结合语言学特征和自编码器的英语作文自动评分①

魏扬威, 黄萱菁

(复旦大学 计算机科学与技术学院, 上海 201203)
(复旦大学 上海市智能信息处理重点实验室, 上海 201203)

摘 要: 近年来, 越来越多的大规模英语考试采用了自动评分系统. 因此, 对英语作文自动评分的研究有着非常重要的价值. 我们先依据英语作文写作技巧提取了大量语言学特征, 再分别使用自编码器, 特征值离散化方法对特征进行重构, 最后我们使用分层多项模型来输出文章的最终得分. 实验表明, 该方法能取得很好的预测效果, 而且面对不同主题的作文进行预测时也能显示出较好的鲁棒性. 相比于传统自动评分方法皮尔森相关系数高出 9.7%, 具有良好的实际应用价值.

关键词: 自动评分; 自编码器; 离散化; 文本特征提取

Automatic Essay Scoring Using Linguistic Features and Autoencoder

WEI Yang-Wei, HUANG Xuan-Jing

(School of Computer Science, Fudan University, Shanghai 201303, China)
(Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai 201303, China)

Abstract: In recent years, more and more large-scale English tests begin to use the automatic scoring system. Therefore, the research of this system is of great value. In this paper, we first extract a lot of features according to English writing guide. Then we use autoencoder and discretization algorithm to learn a different representation of features. Finally, we use a hierarchical multinomial model to output the final scores of articles. Experimental results indicate that this method not only achieves great performance for those essays of the same topic, but also shows good robustness when predicts essays of different topics. Compared with the traditional automatic score method, our approach achieves higher than 9.7% in term of Pearson Correlation Coefficient, with good practical values.

Key words: automatic essay scoring; autoencoder; discretization; textual feature extraction

英语作文自动评分使用自然语言处理相关技术, 让计算机系统对于目标文章给出合适的得分. 随着很多英语等级认证考试报名人数的增加和计算技术的发展, 一些自动评分的软件已经正式被使用. 国外最有代表性的自动评分系统有: Project Essay Grade(PEG), 于 1966 年由美国的杜克大学(University of Duke)的 Ellis Page 等人开发[1]; intelligent Essay Assessor(IEA), 由美国科罗拉多大学(University of Colorado)开发[2]. e-rater 评分系统, 已经正式被用来评测 TOEFL 和 GRE 考试中文章的质量[3]. 性能优异的自动评分系统结合文本纠错的功能[4,5]能减少人的工作量, 极大地节约人力物力资源.

英语作文自动评分的方法研究一直是一项具有挑战性的, 且不断被完善的任务. 1996 年 Arthor Daigon 通过对文章语言形式的考察进行文章质量评测[6]; 1998 年, Leah S. Larkey 使用了基于文本分类的方法取得了性能的提升[7]; 2011-2014 年, Isaac Persing 和 Vincent Ng 等人发表了一系列的文章, 使用了回归方法分别从文章的组织结构[8], 文章和对应主题的相关性[9], 还有文章表达的清晰度方面[10]对文章质量进行评估; 2013 年, Hongbo Chen 和 Ben He 使用了排序的方法, 通过先对文章质量进行排序再进行划分等级来对

① 基金项目:国家自然科学基金(61472088)
收稿时间:2016-04-22;收到修改稿时间:2016-05-23 [doi:10.15888/j.cnki.csa.005535]

© 中国科学院软件研究所 http://www.c-s-a.org.cn

文章评分[11].

自编码器(autoencoder)是人工神经网络的一种, 通常用来学习特征的有效编码. 2006 年 Hinton 发表在 science 上的文章[17]提出了自编码器, 引发了这几年科学界对人工神经网络研究的热潮. Hinton 在文中使用了自编码器对图像的特征矩阵进行压缩编码. 自编码器也可以用于我们的英语自动评分任务, 一方面可以降低特征的维数, 另一方面可以通过重构捕捉到原始特征中最重要的信息.

1 自编码器

一篇英语作文的原始特征直接用来进行分类或者回归, 往往很难得到很好的评分预测结果. 我们可以先使用自编码器对原始特征进行重编码, 再使用编码结果来对文章的评分进行预测.

