CN 43-1258/TP
ISSN 1007-130X

计算机工程与科学
Computer Engineering & Science

第 39 卷第 3 期 2017 年 3 月
Vol. 39, No. 3, Mar. 2017

文章编号:1007-130X(2017)03-0572-08

基于 PAC-Bayes 理论的 Web 文档数据质量评估方法*

汤　莉,何　丽

(天津财经大学理工学院信息科学与技术系,天津 300222)

摘　要:为了更好地评估 Web 文档数据质量,提出一种基于 PAC-Bayes 理论的 Web 文档质量评估指标体系和评估方法。PAC-Bayes 理论融合了 PAC 理论和贝叶斯定理,在充分利用样本先验信息的基础上,推导出了最紧的泛化风险边界,用于衡量学习算法的泛化性能。首先阐述了文档数据质量评估的研究现状,介绍了 PAC-Bayes 理论框架及其在支持向量机上的应用;其次提出一种基于 PAC-Bayes 理论的 Web 文档数据质量评估方法(DQAPB),将 SVM 算法及其 PAC-Bayes 边界应用于 Web 文档的质量评价中,并构建了基于 PAC-Bayes 理论的 Web 文档质量评估指标体系;最后采用 Wikipedia 文档进行实验,实验结果表明该方法具有简便快速、稳定性和鲁棒性较强的优点。

关键词:PAC-Bayes 边界;支持向量机;泛化能力;数据质量评估

中图分类号:TP181　　**文献标志码**:A

doi:10.3969/j.issn.1007-130X.2017.03.025

A data quality assessment method of Web articles based on PAC-Bayes theory

TANG Li, HE Li

(Department of Information Science and Technology, School of Science and Technology, Tianjin University of Finance and Economics, Tianjin 300222, China)

Abstract: We propose an assessment index system and a method based on the PAC-Bayes theory for better data quality assessment of Web articles. Making full use of prior information of samples, the PAC-Bayes theory integrates the theories of Probably Approximately Correct and the Bayesian paradigm, and derives the tightest generalization bounds to assess the generalization capability of classifiers. We analyze the research status of data quality assessment of articles in detail, and then introduce the theoretical framework of the PAC-Bayes theory and its application for SVM. Furthermore, we propose a method for data quality assessment of Web articles based on the PAC-Bayes theory (DQAPB), and apply the SVM algorithm and its PAC-Bayes bound to the data quality assessment of Web articles. Moreover, we establish a quality assessment index system of Web articles based on the PAC-Bayes theory. Experiments on Wikipedia document show that the proposed method is simple and fast with strong stability and robustness.

Key words: PAC-Bayes bound; support vector machine (SVM); generalization capability; data quality assessment

* 收稿日期:2015-05-15;修回日期:2015-11-26
基金项目:天津市自然科学基金(15JCYBJC16000);教育部人文社会科学研究一般项目(14YJA630025);天津市社会科学基金(TJYY15-017);国家自然科学基金(61502331)
通信地址:300222 天津市河西区珠江道 25 号天津财经大学理工学院信息科学与技术系
Address: Department of Information Science and Technology, School of Science and Technology, Tianjin University of Finance and Economics, 25 Zhujiang Rd, Hexi District, Tianjin 300222, P. R. China

1　引言

随着互联网和云计算的快速发展,数据呈现爆炸性地增长,这预示着大数据时代已经到来。大数据具有的主要特点是:数据量巨大;数据结构复杂且形式多样;数据的增长速度和变化速度飞快等[1],因此,在大数据环境下,人们想要利用大数据来做出准确分析和正确决策,就必须从海量大数据中挖掘出有用的信息和高质量的数据。对于大数据来说,数据质量的管理显得尤为重要。数据质量评估是数据质量管理的重要环节,也是确保数据质量的首要前提和基础。

数据质量被认为可分解成若干个具体的数据质量维度来衡量,主要包括:正确性、准确性、一致性、完整性和新鲜性等[2]。数据质量评估的本质就是从这些质量维度对数据进行评价和分析。关于数据质量的评估,研究人员进行了很多深入研究[3-5]。特别地,由于互联网的开放性及其缺乏有效的监督机制和规范的约束,对 Web 文档的数据质量进行评价,具有重要的研究意义。Web 文档作为一种非结构化数据,其数据质量的评估已成为目前热点研究问题[6-14]。

针对 Web 文档的数据质量评估问题,本文提出以"计算学习理论"视角,利用学习算法的计算复杂性理论,来评估 Web 文档数据质量,进一步提出基于"学习算法的泛化误差界"理论—即 PAC-Bayes 理论,来评估和度量 Web 文档的数据质量。

