多分类器融合技术在自动作文评分中的应用

陆 军[1] ,梁颖红[2] ,陆玉清[1] ,李 斌[1] ,姚建民[1]
(1 苏州大学 计算机科学与技术学院 ,江苏 苏州 215006 ;
2 江苏省现代企业信息化应用支撑软件工程技术研究开发中心 ,江苏 苏州 215104)

摘 要: 从作文的内容和语言学两个方面抽取了作文中相关的特征 ,并利用多种分类器(贝叶斯、K 近邻和支持向量机)根据各方面的特征实现了对作文的分类(评分) . 最后利用多分类器融合技术对多个分类器进行了融合处理. 通过实验分析 ,利用文本分类的方法对作文进行评分是完全可行的 ,在采用融合技术以后的评分性能有了较大的提高.

关键词: 自动作文评分 ;特征提取 ;文本分类 ;多分类器融合
中图分类号: TP391 **文献标识码**: A **文章编号**: 1000 - 7180(2009)10 - 0069 - 05

The Application of Classifier Combining in Automated Essay Scoring

LU Jun[1] ,LIANG Ying-hong[2] ,LU Yu-qing[1] ,LI Bin[1] , YAO Jian-min[1]
(1 School of Computer Science and Technology , Soochow University , Suzhou 215006 , China ;
2 Jiangsu Province Support Software Engineering R &D Center for Modern Information Technology Application in Enterprise , Suzhou 215104 , China)

Abstract : We aim to abstract related features from an essay by analyzing its content and structures. Meanwhile , classifiers including Bayes , KNN and SVM are adopted to realize better classification of essays (scoring) based on their features from various aspects. The multi-combination technology is also used in combining different classifiers. Through experimental analysis , it is indicated that scoring for essays is highly feasible via text-classifying method , and a higher performance is obtained by adopting multi-classifiers technology than previously single-classifier one.

Key words : automated essay scoring ; feature selection ; text classification ; classifier combining

1 引言

自动作文评分(Automated Essay Scoring ,AES)就是利用计算机技术对作文进行评估与记分[1] ,它已经成为电子化学习(e-learning)的重要组成部分. 目前 ,国外已经有一些主观题自动评分系统在运行 ,主要应用在 GMAT 等大型的考试中 ,主要有 E-rater、IEA、PEG 等. 在国内 ,对自动作文评分研究稍稍滞后 ,最早涉足自动作文评分领域的是梁茂成教授[2].

本研究拟探索一种自动化、低成本的自动作文评分方法. 利用相对成熟的自动文本分类的方法将作文划分到不同分数类型 ,最终实现对作文的自动评分.

2 系统框架

对语料库中的作文先预处理 ,并将同一分值的作文归为一类. 抽取作文的特征 ,并利用多种特征选择方法对抽取的特征进行筛选. 系统采用向量空间模型来表示每一篇作文;利用三种文本分类算法对待测作文进行分类(评分) . 最后将这些多个特征和多个分类器进行融合处理 ,最终得到分类(评分)结

收稿日期: 2009 - 04 - 30
基金项目: 江苏省现代企业信息化应用支撑软件工程技术研究开发中心项目(SX200907)

果.其整个处理流程如图 1 所示.

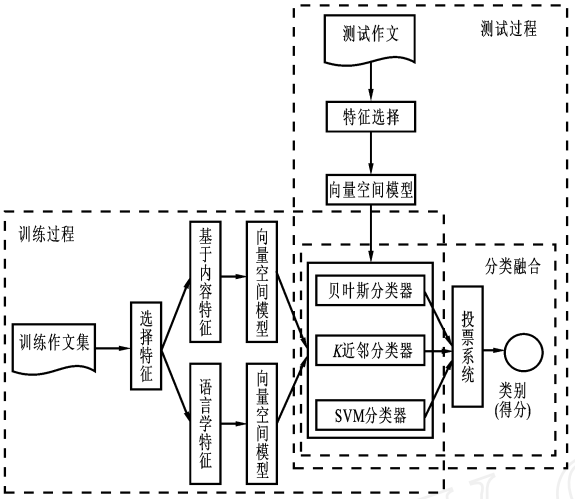

图 1 自动作文评分流程图

3 作文语料组织与特征的提取

3.1 作文类别的划分

本研究采用的语料库是中国学习者英语语料库[3](Chinese Learner English Corpus, CLEC).该语料库对中国学习者的英语作文中的错误进行了详细的标注、分类与统计.本研究选取了 CLEC 中主题为“Global shortage of Fresh Water”的四级作文作为研究语料,该主题的作文数量最多.由于高分段和低分段的作文数量都太少,所以选取了其中 7 至 12 分的作文作为本次研究的对象.从这些作文中选取 80 % 构成训练集,其余的作文构成测试集.每个分值的作文数量分布情况如表 1 所示.