自编码器主要是学习一个近似等式:

$$h_{w,b}(X) \approx X \tag{1}$$

这里的 X 表示输入矩阵, w 表示权重矩阵, b 表示偏置. 自编码器包括编码和解码的两层结构. 通过编码可以得到特征的另外一种表示方式, 再通过解码将编码结果还原出来. 如果最终输出的还原结果和输入非常接近, 那么编码结果就可以看成是输入的近似代替.

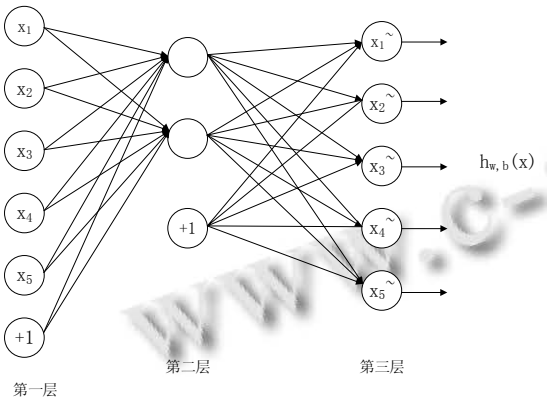

图 1 自编码器

自编码器的意义不在于还原输入数据, 而是体现在对隐层神经元的限制. 如图 1 所示, 为了进行压缩编码, 我们将隐层神经元的数量设置为 2, 这样就可以将输入的 5 维特征压缩到 2 维. 当隐层神经元的数量大于输入特征的维度时, 可以得到特征的高维稀疏编码结果.

显然, 自编码器的目标函数是输出结果和输入之间的重构误差尽可能小. 其计算公式如式(2)所示:

$$J(W,b) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}|x^{(i)} - h_{w,b}(x^{(i)})|^2 \tag{2}$$

当然, 为了避免系统的过拟合, 我们还要加入一个正则化项来控制模型的复杂度增长:

$$J'(W,b) = J(W,b) + \frac{1}{\lambda}\sum_{i,j,l}(W_{i,j}^{(l)})^2 \tag{3}$$

如果我们训练的是稀疏自编码器, 需要在目标函数中再增加一个约束项, 控制模型的稀疏性. 这里引入激活的概念, 如果最后传递函数的输出结果非常接近于 0, 那么我们认为该神经元没有被激活. 而如果最后传递函数输出的结果接近于 1, 那么该神经元被激活了. 通常来说, 传递函数为 sigmoid 函数:

$$f(x) = \frac{1}{1+e^{-x}} \tag{4}$$

或者是:

$$f(x) = \begin{cases} 0, & x \le 0 \\ x, & 0 < x < 1 \\ 1, & x > 1 \end{cases} \tag{5}$$

我们把对稀疏性的控制具体量化为对平均激活度的控制, 令 $a_j^{(2)}(x)$ 表示给定输入 x 时, 第 j 个隐层神经元的激活度, 具体地, 其平均值(整个训练数据上)为:

$$\hat{\rho}_j = \frac{1}{m}\sum_{i=1}^{m}[\alpha_j^{(2)}(x^{(i)})] \tag{6}$$

再引入稀疏性参考 ρ, 通常是一个非常接近于 0 的值, 比如 0.05. 然后计算 ρ 和 $\hat{\rho}_j$ 的信息增益, 用来描述这两者之间分布的差别.

$$\sum_{j=1}^{s_2} KL(\rho \| \hat{\rho}_j) = \sum_{j=1}^{s_2}\rho\log\frac{\rho}{\hat{\rho}_j} + (1-\rho)\log(\frac{1-\rho}{1-\hat{\rho}_j}) \tag{7}$$

其中, S_2 表示隐层中神经元的总数, j 是对隐层神经元的索引. 对于稀疏编码, 我们将上式的信息增益也作为惩罚项加入目标函数中. 因此对于稀疏自编码器, 其目标函数为式(8)所示. 其中 β 是一个系数, 表示对稀疏性惩罚的力度, 这个值越大表示对稀疏性要求越高.