"计算学习理论"源于 Valiant[15] 在 1984 年提出的"概率近似正确性学习"PAC learning(Probably Approximately Correct learning)理论。PAC 理论融合了计算复杂性理论和概率理论,来评价机器学习算法的性能,是"计算学习理论"中评价学习问题的第一个理论框架。PAC 理论确定了以高概率选择具有较少损失的泛化函数作为学习目标,却不能估计出学习算法的泛化误差边界。虽然 VC 维理论可以给定学习算法的误差边界,但是无法充分利用样本的先验知识。由此,1999 年 McAllester[16] 提出 PAC-Bayes 理论,该理论将 PAC 性能度量和贝叶斯定理有效结合起来,既能有效利用样本的先验知识,又能为衡量学习算法的泛化性能给出一种最紧的风险边界—即 PAC-Bayes 边界。该边界作为"Occam's razor"边界的一种推广,最初应用于线性分类器[17-19],经过研究人员的发展,它已经成功应用于多类分类、回归、聚类等各类学习问题的泛化性能研究[20-22]。

本文的组织结构如下:首先探讨了数据质量评估的研究现状,并阐述了 PAC-Bayes 理论框架及其在支持向量机 SVM(Support Vector Machine)上的应用;其次建立了 PAC-Bayes 理论应用于数据质量评估的模型,提出一种基于 PAC-Bayes 理论的 Web 文档数据质量评估方法 DQAPB(Data Quality Assessment of web article based on PAC-Bayes theory);再次构建了一种基于 PAC-Bayes 理论的评估指标体系,利用 PAC-Bayes 边界 PBB(PAC-Bayes bound)、敏感性 SEN(SENsitivity)、特异性 SPC(SPeCificity)和正确率 ACC(ACCuracy)等性能指标来度量和评估数据质量;最后以 Wikipedia 文档为例进行实验,结果表明该方法简便快速,并具有较强的稳定性和鲁棒性。

本文的创新特色在于,构建了一种基于 PAC-Bayes 理论的评估指标体系,同时提出一种基于 PAC-Bayes 理论的 Web 文档数据质量评估方法 DQAPB,该方法由三个步骤构成:(1)根据目标文档,进行特征提取和归一化处理;(2)采用 SVM 算法进行文档分类,并计算相关的性能指标;(3)将 PAC-Bayes 理论应用于 Web 文档的数据质量评估。

2　文档数据质量评估的相关研究

数据质量评估的研究主要包括构建评估指标体系和确定评估方法的两个方面内容。杨青云等[3]提出了一种数据质量评估模型,并阐述了该模型的构造技术和计算方法。在数据质量评价领域中,通常认为数据质量是一个层次分类的概念[2],采用的评估方法是层次分析法。严浩等[4]以正确性和准确性等质量维度构建了数据质量评估指标体系,通过引入二次变权思想,提出一种改进的层次分析方法来综合评估数据质量。杨栋枢等[5]基于熵权与层次分析法构建了一种数据质量组合权重的评价模型。然而,由于不同应用领域中的数据采用的评价指标也不尽相同,因此目前尚未形成统一的评估指标体系。

针对在线 Web 文档的数据质量评价,研究人员进行了很多相关的研究。韩京宇等[6]提出一种基于模拟退火的 Web 文档内容数据质量评估 QASA(Quality Assessment based on Simulated Annealing)方法,来满足实时响应的要求,实现在线评价文档质量;同时提出一种基于事实的质量评

估方法 FQA(Fact-based Quality Assessment)[7],基于事实内涵来量化数据质量维度。这两种方法都是从文档中提取信息,根据文档内容事实进行数据质量评估。

此外,研究人员对在线文档数据质量的评价方法,还包括基于文档的基本特征、基于文档的编辑历史、基于用户交互三个方面。文献[8]采用多种机器学习算法,根据文本长度来确定 Wikipedia 文档的数据质量。文献[9]采用一种基于 SPEA2 多目标遗传算法的特征选择方法,实现了根据文档最少的特征信息较好地评价协同网站的文档质量。文献[10]采用 SVM 方法,根据文档的三个特征即文本、评论和网络,来评估文档的数据质量。文献[11]根据在线文档修订的动态性质,并以文档的修订历史来构建文档质量的评估模型。文献[12]基于文档质量和其作者的学术权威性之间的依赖关系,提出一种文档质量评价模型,并从文档的编辑历史及其作者学术权威性的角度来评价文档质量。文献[13]从文档的编辑历史中提取文本和作者信息,计算作者信誉分数,并以此来调整文档文本质量的分数,从而改进了文档质量评估的方法。文献[14]基于用户的交互模式来判断文档的数据质量。这些研究方法主要是通过提取文档某些方面的特征来进行建模,实现质量评估。

本文构建了一种基于 PAC-Bayes 理论的评估指标体系,同时提出一种基于 PAC-Bayes 理论的 Web 文档数据质量的评估方法,该方法具有简便快速、稳定性和鲁棒性强的优点。