表 1 各分值(类别)作文数量分布

分值	7	8	9	10	11	12	合计
训练集	24	28	40	35	31	26	184
测试集	4	8	12	11	6	5	46
合计	28	36	52	46	37	31	230

3.2 作文的表示和特征提取

一般情况下,在对作文进行评分之前需要对文本进行形式化处理,以便于计算机处理.在这里,需要考虑文本的表示和特征的权重表示两方面的问题.在文本表示中,本研究采用向量空间模型(Vector Space Model, VSM).从两个大的方面来选取特征:作文内容的特征和语言学的特征.

基于内容方面:本研究选取了作文中的单词和短语作为内容特征.在这里,作文中的单词是指英文作文中的 words,而短语则由作文中两个或三个相邻的 words 组合而成.这些特征也就是通常所说的 Uni-Gram、Bi-Gram 和 Tri-Gram 模型.这一方面的特征主要是考察作文的主题和内容.若作文中的主题、内容与写作要求不一致或不全面,则在这一方面有助于得分的特征会比较少.这些特征的权重使用布尔型权重.

基于语言学方面:本研究选取了浅层的语言学特征和复杂的语言学特征,其中浅层语言特征包括作文中句子的个数、句子的平均长度、单词的平均长度、单词的个数等;复杂的语言学特征包括作文中的句法、单词的词性、连接词、作文中的各种类型错误个数等,共有 25 种语言学特征.浅层语言特征简单地考察了作文的形式,本评分系统主要针对学生的习作,对于大量使用简单单词、句子或词数较少之类的作文不会有高的分数;而复杂语言特征从更深的层次上考察作文的语言,是对浅层语言特征的补充和深化,它主要考察作文的语法、连贯性和错误等方面.这些特征本身具有一个属性值,所以把这些属性值作为对应特征的权重.

以上两大方面的特征相互独立、相互补充,考察了作文的不同方面.参考四、六级作文的评分细则,这些特征比较全面地反映了一篇作文的水平.

为了有效降低高维特征空间的维数,降低计算复杂度,并希望能通过选择一些和类别相关性大的特征来提高分类的准确率,本系统采用了多种特征选择方法来对作文的特征进行筛选.系统分别采用了文档频率(Document Frequency, DF)、信息增益(Information Gain, IG)和统计量(Chi-square Statistic, CHI)对抽取的特征进行筛选[4].

4 文本分类算法应用于作文评分

4.1 朴素贝叶斯

朴素贝叶斯(Naive Bayes)分类模型源自古典概率中的贝叶斯公式,有着较为坚实的数学基础.该分类模型假定文本的各种特征之间相互独立,但是这个假设往往不完全成立.在实践中,朴素贝叶斯分类器的性能不稳定,易受分类任务的影响.

4.2 K 近邻

K 近邻(K-Nearest Neighbor)分类方法是一种传统的文本分类方法,该算法实现简单,分类结果稳定有效,其基本思想:对于待分类文档 X,在训练集中找到 K 篇与 X 最相似的文档.根据这 K 篇文档所属的类别来判定文档 X 所属类别. K 近邻方法要求有较高质量的训练集(精度要高),这对作文的分类是不利的;另外, K 值的确定也比较困难.

在本系统中,采用余弦法计算文档的相似度. K 值根据实验测试的结果调整得到. 在得到 K 篇最相似的文档后,使用以下方法判定待测文档的类别:

(1) 对这 K 篇文档根据与测试文档的相似度从小到大排序,则每篇文档就有一个序号(从 1 到 K).

