最近邻分类、回归模型、CART 算法、推导规则、贝叶斯分类、神经网络、支持向量机以及相关反馈。组合分类器方法是近年来流行的一种分类方法, 它是将多个分类器的判定结果合并为一个分类器的方法。即: 对于需要专家介入的任务, n 个独立的专家的判断经过适当归并, 比单个人作出的判断要好。

Boosting 算法是一种特殊的组合分类器方法, 其组合分类器中的 n 个分类器 s_1, s_2, \dots, s_n (称为弱假设) 是由相同的学习算法(称为弱学习器)形成的, 并且 n 个弱学习器都采用相同的文本表示方法。该组合分类器中的 n 个弱学习器是一个接一个序列式地进行训练。即弱假设 s_i 需要考虑弱假设 s_1, \dots, s_{i-1} 在训练集上的分类效果, 并重点处理 s_1, \dots, s_{i-1} 分类效果不佳的文本。最后采用归并规则将 n 个分类器进行合并形成终假设 S 。基于 Boosting 分类算法的分类器较其他分类器有更好的分类精度。一个典型的邮件自动分类系统的功能流程如图 1 所示。

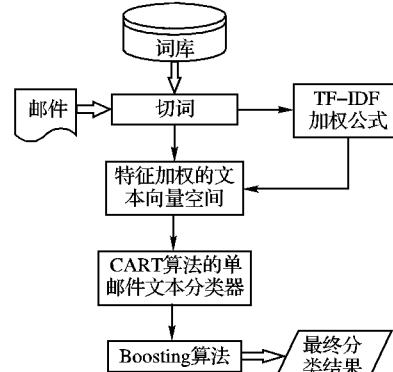


图 1 邮件分类系统功能流程

1 邮件特征提取

1.1 邮件预处理

在对邮件进行分类之前, 首先要对待判断的邮件和已知分类的邮件进行一系列的预处理, 转化为计算机易于处理的向量形式。本文采用文献[1]中提出的向量空间模型。

首先对邮件的内容进行分词处理: 对于外文邮件, 采用空格作为单词之间的分隔符进行分词。对于中文文本, 除了标点符号, 词语之间没有明显的分隔符。中文文本分词常用的方法有基于字符串匹配的分词方法, 基于理解的分词方法和基于统计的分词方法等。本文采用的是基于词频统计的分词方法。分词后, 得到一系列文档中词组成的表, 称为邮件特征词典。在邮件特征词典中将停用词(各类邮件中都频繁出现的词)、低频词(只在极少数的邮件中出现的词)从邮件特征

收稿日期:2008-08-27;修回日期:2008-10-17。 基金项目:国家 863 计划项目(2007AA01Z197)。

作者简介: 孔颖(1980-), 女, 浙江杭州人, 讲师, 硕士, 主要研究方向: 信息处理、图像技术、数据融合、机器学习; 裘彬强(1980-), 男, 浙江天台人, 助理研究员, 硕士, 主要研究方向: 垃圾邮件处理; 徐从富(1969-), 男, 浙江衢州人, 副教授, 主要研究方向: 人工智能、机器学习、文本分类、信息融合。

词典中去除并进行数字、人名等的合并以减少邮件特征词典可能带给特征向量的数据噪声,缩小特征词典的规模,提高邮件分类器的训练和分类效率,重新得到邮件特征词典。

1.2 邮件内容特征表示

目前,文本表示主要采用向量空间模型^[5](Vector Space Model, VSM)。因此本文在邮件内容特征表示时也用VSM。VSM是将每一个文本表示为一个向量: $\mathbf{d} = (w_1, w_2, \dots, w_n)$, 其中 w_i 为第 i 个特征项在邮件 \mathbf{d} 中的权重。在计算权重 w_i 时, 使用目前最广泛使用的TF-IDF加权算法。

$$W(t, \vec{d}) = tf(t, \vec{d}) \times \log\left(\frac{N}{n_t}\right)$$

其中, $W(t, \vec{d})$ 为特征项 t 在邮件 d 中的权重, 而 $tf(t, \vec{d})$ 为特征项 t 在邮件内容中的词频, N 为训练文本的总数, n_t 为训练邮件集中出现特征项 t 的邮件数。用TF-IDF算法来计算特征词的权重值是表示当一个词在这篇邮件中出现的频率越高, 同时在其他文档中出现的次数越少, 则表明该词对于表示这篇文档的区分能力越强, 所以其权重值就应该越大。将所有词的权值排序, 根据需要选择特征项。