3 PAC-Bayes 理论框架

3.1 PAC-Bayes 基本理论

PAC-Bayes 理论融合了 PAC 学习理论和 Bayes 定理两者优势,下面对该理论做简要介绍。

假设样本空间为 D,样本集为 $S=\{x_1,x_2,\cdots,x_n \mid x_i \in S_i\}$,对应的类标签为 $y_i \in \{-1,1\}$,分类器为 c 及其分布为 Q,得出以下定义:

定义 1 分类器的真实误差[17],指的是 c 从 D 中随机分类样本对 (x,y) 的误分概率。

$$C_D \equiv Pr_{(x,y)\in D}(c(x) \neq y) \tag{1}$$

定义 2 分类器的经验误差[17],指的是 c 在包含 m 个样本的样本集 S 上的分类错误率。

$$\hat{C}_S \equiv Pr_{(x,y)\in S}(c(x) \neq y) = \frac{1}{m}\sum_{i=1}^{m} I(c(x_i) \neq y_i) \tag{2}$$

其中,$I(\cdot)$ 是布尔特征函数,其参数为真时值为 1,否则为 0。

定义 3 平均真实错误率[17],指的是分类器 $c \in Q$ 对样本 $x \in D$ 的误分概率。

$$Q_D \equiv E_{c\sim Q} C_D \tag{3}$$

定义 4 平均样本错误率[17],指的是分类器 $c \in Q$ 对样本 $x \in S$ 的误分概率。

$$\hat{Q}_S \equiv E_{c\sim Q} \hat{C}_S \tag{4}$$

由以上定义推导出 PAC-Bayes 的基本理论。

定理 1 (PAC-Bayes 定理[17]) 分类器 c 的先验分布为 $P(c)$,置信度为 $\delta \in (0,1]$,有:

$$Pr_{S\sim D^m}(\forall Q(c): KL(\hat{Q}_S \parallel Q_D) \leqslant \frac{KL(Q(c) \parallel P(c)) + \ln(\frac{m+1}{\delta})}{m}) \geqslant 1-\delta \tag{5}$$

这里 KL 是 Kullback-Leibler 差,也称为相对熵:

$$KL(p \parallel q) = q \ln \frac{q}{p} + (1-q)\ln \frac{1-q}{1-p} \tag{6}$$

$$KL(Q(c) \parallel P(c)) = E_{c\sim Q}\ln(Q(c)/P(c)) \tag{7}$$

PAC-Bayes 定理推导出泛化误差界基于概率近似正确的表达式。给定一个任意的分类器 c,其经验误差与真实误差的 KL 相对熵可由其先验分布与真实分布的 KL 相对熵来界定[23]。假设已知学习算法和数据集,就能确定出经验误差;当经验误差确定时,真实误差与两个分布的 KL 相对熵是单调递增的。因此,该理论推导出了真实误差的上界,该误差界是衡量学习算法泛化性能的重要指标。

3.2 PAC-Bayes 理论在 SVM 上的应用

研究表明,PAC-Bayes 边界是多种分类器上最紧的泛化误差边界,并能有效评价分类器的泛化性能[17-22]。这里仅分析 PAC-Bayes 边界应用于 SVM 分类器[17-19]。

采用 SVM 算法对文档数据质量进行分类和评估,能够实现较高的分类正确率,同时能较好地避免神经网络中常见的过拟合问题。对于线性不可分的数据,SVM 使用核函数将其映射到高维空间,变成线性可分数据,这里涉及到核函数的种类、核函数的参数 σ 和惩罚系数 C[24]。SVM 的各种核函数包括:线性核函数、多项式核函数、RBF 核函数和 sigmoid 核函数。本文拟研究不同类型核函数的 SVM 分类效果和不同参数的 SVM 分类效果,并将 PAC-Bayes 边界应用于 SVM,计算相应的 PAC-Bayes 边界。

定理 2　(PAC-Bayes 边界应用于 SVM[17])假设先验分布 P 和后验分布 Q 均为高斯分布,且协方差矩阵均为仅对角元为 1 的单位矩阵,先验分布 P 的中心是原点,后验分布 Q 的方向矢量为 w,该方向上距离原点为 μ。对于任意的 D 分布,给定 w 和 $\mu>0$,置信度为 $\delta\in(0,1]$,概率至少为 $1-\delta$:

$$KL(\hat{Q}_S(\boldsymbol{w},\mu)\parallel Q_D(\boldsymbol{w},\mu))\leqslant\frac{\frac{\Delta\mu^2}{2}+\ln(\frac{m+1}{\delta})}{m} \tag{8}$$

在定理 2 中,假设概念的先验分布 P 和后验分布 Q 都是权空间上的多元正态分布,从而使两种分布的相对熵 KL 得到简化。

设 $Q=N(\boldsymbol{X},\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$,$P=N(\boldsymbol{X},\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$,则 $KL(Q\parallel P)$表示为式(9):

$$KL(Q\parallel P)=\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)+\frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}+\frac{1}{2}\mathrm{tr}\{\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}-\boldsymbol{I}_d\} \tag{9}$$