(2) 用以下公式确定测试文档 X 与每一类别的权重 $P(X, C_i)$:

$$P(X, C_i) = \sum_{j=1}^{T_i} 2n_j/((K+1)K) \quad (1)$$

式中, T_i 表示在 K 篇相近文档中属 C_i 类的文章数, n_j 表示在这 T_i 文档中第 j 篇的序号.

(3) 将待测文本分到权重最大的那个类中.

4.3　支持向量机

支持向量机(Support Vector Machine, SVM)是在统计学习理论的基础上发展起来的一种新的机器学习方法,它在解决有限样本、非线性以及高维模式识别问题中有较好的表现,这些特性都有利于作文的分类. SVM 基于结构风险最小化理论之上,在特征空间中建构最优分割超平面,使得学习器得到全局最优化,具有较好的通用性.

5　多分类器融合技术

5.1　多分类器概述

传统的文本分类系统往往只使用某一个特定的分类器进行分类,这种系统对于类别数较大的问题很难获得好的分类效果. 多分类器的融合方法常用来获得更好的分类效果,其基本假设:对一个需要专家进行的任务, k 个专家个人判断的有效组合应该优于个人的判断[5]. 目前,它在模式识别的多个方面取得得了较好的应用效果.

从第 4 节的介绍中可以看出贝叶斯、K 近邻和 SVM 这三种分类方法有着不同的优势和缺陷,并存在一定的互补关系,本研究对这三种分类器进行融合处理,使得它们之间能够"取长补短".

5.2　多分类器融合技术

5.2.1　多分类器融合模型

假设每个分量分类器都取自混合模型,首先根据分布 $P(r \mid x, \theta_0)$ 随机选取一个用 r 标记的过程或函数($1 \le r \le k$), θ_0) 是参数向量. 然后被选出的过程将根据 $P(y \mid x, \theta_0)$ 产生一个输出 y(即一个类别标记),其中的参数向量 θ_r 表示该过程的自然状态(上标 0 代表的是产生的模型的特性. 在下面的式子中,没有上标的项用来表示分类器中的参数). 这样产生 y 的总的概率可由下式对全部过程的求和得出[6]:

$$P(y \mid x, \Theta^0) = \sum_{r=1}^{k} P(r \mid x, \theta_0^0)\, P(y \mid x, \theta_r^0) \quad (2)$$

式中, $\Theta^0 = [\theta_0^0, \theta_1^0, \theta_2^0, \cdots, \theta_k^0]^{\mathrm{T}}$ 表示全部有关的参数向量. 图 2 显示出一个这样的系统分类器的基本结构,其任务是将待分类文档分成 $C_i(i = 1, 2, \cdots, M)$ 类中的 1 类.

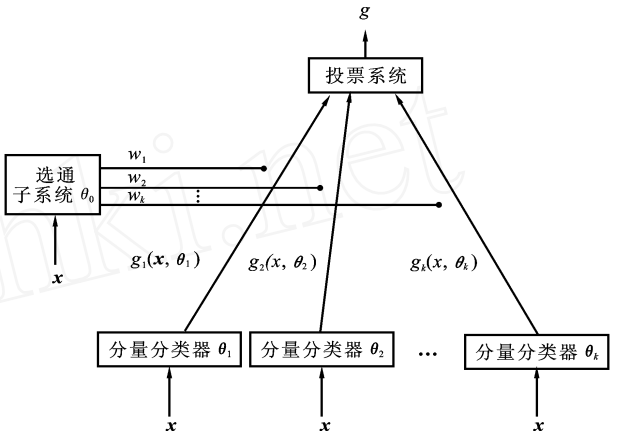

图 2　基于投票的多分类器融合系统

上述结构表示的是由 k 个"分量分类器"或"专家"组成的"混合专家"组合分类器结构. 其中每个分量分类器都有一个可训练的参数 $\theta_i, i = 1, \cdots, k$. 对每个输入样本 x,每个分量分类器 i 都给出一个类别隶属度 $g_{ir} = P(C_r \mid x, \theta_i)$ 的模型.