在有了压缩编码和稀疏编码自编码器的目标函数之后, 我们可以进一步利用优化算法, 如梯度下降法, 来对目标函数进行优化以得到最优的网络结构. 在英语作文自动评分任务中, 对于提取的原始特征, 我们可以进一步使用自编码器进行重构. 通过控制隐层神

© 中国科学院软件研究所 http://www.c-s-a.org.cn

经元的数量, 一方面压缩编码进行特征压缩, 另一方面稀疏编码将特征重构到高维.

$$J_{sparse}(W,b)=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}|x^{(i)}-h_{w,b}(x^{(i)})|^2+\frac{1}{\lambda}\sum_{i,j,l}(W_{i,j}^{(l)})^2+\beta\sum_{j=1}^{S_2}KL(\rho\|\hat{\rho}_j) \tag{8}$$

2 特征值离散化

机器学习系统进行数据训练时, 有时候会遇到少量的异常样本. 比如英语作文自动评分任务, 其中一维特征是平均每句话中第一人称代词所占的比例. 这个比例不会太高, 一般来说低于 0.25, 一些异常学生作文在该维取值可能达到了 0.8, 0.9 或者更高. 为了削弱这些异常样本的影响, 我们可以使用不同的区间来对特征值进行分段. 比如这里我们可以取 0~0.1, 0.1~0.2, 0.2~0.3, 0.3~0.4 和 0.4~1 这几个区间. 不论异常作文在该维度的取值是 0.8 还是 0.9 统统归到 0.4~1 这个区间中, 其本身的特征值并不会加入系统训练. 这样可以大大减少异常样本对系统整体性能的干扰.

特征离散化关键问题就在于分割区间的选择[18], 不同的分割区间直接影响到系统的性能. 我们首先将所有样本都归为一个区间中, 使用信息增益的方式, 来决定是否进一步分割区间, 再递归地分割其每个子区间. 首先是特征 A 对应的熵, 如式(9)所示:

$$E(A,S)=-\sum_{i=1}^{k}p(F_i,S)\log p(F_i,S) \tag{9}$$

其中 S 是特征 A 对应的取值的集合, $p(F_i,S)$ 是 S 上取值 F_i 对应的比例. 下面我们使用分割边界 T 对特征 A 划分, 划分之后其熵值计算方法为:

$$E(A,T;S)=\frac{|S_1|}{|S|}E(A,S_1)+\frac{|S_2|}{|S|}E(A,S_2) \tag{10}$$

其中 $S1$ 和 $S2$ 分别是集合 S 对应分割边界 T 的两个子集. 因此信息增益为:

$$Gain(A,T;S)=E(A,S)-E(A,T;S) \tag{11}$$

当然我们不能无限对特征值区间进行分割, 因此, 我们需要增加一个停止分割条件[19]:

$$Gain(A,T;S)<\frac{\log(N-1)}{N}+\frac{\Delta(A,T;S)}{N} \tag{12}$$

其中, N 是集合 S 中的元素个数, 使用以下公式进行计算:

$$\Delta(A,T;S)=\log(3^k-2)-[k\cdot E(A,S)-k_1\cdot E(A,S_1)-k_2\cdot E(A,S_2)] \tag{13}$$

其中 k 表示 S 中元素的个数, $k1$ 和 $k2$ 分别表示 $S1$ 和 $S2$ 中元素的个数. 有了上述条件之后, 我们在对连续特征进行分割的时候就会逐渐收敛, 最终停止得到最优的分割结果.

离散化能够进一步提升特征值的表达能力, 离散化之后的多维特征, 每个特征都可以有不同的权重, 因此特征的表达能力得到进一步提升, 系统更加稳定. 我们在进行自动评分时, 可以对于提取到的原始特征首先进行离散化, 离散化后的高维特征再使用自编码器重编码. 重构的特征最后分别使用支持向量机进行回归或者分层多项模型进行分类, 以输出一篇学生作文的最终得分.

3 文本的特征表示

一般地, 对英语作文的评价主要基于三个大的方面: 词汇的使用, 即词汇使用是否正确, 词汇量是否丰富, 是否高级优美; 语法的使用, 即语法结构是否正确, 语法结构是否复杂(不能过于单一), 句子是否通顺; 话语的长短和连贯性, 即句子和句子之间结构是否紧凑.