协方差矩阵 $\boldsymbol{\Sigma}_1$ 和 $\boldsymbol{\Sigma}_2$ 是单位矩阵,则概念空间的两个分布的相对熵 KL 可表示为 $\Delta\mu^2/2$。

4　PAC-Bayes 边界应用于 Web 文档数据质量评估

本文提出将 PAC-Bayes 理论应用于 Web 文档数据质量评估。首先,建立评估模型,提取文档的综合特征;其次,提出一种基于 PAC-Bayes 理论的 Web 文档数据质量评估方法 DQAPB;再次,构建一种基于 PAC-Bayes 理论的评估指标体系,将 PAC-Bayes 边界、特异性和敏感性等作为在线文档质量评价的指标,来实现在线文档的质量评价。

4.1　建立模型

为了评估 Web 文档的数据质量,建立一个五元组的数据质量评估模型:

$$M=\langle D,F,W,A,I\rangle$$

其中,D 表示 Web 文档的数据集;F 表示 Web 文档数据的各项特征;W 表示各项特征相应的权值,这里默认为 1;A 表示对 Web 文档数据的各项特征进行分类和预测所采用的学习算法,这里采用 SVM 算法;I 表示评估 Web 文档数据质量的性能指标,这里包括 PAC-Bayes 边界、敏感性、特异性、正确率、正例预测值 PPV(Positive Predictive Value)、负例预测值 NPV(Negative Predictive Value)和马休斯相关系数 MCC(Matthews Coefficient of Correlation)等。

4.2　特征提取

考虑到不同质量水平的文档可能具有不同的属性和特征,本研究先对文档的属性进行特征提取,再使用 SVM 算法及 PAC-Bayes 边界来评估 Wikipedia 文档的数据质量。

首先定义 Web 文档数据质量维度包括:完整性、一致性、正确性、新鲜性、可用性和有效性。其次,作为非结构化数据,Web 文档的读写具有开放性,允许任何用户随时进行编辑,文档的各项特征会产生动态变化,而文档质量也会随之变化。影响 Web 文档质量的特征主要包括:文档的基本特征、文档的编辑历史、文档的网络链接特征和用户交互的四个方面。因此,基于这四个方面和以上六个数据质量维度,选择文档的 15 个特征,来描述一篇文档的质量,如表 1 所示。

Table 1　Data features of Web article

表 1　Web 文档数据特征

类别	文档特征	质量维度
文档的基本特征	文档长度(字节数)	完整性
	图片数	完整性
	文档的节数	完整性
	文档子节数	完整性
	文档建立时长	新鲜性
文档的编辑历史	文档编辑的日均值	正确性
	30 日内文档编辑数	正确性
	文档不同作者总数	一致性
	30 日内文档不同作者数	一致性
文档的网络链接特征	重定向该文档数	可用性
	网站内链接文档的入度	可用性
	网站外部链接文档的入度	可用性
	文档的出度	可用性
用户交互	文档读者数	有效性
	文档翻译语言数	有效性

对文档以上 15 个特征进行提取,再进行归一化处理,得到文档质量数据集。

4.3　性能度量

为了更好地评价 Web 文档质量,构建出一种基于 PAC-Bayes 理论的评价指标体系:计算 SVM 算法相应的 PAC-Bayes 边界值来衡量算法的泛化性能,同时引入敏感性、特异性、正例预测值和负例预测值、正确率等指标。

定义正例样本为 Positive,负例样本为 Negative,计算相关参数:真阳性的数量(TP),假阳性的

数量(FP),真阴性的数量(TN),假阴性的数量(FN),参数含义如表2所示。

Table 2 Description of positive sample and negative sample

表2 正负例样本描述

预测类	Positive	Negative
Positive	True Positive(TP)	False Positive(FP)
Negative	False Negative(FN)	True Negative(TN)

在此基础上分别计算以下参数来评估SVM分类器的性能:敏感性、正例预测值、特异性、负例预测值、正确率和马休斯相关系数[14]。其中,敏感性表示预测出的正例样本占全部正例样本的百分比。正例预测值表示实际的正例样本占预测出正例样本总数的百分比。特异性表示预测出的负例样本占全部负例样本的百分比。负例预测值表示实际的负例样本占预测出负例样本总数的百分比。计算公式如下所示:

$$SEN = [TP / (TP + FN)] \quad (10)$$

$$PPV = [TP / (TP + FP)] \quad (11)$$

$$SPC = [TN / (TN + FP)] \quad (12)$$

$$NPV = [TN / (TN + FN)] \quad (13)$$

$$ACC = [(TP + TN) / (TP + TN + FP + FN)] \quad (14)$$

$$MCC = \frac{[(TP \times TN) - (FP \times FN)]}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FN) \times (TN + FP)}} \quad (15)$$