这种结构适合于假设的混合模型,一个待分类样本 x 将被提供给 k 个分量分类器,每一个都输出 M 个标量的判别函数值(每个对应一类). 这种对分量分类器 r 的 M 个判别值组织在一起,记作 $g(x, \theta_r)$,并且有

$$\sum_{i=1}^{M} g_{rj} = 1, \text{对所有 } r \quad (3)$$

分量分类器 r 输出的全部判别值都乘上一个标量系数 w_r, w_r 的值由一个"选通子系统"给定."混合专家"结构被训练使得每个分量分类器都能对应于混合模型中的一个过程,而"选通子系统"则表达了式(2)中的混合参数 $P(r \mid x, \theta_0^0)$ 的模型. 在本系统中最终的判决是选择经投票系统后的具有最大判别值的那个类别.

下面通过具体实例来描述基于栈的多分类器融合过程:对于 M 个模式类别 $C_i(i = 1, 2, \cdots, M)$ 和 K 个分类器 $\theta_r(r = 1, 2, \cdots, K)$ 的分类问题,设分类器 θ_r 的度量矢量为 X_r,对于一个待测的类别 X, X 将被判为具有最大后验概率的类别,即

$$X \rightarrow C_j, \text{当 } P(C_j \mid X_1, X_2, \cdots, X_k) = \max_{t=1}^{k} P(C_t \mid X_1, X_2, \cdots, X_k) \quad (4)$$

利用式(2),可以推导出判决公式:

$$X \rightarrow C_j, \text{当} \sum_{r=1}^{K}(P(X_r)\,p(C_j \mid X_r)) = \max_{t=1}^{M}(P(X_r)\,P(C_t \mid X_r)) \quad (5)$$

5.2.2 分类器权重计算

由于各分量分类器的分类效果是随着待定样本的特征的不同而变化的,因此需要针对不同样本进行分类器组合及权重的调整. 文中方法考察分类器在待定样本的分类准确率,各个分量分类器的权重与其准确率成正比,并且不考虑准确率太低的分量分类器[7]. 准确率计算可以通过构造混乱矩阵后进行,计算公式如下:

$$P(X_r) = Acc_r = \sum_{i=1}^{M} r_{jj}^{(i)} / \sum_{j=1}^{M}\sum_{l=1}^{M} r_{jl}^{(i)} \quad (6)$$

式中,$r_{jl}^{(i)}$ 是混乱矩阵中的某一元素,表示分类器 i 将类别 C_j 中的样本识别为 C_l 类的概率.

将式(6)代入式(5)便能得出最后的分类结果.

6 实验结果与分析

6.1 性能评价标准

通常采用准确率、召回率和 F_1 测度来评价分类结果的好坏,这三种评价标准的公式表示如下:

准确率: p = 分类的正确文本数 / 实际分类的文本数 (7)

召回率: r = 分类的正确文本数 / 应有的文本数 (8)

F_1 测度: $F_1 = 2 \times p \times r/(p + r)$ (9)

此外,有微平均和宏平均两种计算准确率、召回率和 F_1 测度值的方法. 具体到作文评分中,对同一篇作文不同的人有不同的评分是常见的,像 E-rater 等商用机器评分系统通行的准确率衡量方法都是计入相邻分数,即上下只差一分的,也算准确评分,因此采用这种准确率衡量方法.

6.2 实验结果

经过测试,从整个测试集上的统计来看,贝叶斯、K 近邻和 SVM 三种分类方法的准确率分别为 0.58、0.63 和 0.67,而在把这些分类器恰当融合后评分准确率提高到了 0.73. 详细的实验结果见表 2.

在实验测试中,一些分类器把作文判定某一分数的篇数为 0,比如贝叶斯分类器判定测试集中无 7 分作文,此时准确率和 F_1 测度无法计算,对应的召回率为 0. 表 2 中用 Def 表示无法计算的数据.

表 2 作文评分测试结果

分数	准确率				召回率				F_1 测度			
	贝叶斯	K 近邻	SVM	融合	贝叶斯	K 近邻	SVM	融合	贝叶斯	K 近邻	SVM	融合
7	Def	0.0	Def	Def	0.0	0.0	0.0	0.25	Def	0.0	Def	Def
8	1.0	1.0	Def	0.50	0.75	0.38	1.0	1.0	0.86	0.55	Def	0.67
9	0.63	0.57	0.67	0.71	0.75	0.92	1.0	0.92	0.68	0.70	0.80	0.80
10	0.73	0.77	Def	1.0	0.73	1.0	1.0	1.0	0.73	0.87	Def	1.0
11	0.50	0.14	Def	0.75	0.50	0.40	0.0	0.33	0.50	0.20	Def	0.45
12	0.20	Def	Def	1.0	0.20	0.0	0.0	0.20	0.20	Def	Def	0.33