2002 年 Eli Hinkel 研究了母语是英语的学习者和母语非英语的学习者的英语写作在词法、从句和句子间关系三个方面应用的差异, 提供了指导英语写作的一些技巧[12]. 我们认为这些差异和技巧能反映英语学习者的文章质量, 因此从这些角度出发, 提取了一些语言学特征. 大多数现有的评分系统往往给出的只是简单的特征, 如文章长度, 句子长度, 停用词的个数. 但是这些特征都不能直接反应文章的写作水平, 我们这里提取的特征, 每一项都旨在考察文章的表述能力和语言的运用能力, 更加细致地考察了作者的写作功底. 因此, 我们的特征能更好地完成作文评分的任务.

3.1 词法特征

词法特征是对于英语写作水平最基本的考察, 词法的特征能反应文章作者对于词汇和短语的掌握能力. 如表 1 所示, 其中列举类、语言活动类、分析类、结果类和模糊类等是作者表述中常用的关键性名词词汇; 动词的不同时态以及动词不定式和动名词能够考察作者对于动词形式变换的熟练程度; 形容词和副词在句法中常用作修饰成分, 能考察作者对于

© 中国科学院软件研究所 http://www.c-s-a.org.cn

不同修饰词其修饰程度的把握. 学生英语作文中　　词汇量不能过于狭窄, 不能仅仅使用某一类的词.

表 1　词法特征

特征类别	举例或说明
名词分类	列举; 进展/追溯; 语言活动; 言外行为; 分析; 结果; 模糊
人称代词	第一人称; 第二人称; 第三人称
泛指和否定	every one, no one, nothing
断言	anyone, some, something
形式主语	it 或 there 做主语
动名词	动词的-ing 形式具有代词的性质和名词的功能
动词时态	过去式, 现在式, 将来式, 进行时态, 完成时态
动词分类	公共; 私有; 劝说; 逻辑/语义关系; 期望/打算
情态动词	can, may, might, could, must, have to, should, ought to
被动式	by+动词过去分词
动词不定式	通常是 to+动词原形
分词用作形容词或副词	现在分词, 过去分词
形容词作定语	词性为形容词, 句中成分为定语
形容词作表语	词性为形容词, 句中成分为表语
副词分类	时间; 频率; 地点; 语气加强; 语气缓和

3.2 从句特征

传统的特征提取往往只有词汇级别的考察. 但是仅仅考察作者对于词法使用的能力是不够的, 假如一篇作文通篇堆砌高级的词汇或者精美的短语, 可是全部使用单一的简单句、短句子, 按照作文评测的标准不能给予高分. 另一方面, 如果只考察词汇, 系统很容易被学生作文刻意使用一些词汇所欺骗[20]. 从句的特征考察的正是作者运用复杂句式的能力, 如果文章中使用的词汇优美准确, 而且能够很好地运用各类从句使句法不再单一, 这样的文章是有理由给予较高分数的.

表 2　从句级别的特征

特征类别	功能
名词性从句	作主语, 作宾语, 作形容词的修饰, 作介词的修饰
形容词性从句	完整的从句, 简化的从句
副词性从句	表示原因, 表示让步, 表示条件, 表示目的, 简化的从句

3.3 句子间关系

如果作者的文章中对于词汇和从句已经能够很好的掌握, 可是句子和句子之间不连贯没有逻辑, 我们显然不能给予这篇文章很高的得分. 因此我们加入了句子间关系的特征, 用来考察作者文章对于前后句子连贯性句子间逻辑性的掌握. 如表 3 所示, 主要考察前后句的并列, 平行, 递进, 因果, 转折关系, 以及后一句是否是对前一句的说明或者限制. 以上就是我们全部的语言学特征. 首先从词法方面, 考察了英语作文中对各类词汇的掌握情况, 每类词汇都有其特定的表达含义和语气. 同时还考察了动词和形容词的词法活用, 反映了学生对于基本的语法知识、句子成分的理解. 然后我们考察了英语作文中从句的使用情况, 各类从句运用得是否恰当能极大地反映英语写作水平的高低. 最后考察的是句子的前后关系, 我们认为好的文章不仅要能有好的词汇表达, 好的从句使用, 还要在文章的组织结构上要有一定的逻辑性, 连贯性. 其中从句级别特征和句子间关系特征在提取的时候, 我们先使用 Stanford parser 进行句法分析[13], 再从句法分析树上进行匹配.