4.4 算法实现

提出一种基于PAC-Bayes理论的Web文档数据质量评估方法DQAPB。DQAPB算法具体过程为:先根据目标文档提取特征,并进行归一化处理;然后设置SVM参数(C,σ),采用SVM算法对文档进行分类,经过训练生成不同的模型,利用产生的模型对文档的类别进行预测;再将PAC-Bayes理论应用于Web文档的数据质量评估,计算所对应的PAC-Bayes边界。该方法具有简便快速的优点,并具有较强的稳定性和鲁棒性。算法实现流程图如图1所示。

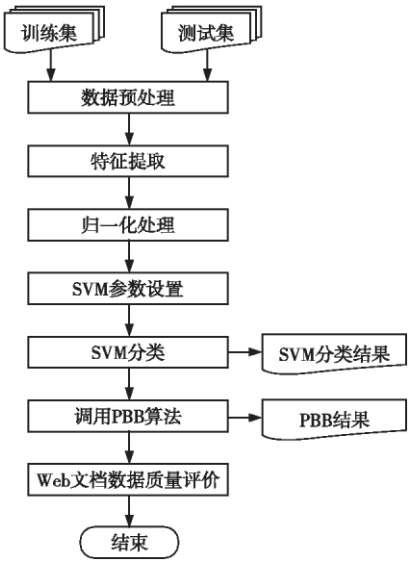


Figure 1 Flow chart of the algorithm

图1 算法流程图

算法1 基于PAC-Bayes理论的Web文档数据质量评估方法(DQAPB算法)

Input:训练文件、待预测文件。

Output:PAC-Bayes边界、正确率、敏感性、特异性等。

Step 1 初始化,确定要进行质量评估的文档;

Step 2 对文档的15个属性进行特征提取;

Step 3 对特征值进行归一化处理;

Step 4 读入训练数据文件和待预测文件,设置SVM核函数和参数(C,σ);

Step 5 利用SVM算法对训练数据执行训练,产生模型,再执行预测,计算出预测正确率、经验错误率、敏感性、特异性等;

Step 6 调用PBB算法,在SVM算法上应用PAC-Bayes边界,返回PAC-Bayes边界值;

Step 7 采用五折交叉验证方法,返回Step 4,至循环结束;

Step 8 利用各性能指标对算法的泛化性能和Web文档数据质量进行度量和评价。

算法2 PAC-Bayes边界算法(PBB算法)

Input:训练文件、预测文件。

Output:最优的PAC-Bayes边界。

```
delta ←0.1,eps ← 0.00001,PBbound ← 1;//初始化
KL(p,q)=p*log(p/q)+(1-p)*log((1-p)/(1-q));//KL函数定义
read (train file,test file);//读训练文件和预测文件
Foreach mu∈(0.01,100) do
  calculate Qs(train file,test file);//经验误差率 Qs
  lower ← Qs;upper←Qs;
  klbound=(mu*mu/2+log((m+1)/delta))/m;//式(9)右端
  While(KL(Qs,upper)< klbound & upper≤1-eps/2) do
    upper ← upper+(1-upper)/2;
  End While
  While(upper-lower > eps) do //二分法求边界
    If (KL(Qs, lower + (upper - lower)/2) > klbound) then
      upper ← lower+(upper-lower) / 2;
    else
      lower ← lower+(upper-lower) / 2;
```

End if
End While
If $upper < PBbound$ then $PBbound \leftarrow upper$;
End Foreach
return $PBbound$; //返回 PAC-Bayes 边界

5　实验结果及分析

5.1　实验数据

本文使用的 Web 文档数据集来自于 Wikipedia 中的 205 篇文档,分别为 FA 类、GA 类和 B 类,各类文档数据集的样本数如表 3 所示。其中,把 FA 类和 GA 类文档归为 A 类,定义 A 类文档为正例,定义 B 类文档为负例。实验中采用 libSVM 算法[25]。

Table 3　Data sets

表 3　数据集

文档类别	数据量	定义类
FA	61	A 类(正例)
GA	75	
B	69	B 类(负例)

5.2　五折交叉验证方法的实验结果及分析

将 SVM 算法及其 PAC-Bayes 边界应用于 Web 文档的数据质量评估,这里 SVM 使用 RBF 核函数,核函数的参数 σ 和惩罚系数 C 使用默认值。为了提高分类精度,采用五折交叉验证方法,实验结果如表 4 所示。

Table 4　Experimental results of 5-fold cross validation

表 4　五折交叉验证的实验结果

测试指标	均值	方差
SEN	91.93%	0.006 4
PPV	75.46%	0.001 2
SPC	40.55%	0.016 4
NPV	77.95%	0.044 7
ACC	74.63%	0.002 3
MCC	0.408 8	0.001 6
PBB	0.381 5	0.002 7