为消除文档长度不一对文本表示方式的可能影响, 往往需要对加权后的向量进行规范化处理, 使权值落在 $[0, 1]$ 中。即:

$$W(t, \vec{d})' = \frac{W(t, \vec{d})}{\sqrt{\sum_{i=1}^n W_i^2}}$$

2 基于CART算法的邮件单分类器设计

2.1 CART算法^[2]

回归分类树(Classification and Regression Trees, CART)算法是分类数据挖掘算法的一种。它描述给定预测向量值 \mathbf{X} 后, 变量 \mathbf{Y} 条件分布的一个灵活的方法。该模型使用了二叉树将预测空间递归划分为若干子集, \mathbf{Y} 在这些子集上的分布是连续均匀的。树中的叶节点对应着划分的不同区域, 划分是由与每个内部节点相关的分支规则(Splitting Rules)确定的。通过从树根到叶节点移动, 一个预测样本被赋予一个唯一的叶节点, \mathbf{Y} 在该节点上的条件分布也被确定。CART模型最早由 Breiman等人提出并已在统计学领域普遍应用。

CART是一种有监督学习算法, 即用户在使用CART进行预测之前, 必须首先提供一个学习样本集对CART进行构建和评估, 然后才能使用。CART使用如下结构的学习样本集:

$$L = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m, \mathbf{Y}\}$$

$$\mathbf{X}_1 = (x_{11}, x_{12}, \dots, x_{1t}), \dots, (x_{m1}, x_{m2}, \dots, x_{mt})$$

$$\mathbf{Y} = (y_1, y_2, \dots, y_k)$$

其中, $\mathbf{X}_1 \sim \mathbf{X}_m$ 称为属性向量(Attribute Vectors), 其属性可以是有序的也可以是离散的; \mathbf{Y} 称为标签向量(Label Vectors), 其属性可以是有序的也可以是离散的。当 \mathbf{Y} 是有序的数量值时, 称为回归树; 当 \mathbf{Y} 是离散值时, 称为分类树。

根据给定的样本集 L 构建分类树由以下3步组成:

步骤1 使用 L 构建树 T_{\max} , 使得 T_{\max} 中每一个叶节点要么很小(节点内部所含样本个数小于给定值 N_{\min}); 要么是纯节点(节点内部样本的 \mathbf{Y} 属于同一个类); 要么只有唯一属性向量作为分支选择。

其核心算法是确定一个决策树分枝准则, 该准则涉及到两方面问题: 1) 如何从众多的输入变量中选择一个最佳的分组变量; 2) 如何从分组变量的众多取值中找到一个最佳的分割阈值。

生成原始树首先引入差异系数的概念。这个系数用于测度每一个节点内 n ($n \geq 2$) 个类样本的差异程度。在这里我们

采用的是基尼系数 $Gini$, 如果集合 T 包含 N 个类别的记录, 那么 Gini 指数就是:

$$Gini(T) = 1 - \sum_{j=1}^N p_j^2$$

如果集合 T 在 X 的条件下分成两部分 N_1 和 N_2 , 那么这个分割的 Gini 指数就是:

$$Gini_{split(x)}(T) = \frac{N_1}{N} Gini(T_1) + \frac{N_2}{N} Gini(T_2)$$

步骤2 使用修剪(Pruning)算法构建一个有限的递减(节点数目)有序子树序列。

当判定树创建时, 由于数据中的噪声和孤立点, 许多分枝反映的是训练数据中的异常。剪枝方法处理这种过分适应的数据问题。通常, 这种方法使用统计度量, 剪去最不可靠的分枝, 这将导致较快的分类, 提高树独立于测试数据正确的分类的能力。

这里采取后剪枝(postpruning)方法, 它由“完全生长”的树剪去分枝。通过删除节点的分枝, 剪掉树叶节点, 使得剪枝后的树能够对新数据进行更准确的分类。在删减中我们采用CART系统的成本-复杂度最小(Minimal cost-complexity pruning)原则, 其测度表示为:

$$R_a(T) = R(T) + a \times |T|$$

其中, $|T|$ 为该树的叶节点的个数; a 为复杂度参数, $R_a(T)$ 理解为该树加权错分率与对复杂度处罚值之和的复合成本。当 a 一定时, 由 T_{\max} 删减而生成的两个具有相同 $R(T)$ 值的树, 叶节点越多, 则树的复杂度越高 $R_a(T)$ 越大, 其可取性也就越小。按照赤池信息准则(Akaike Information Criteria, AIC)原则, $a = 2(k-1)$, k 为分类数, 在两分类问题中取 $a = 2$ 。