6.3 讨论与分析

从总体来看,所有用到的分类方法的效果都要低于它们在其他方面的分类效果,文中认为这是由两方面原因造成的,一是四级作文本身的特点,二是各分数作文的篇章数的分布情况. 选取的是同一主题的作文,不同分数作文之间的差别比较小,尽管使用了多种特征提取方法,尽量去除对分类贡献不大的特征,但是从测试结果来看,特征选择并不令人满意,这也是作文评分中一个较大的难点. 其次,各分数作文的篇章数的分布也对实验结果产生了很大的影响. 在作文语料中,7、8、11 和 12 分的作文比较少,9 和 10 分的作文比较多,这样导致了高、低分段作文评分模型训练不足. 从测试结果中明显可以看出中间分数段作文的评分效果比高、低分段的作文评分效果好很多.

虽然实验的准确率与 E-rater 等商用机器评分系统的准确率有一定的差距,但是仍然得到了许多有指导意义的结果. 首先,从实验结果来看,用文本分类的方法对作文评分是可行的. 其次,单独的分类器对作文评分有不同的性能表现,而且有时很不可靠稳定,比如 K 近邻算法对 10 分和 11 分作文的确率分别为 0.77 和 0.14,波动很大. 在采用多分类器融合技术后,评分的准确率得到了较大的提高,在评分的可靠稳定性方面也有较大改善. 使用单独的分类器评分时,高、低分作文篇数的减少对评分效果有很大的影响,但是采用融合分类器后,发现这些影响有所减小.

7　结束语

本次研究表明,利用文本分类方法对作文评分是可行的,相对于单独的分类器,采用多分类器融合技术可以较大的提高评分性能.为了获得更好的评分性能,还有许多方面值得研究和改进.首先,目前仍然缺乏足够的作文语料.其次,在特征选择方面,语言学特征更能体现作文的水平,对于这些特征的选取还可以做很多,这需要运用现有的 NLP 技术.最后,在本次研究中主要使用了通用的融合技术,将来可以根据作文评分的自身特点对融合技术做一些调整和改进.

参考文献:

[1] Shermis M D, Burstein J. Automated essay scoring: a cross - disciplinary perspective [M]. Mahwah, NJ: Lawrence Erlbaum Associates, 2003.

[2] 梁茂成. 中国学生英语作文自动评分模型的构建[D]. 南京: 南京大学, 2005.

[3] 桂诗春,杨惠中. 中国学习者英语语料库[M]. 上海:上海外语教育出版社,2003.

[4] Yiming Yang, Jan O, Pedersen. A comparative study on feature selection in text categorization[C]// Proceedings of the Fourteenth International Conference on Machine Learning. USA: Carnegie Mellon University, 1997, 8 (12) :412 - 420.

[5] Sebastiani F. A tutorial on automated text categorization [C]// Proc. Of Argentinian Symposium Artificial Intelligence. Buenos Aires, 1999:7 - 35.

[6] Richard O, Duda, Peter E Hart, David G Stork. Pattern classification[M]. 2nd ed. Beijing: China Machine Press, 2004.

[7] 唐春生,金以慧. 基于全信息矩阵的多分类器集成方法[J]. 软件学报, 2003, 14(6): 1103 - 1109.

作者简介:

陆　军　男,(1987 -).研究方向为计算机科学与技术.

(上接第 68 页)

[7] 韩艳,林煜熙,姚建民. 基于统计信息的未登录词扩展识别方法[C]//第四届全国信息检索与内容安全学术会议,上卷. 北京,2008:102 - 110.

[8] Ferreri Cancho R, Sole R V. The small world of human language[J]. Biological Sciences, 2001, 268(1482):2261 - 2265.

[9] 刘涛,陈忠,陈晓荣. 复杂网络理论及其应用研究概述[J]. 系统工程,2005,23(6):165 - 168.

作者简介:

林煜熙　男,(1987 -).研究方向为自然语言处理.