在表 4 中,SVM 算法正确率均值为 74.63%,*PBB* 均值达到 0.381 5,PBB 给出了算法的泛化误差边界,体现了误差率,该实验结果中的 *PBB* 与正确率之和近似于 1,验证了两者指标是一致的。同时,*SEN* 取值最高,*PPV* 和 *NPV* 值较高,分别为 75.46%和 77.95%。SVM 算法在该数据集上能取得较高的预测正确率、敏感性和较低的泛化误差边界,这表明,A 类和 B 类文档在质量水平上具有显著差异,该评估方法能够较好地实现对不同类别文档的分类和预测。各性能参数的方差值较低,体现了该方法具有较强的稳定性和鲁棒性。然而 *SPC* 和 *MCC* 取值较低,这说明还有某些其他文档特征或因素,对于文档的分类产生一定的影响。

5.3　核函数的实验结果及分析

为了研究核函数的分类效果,分别采用不同类型的核函数进行 SVM 分类实验,并计算相关的性能指标,如表 5 所示。

Table 5　Experimental results of kernel function

表 5　核函数的实验结果

核函数	*SEN* /%	*PPV* /%	*SPC* /%	*NPV* /%	*ACC* /%	*MCC*	*PBB*
线性核	94.85	**81.13**	**56.52**	84.78	81.95	0.581 9	0.329 6
多项式核	97.06	80	52.17	**90**	81.95	0.587 1	0.329 6
RBF 核	96.32	80.86	55.07	88.37	**82.44**	**0.596 5**	**0.323 9**
sigmoid	95.59	80.25	53.62	86.05	81.46	0.571 2	0.335 2

总体上看,采用 RBF 核函数,能够取得最好的分类效果,实现最佳分类正确率 82.44%、最低的泛化误差值 0.3239 及最优 *MCC*,此时其他指标也能得到较优值:*SEN* 和 *NPV* 仅略低于多项式核函数获得的最优相应值,而 *SPC* 和 *PPV* 也仅次于线性核函数获得的最高值。

5.4　模型选择的实验结果及分析

为了研究不同惩罚系数和核函数参数的分类效果,采用网格搜索法进行模型选择实验。这里默认采用 RBF 核函数,初始化 SVM 参数(C,σ),设 $C=C_i,\sigma=\sigma_j$,其中 $C_i\in\{0.01,0.1,1,10,100\}$,$\sigma_j\in\{0.1,1,10,50\}$,选择一组$(C_i,\sigma_j)$进行实验。实验表明,$(C,\sigma)$参数中惩罚系数 C 的取值并不会影响正确率等指标,而核函数参数 σ 会造成正确率等指标的改变,这里仅根据 $C=1$ 时的实验结果绘制图表,如图 2 所示。

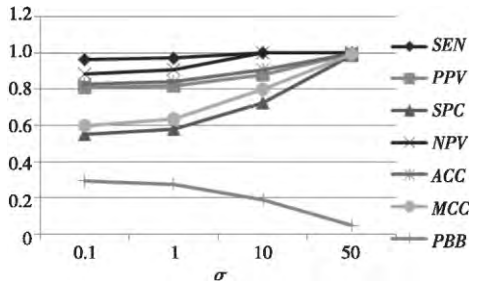


Figure 2　Experimental results of model selection

图 2　模型选择实验结果

图 2 表明,σ 越大,*ACC*、*SEN* 和 *SPC* 等指标

越高,而 PBB 越低。泛化误差界与分类正确率、敏感性和特异性等其他指标呈现出了较高的负相关性,比较符合理论期望。

5.5 开放测试和封闭测试的实验结果及分析

将数据集分为不同比例的训练数据和测试数据,分别采用封闭测试和开放测试进行实验。封闭测试是指利用训练数据本身进行预测;而开放测试是指对训练数据执行训练,对测试数据进行预测。这里 SVM 使用 RBF 核函数,参数 σ 和惩罚系数 C 使用默认值,实验结果如表 6 和表 7 所示。

Table 6 Experimental results of close tests

表 6 封闭测试的实验结果

数据集	*SEN* /%	*PPV* /%	*SPC* /%	*NPV* /%	*ACC* /%	*MCC*	*PBB*
20%	100	81.82	57.14	100	85.37	0.683 8	0.258 7
40%	96.30	85.25	67.86	90.48	86.59	0.697 0	0.243 8
60%	92.59	82.42	61.90	81.25	82.11	0.589 0	0.297 4
80%	96.30	80	53.57	88.24	81.71	0.583 3	0.302 2
100%	97.06	80.98	55.07	90.48	82.93	0.610 3	0.287 9

使用不同比例数据进行封闭测试时,SEN 均为 90%以上,PPV、NPV 和 ACC 均为 80%以上,PBB 均小于 0.3,这表明 SVM 分类效果较好,泛化误差界较小,不同类型的文档具有显著差异,通过该评估方法能够实现文档分类,并能利用这些指标有效评价文档的数据质量。