步骤3 利用文献[9]中的评估(Estimate)算法从步骤2产生的子树序列中选出一棵最优树作为最终的决策树。

2.2 CART算法在垃圾邮件过滤中的适用性分析

由于垃圾邮件过滤自身的特点, 给模型提出了较高的要求。其中包括整体样本的动态性要求、个体样本的时间序列动态性要求, 解决信息缺失的要求。CART算法模型简单直观, 误差率非常低, 特别是当处理许多指标组成的复杂数据时, 用CART算法产生的分类树的误差率要比通常的参数方法低得多; 同时分类树又非常稳健, 受少数异常数据的影响非常小, 处理特别数据的效果非常突出。CART算法模型可以运用于多指标海量数据的处理, 因此在解决垃圾邮件过滤问题时有很大的优势。

CART算法模型采用非参数估计的形式, 在计算的过程中自动选取变量避免了事先选好变量所可能带来的一些主观因素的影响。CART算法模型所应用的变量个数可以是很大的, 树在选择分割的过程中, 自动选择对正确分类提供信息量最大的分割, 排除那些有缺失数据的变量; 也可以自动进行样本的调整, 可以排除一些有相互作用的变量, 提高分类的准确率。CART算法模型具有灵活性, 反映在分类树不仅可以用于连续变量和类别变量上, 而且可以用于两者的任意组合, 也可以用变量的线性组合来分割树点。

2.3 CART算法邮件分类模型评价体系

模型评价是评价模型分类错误概率情况, 衡量分类器有效性, 决定模型是否满足应用要求或达到预期作用的重要手段。评价一个解决分类问题的模型是否适用的一个直接手段就是看它的错分率, 即错误分类数与总记录数的比值。我们必须测量用来建立模型和在建立模型过程中没有用到的记录组成的测试样本, 选择最具有普遍意义的而不是最适合训练样本的模型。

考虑 N 封待测试邮件 (N_s 封垃圾邮件和 N_b 封正常邮件), $N = N_s + N_b$, 在 CART 算法邮件分类模型中, 垃圾邮件被分类器正确判定的有 A 封, 误判的有 B 封; 正常邮件被分类器正确判定的有 C 封, 误判的有 D 封, 显然 $N_s = A + B, N_b = C + D$ 。根据定义, 有下列各式成立:

$$ham = \frac{C}{(C + D)} \times 100\%$$

$$spam = \frac{A}{(A + B)} \times 100\%$$

$$Accuracy = \frac{(A + C)}{N} \times 100\%$$

CART 算法的垃圾邮件过滤应用模型采用 ham、spam 和 Accuracy 等传统分类指标来分析特征提取和特征值计算方法、词权重的定义和训练模型的选择对邮件分类模型的影响。

同时要考虑错分成本。预测错误可分为两类:第一类错误是指将垃圾邮件预测为正常邮件, 第二类错误为将正常邮件预测为垃圾邮件, 研究表明在垃圾邮件过滤管理中, 每个用户宁愿收到垃圾邮件, 不愿意一封正常的邮件被当成垃圾邮件而导致无法收到。所以第二类错误的成本要远远高于第一类错误, 因此应该做到在保证模型预测精度的情况下, 尽量减少第二类错误发生。

2.4 CART 算法模型在垃圾邮件过滤中的应用分析

在进行垃圾邮件过滤时, 各服务器根据自身情况选择合适的指标建立指标体系, 运用专家评估法和统计模拟法, 通过利用一定数目的指标进行模拟, 确定各级指标的权重, 并得出标准分值, 就可以建立垃圾邮件过滤系统, 科学、客观完整和快速地对邮件进行分类预测, 做出是否是垃圾邮件的决策。

对不同规模大小的样本数据, 采用不同的数据建模和检验方法。在邮件样本充分反映总体邮件的特征并且邮件样本数据比较多, 所以采用独立估测方法, 即数据分为两个部分, 一部分用来建立模型, 一部分用于检验模型效果, 这种方法可以产生错分率的无偏估计。

2.5 基于 Boosting 的邮件多分类器模型设计

单分类器往往出现偏好的现象, 即模型不稳定。因此, 大多自动分类系统都采用多分类器模型来进行文本分类。在单分类器功能设计完善的基础上, 多分类器模型主要考虑如何综合处理组成多分类器模型中各个单分类器的分类结果。在对各种分类方法进行对比的基础上, 本文使用基于 Boosting 算法多分类器模型, 具体算法如图 2 所示。

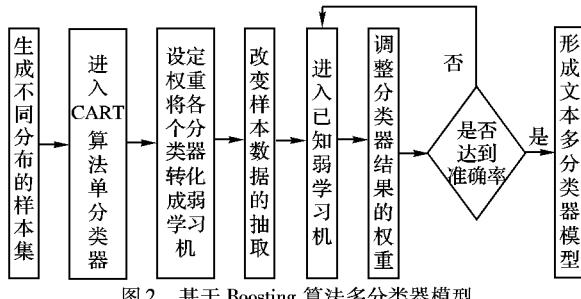


图 2 基于 Boosting 算法多分类器模型

3 实验结果分析

本文样本是取自 2008 年全国搜索引擎和网上信息挖掘学术研讨会(SEWM2008)的比赛中第二阶段的比赛数据集作为评测数据。邮件样本总共 30 000 封, 其中正常邮件有 7 676 封, 垃圾邮件有 22 324 封。每封邮件有 95 949 个特征项反映了邮件的内容信息。随机抽取样本的 80% 的数据用做训练样本, 其余 20% 数据用于分类器准确率的检验样本。

1) 单分类器训练结果。使用训练样本对单分类器进行训

练, 然后用检验样本计算分类器的准确率, 结果如表 1 所示。

表 1 单分类器样本预测成功率

样本	邮件总数	判断正确数	判断错误数	判断正确率
学习	垃圾	6 263	6 044	96.51
样本	正常	17 737	16 875	95.14
测试	垃圾	1 413	1 330	94.13
样本	正常	4 587	4 311	93.99

其中, 学习样本 average 为 95.83%, Accurate 为 95.50%, 测试样本 average 为 94.04%, Accuracy 为 94.02%。

2) 多分类器训练结果。基于 Boosting 的多分类器分类的准确率, 如表 2 所示。

表 2 多分类器测试样本预测成功率

邮件	邮件总数	判断正确数	判断错误数	判断正确率
垃圾	1 401	1 364	141	97.33
正常	4 599	4 458	37	96.94

其中, 测试样本 average 为 97.14%, Accuracy 为 97.03%。

3) 实验结果对比。基于 CART 算法的单分类器与基于 Boosting 的多分类器的分类结果对比, 如表 3 所示。

表 3 各实验结果比较 %

分类器	Ham	Spam	Average	Accuracy
CART 算法的单分类器	94.32	93.93	94.20	94.00
Boosting 的多分类器	97.30	97.00	97.10	96.91

由表 3 可知, 基于 Boosting 的多分类器不管从正常邮件分类、垃圾邮件分类还是整体分类都具有较高的准确率。此外, 垃圾邮件预测为正常邮件的概率大于正常邮件预测为垃圾邮件的概率, 符合上述提到的错分成本。

4 结语

本文将已在其他领域得到广泛应用的 CART 算法引入垃圾邮件过滤系统——基于 CART 算法建立单分类器模型, 在此基础上, 提出基于 Boosting 的多分类器模型。实验结果表明, 基于 Boosting 的多分类器模型取得更好的垃圾邮件过滤效果。

参考文献:

- [1] SALTON G, WONG A, YANG C. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613 - 620.
- [2] LEWIS R J. An introduction to classification and regression tree (CART) Analysis[EB/OL]. [2007-09-01]. <http://www.saem.org/download/lewis1.pdf>.
- [3] ZEITOUN I K, YEH L. Join indices as a tool for spatial data mining [C] // International Workshop on Temporal, Spatial and Spatio-Temporal Data Mining, Lecture Notes in Artificial Intelligence. France: Springer Press, 2007: 102 - 114.
- [4] 黄萱菁, 夏迎炬, 吴立德. 基于向量空间模型的文本过滤系统 [J]. 软件学报, 2003, 14(3): 12 - 16.
- [5] 庞剑锋, 卜东波, 白硕. 基于向量空间模型的文本自动分类系统的研究与实现[J]. 计算机应用研究, 2000, 18 (9) : 23 - 26.
- [6] RYSZARD S, BRATRO I, KUBAT M. 机器学习与数据挖掘方法和应用[M]. 朱明, 译. 北京: 电子工业出版社, 2004.
- [7] 肖江, 张亚非. Boosting 算法在文本自动分类中的应用[J]. 解放军理工大学学报: 自然科学版, 2003, 14(2): 20 - 23.
- [8] 陈勇, 李卓桓. 反垃圾邮件完全手册[M]. 北京: 清华大学出版社, 2006.
- [9] 于维英. 分类与回归树算法[J]. 宁波职业技术学院学报, 2007, 8 (3): 1 - 3.