Table 7 Experimental results of open tests

表 7 开放测试的实验结果

数据集	*SEN* /%	*PPV* /%	*SPC* /%	*NPV* /%	*ACC* /%	*MCC*	*PBB*
20%	88.99	63.82	40	64.71	72.56	0.223 8	0.404 9
40%	84.15	82.14	63.41	66.67	77.24	0.481 8	0.353 3
60%	81.82	81.82	62.96	62.96	75.61	0.447 8	0.371 5
80%	89.29	75.76	38.46	62.50	73.17	0.325 8	0.398 3

开放测试中,SVM 算法预测的 ACC 达到 70%以上,SEN 达到 80%以上,均比封闭测试稍有下降。NPV 仅为 60%,比封闭测试结果下降显著,而 SPC 和 PPV 的取值比封闭测试也有不同程度的下降,PBB 取值则略有提高。结果表明开放测试的性能略低于封闭测试。

以封闭测试和开放测试中的 ACC 和 PBB 指标为代表,绘制图表进行对比,如图 3 所示。

由图 3 可见,封闭测试中,ACC 较高,相应的 PBB 较低。而开放测试的 ACC 较低,其 PBB 较高。封闭测试和开放测试的 PBB 与 ACC 均具有较高的负相关性。

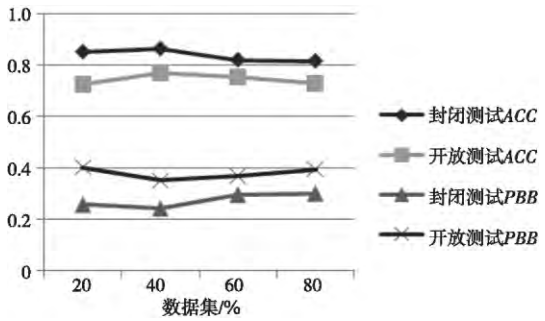


Figure 3 Experimental results of open test and close test

图 3 封闭测试和开放测试实验结果

综上表明,该评估方法能够对不同质量的文档实现较好的预测,PBB 给出了最紧的泛化误差边界,这些指标能够有效评价文档数据质量。SVM 算法取得的预测正确率等多项指标均与 PAC-Bayes 边界呈现较高的负相关性。较高的预测正确率和较低的泛化误差界表明,在文档的质量水平上,A 类和 B 类文档具有显著差异。

6 结束语

本文对文档的 15 个特征进行提取,来评价 Web 文档的数据质量,15 个特征值分别基于文档的基本特征、编辑历史、网络链接特征和用户交互这四个方面。本文将 SVM 算法与 PAC-Bayes 边界应用于文档质量的评估中,提出一种基于 PAC-Bayes 理论的 Web 文档数据质量评估方法 DQAPB,同时提出一种基于 PAC-Bayes 理论的评估指标体系,将 PAC-Bayes 边界、敏感性和特异性等作为在线文档质量评价的指标,实验结果表明,该方法能有效评价不同类别的文档质量,并具有简便快速、稳定性和鲁棒性强等优点。

PAC-Bayes 理论已广泛应用于各类机器学习模型,如何将 PAC-Bayes 边界更好地应用于各类实际问题,是将来研究的重要方向。

参考文献:

[1] Zong Wei, Wu Feng. The challenge of data quality in the big data age[J]. Journal of Xi'an Jiaotong University(Social Sciences), 2013, 33(5): 38-43. (in Chinese)

[2] Han Jing-yu, Xu Li-xhen, Dong Yi-sheng. An overview of data quality research[J]. Computer Science, 2008, 35(2): 1-12. (in Chinese)

[3] Yang Qing-yun, Zhao Pei-ying, Yang Dong-qing, et al. Research on data quality assessment methodology[J]. Computer Engineering and Applications, 2004, 40(9): 3-4. (in Chinese)

[4] Yan Hao, Qiu Hang-ping, Diao Xing-chun, et al. Comprehen-

sive data quality assessment based on improved analytic hierarchy process[J]. Journal of Computer Applications,2014,34(z1):287-290. (in Chinese)

[5] Yang Dong-shu, Yang De-sheng. Data quality assessment based on entropy weight and AHP[J]. Modern Electronics Technique,2013,36(22):39-42. (in Chinese)

[6] Han Jing-yu,Chen Ke-jia. Data quality assessment of web article content based on simulated annealing[J]. Journal of Computer Applications,2014,34(8):2311-2316. (in Chinese)

[7] Han Jing-yu,Chen Ke-jia. Ranking data quality of web article content by extracting facts[J]. Computer Science, 2014, 41(11):247-255. (in Chinese)

[8] Blumenstock J E. Size matters:Word count as a measure of quality on Wikipedia[C]// Proc of the 17th International Conference on World Wide Web,2008:1095-1096.

[9] Dalip D H,Lima H,Gonalves M A,et al. Quality assessment of collaborative content with minimal information[C]// Proc of the ACM/IEEE Joint Conference on Digital Libraries, 2014:201-210.

[10] Dalip D H,Goncalves M A,Cristo M,et al. Automatic assessment of document quality in web collaborative digital libraries[J]. Journal of Data and Information Quality,2011,2(3):1-14.

[11] Zeng H,Alhossaini M A,Fikes R,et al. Mining revision history to assess trustworthiness of article fragments[C]// Proc of the 2006 International Conference on Collaborative Computing: Networking, Applications and Worksharing, 2006:1-10.

[12] Hu M,Lim E P,Sun A. Measuring article quality in Wikipedia:Models and evaluation[C]// Proc of the 16th ACM International Conference on Information and Knowledge Management,2007:243-252.

[13] Yu S,Masatoshi Y. Assessing quality scores of Wikipedia article using mutual evaluation of editors and texts[C]// Proc of the 22nd ACM Conference on Information and Knowledge Management,2013:1727-1732.

[14] Liu J,Ram S. Who does what:Collaboration patterns in the Wikipedia and their impact on article quality[J]. ACM Transactions on Management Information Systems,2011,2(2):1-23.

[15] Valiant L. A theory of the learnable[J]. Communications of the ACM,1984,27(11):1134-1142.

[16] Mcallester D A. Some PAC-Bayesian theorems[J]. Machine Learning,1999,37(3):355-363.

[17] Langford J. Tutorial on practical prediction theory for classification[J]. Journal of Machine Learning Research,2005,6(3):273-306.

[18] Germain P, Lacasse A, Laviolette F, et al. PAC-Bayesian learning of linear classifiers[C]// Proc of the 26th Annual International Conference on Machine Learning,2009:353-360.

[19] Ambroladze A, Parrado-Hern E, Shawe-Taylor J. Tighter PAC-Bayes bounds[C]// Proc of Advances in Neural Information Processing Systems,2007:9-16.

[20] Morvant E,Koco S,Ralaivola L. PAC-Bayesian generalization bound on confusion matrix for multi-class classification[C]// Proc of the 29th International Conference on Machine Learning,2012:815-822.

[21] Gigu S,Marchand M,Sylla K,et al. Risk bounds and learning algorithms for the regression approach to structured output prediction[C]// Proc of the 30th International Conference on Machine Learning,2013:107-114.

[22] Seldin Y,Tishby N. A PAC-Bayesian approach to unsupervised learning with application to co-clustering analysis[J]. Journal of Machine Learning Research,2010,3:1-46.

[23] Tang Li,Gong Xiu-jun,He Li. Survey on PAC-Bayes bound theory and application research[J]. Journal of Frontiers of Computer Science and Technology, 2015, 9(1): 1-13. (in Chinese)

[24] Tang Li,Zhao Zheng,Gong Xiu-jun. Method of SVM model selection based on PAC-Bayes bound theory[J]. Computer Engineering and Applications,2015,51(6):27-32. (in Chinese)

[25] LIBSVM:A library for support vector machines[EB/OL]. [2011-09-15]. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

附中文参考文献:

[1] 宗威,吴锋. 大数据时代下数据质量的挑战[J]. 西安交通大学学报(社会科学版),2013,33(5):38-43.

[2] 韩京宇,徐立臻,董逸生. 数据质量研究综述[J]. 计算机科学,2008,35(2):1-12.

[3] 杨青云,赵培英,杨冬青,等. 数据质量评估方法研究[J]. 计算机工程与应用,2004,40(9):3-4.

[4] 严浩,裘杭萍,刁兴春,等. 基于改进层次分析的数据质量综合评估[J]. 计算机应用,2014,34(z1):287-290.

[5] 杨栋枢,杨德胜. 基于熵权和层次分析法的数据质量评估研究[J]. 现代电子技术,2013,36(22):39-42.

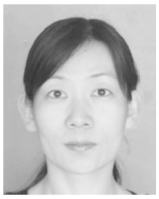
[6] 韩京宇,陈可佳. 基于模拟退火的在线 Web 文档内容数据质量评估[J]. 计算机应用,2014,34(8):2311-2316.

[7] 韩京宇,陈可佳. 基于事实抽取的 Web 文档内容数据质量评估[J]. 计算机科学,2014,41(11):247-255.

[23] 汤莉,宫秀军,何丽. PAC-Bayes 理论及应用研究综述[J]. 计算机科学与探索,2015,9(1):1-13.

[24] 汤莉,赵政,宫秀军. 基于 PAC-Bayes 边界理论的 SVM 模型选择方法[J]. 计算机工程与应用,2015,51(6):27-32.

作者简介:



汤莉(1979-),女,河北景县人,博士,讲师,CCF 高级会员(42781M),研究方向为机器学习和数据挖掘。**E-mail**: tangli0831@tjufe.edu.cn

TANG Li, born in 1979, PhD, lecturer, CCF senior member(42781M), her research interests include machine learning, and data mining.