表 3　句子间关系的特征

特征类别	功能
平行结构	并列, 平行关系
逻辑连接	递进, 因果, 转折关系
范例	对前一句的举例说明
限制	对前一句的限制
疑问句	较为少见, 不应过多出现

4　数据集

本项研究的数据集在kaggle上公开, kaggle是一个机器学习比赛的公共平台, 我们可以免费注册账号下

© 中国科学院软件研究所 http://www.c-s-a.org.cn

载其举办的比赛的训练数据. 该数据集是 7-10 年级的第一语言学习者的英语作文, 一共包含 8 个子集, 每个子集都是独立的数据, 独立的主题, 平均文章长度都不同. 数据集概况见表 1, 其中数据子集 2 在 kaggle 中给出了 2 项评分, 我们在这里选取了第 1 项评分, 即写作应用项作为其最终得分.

表 4 数据集

数据子集	文章类型	文章作者年级	满分	文章数量
1	论述文 / 叙事文 / 说明文	8	12	1783
2	论述文 / 叙事文 / 说明文	10	6	1800
3	根据源文章回答问题	10	3	1726
4	根据源文章回答问题	10	3	1772
5	根据源文章回答问题	8	4	1805
6	根据源文章回答问题	10	4	1800
7	论述文 / 叙事文 / 说明文	7	30	1569
8	论述文 / 叙事文 / 说明文	10	60	723

如表 4 中所示, 文章类型主要是论述类、叙事类、说明类和回答问题类. 论述文、叙事文或者说明文要求作者的文章描述一个故事或者新闻. 而回答问题类则要求作者先阅读一段材料, 再根据阅读材料最后给出的问题和要求写一篇文章. 8 个数据子集的主题各自不同, 其中, 子集 1 要求谈论计算机对生活带来的影响; 子集 2 是谈论图书馆是否需要对图书内容进行审查; 子集 3-6 是先阅读材料再根据提示写作文, 4 篇材料也都不同; 子集 7 要求写一篇关于耐心的故事; 子集 8 说明笑是人际关系中的一个重要元素, 要求写一篇关于笑的文章.

5 实验

5.1 实验评测

结果分别使用皮尔森相关系数 r, 平均偏差 $\bar{d}$ 和均方偏差 σ^2 进行统计[8].

$$r=\frac{N\sum_{i=1}^{N}A_iE_i-\sum_{i=1}^{N}A_i\sum_{i=1}^{N}E_i}{\sqrt{N\sum_{i=1}^{N}A_i^2-(\sum_{i=1}^{N}A_i)^2}\sqrt{N\sum_{i=1}^{N}E_i^2-(\sum_{i=1}^{N}E_i)^2}} \quad (14)$$

$$\bar{d}=\frac{1}{N}\sum_{i=1}^{N}|A_i-E_i| \quad (15)$$

$$\sigma^2=\frac{1}{N}\sum_{i=1}^{N}(A_i-E_i)^2 \quad (16)$$

其中 A_i, E_i 分别表示第 i 篇文章的人工评分和系统评分, N 表示文章的总数. 皮尔森相关系数 r 用来反映系统评分和人工评分的线性相关程度, 取值范围在[-1,1]之间, 值越大说明相关性越高. 接近于 0 表示人工评分和系统评分几乎没有相关性, 接近于 1 表示人工评分和系统评分几乎一致, 而小于 0 则表示人工评分和系统评分负相关. $\bar{d}$ 表示系统得分和人工得分之间的平均偏差, σ^2 表示系统得分和人工得分之间的均方偏差. 相关系数用来作为最主要的评测指标, 平均偏差和均方偏差作为参考. 我们进行 5 折交叉验证, 对于每个数据子集, 随机切分成 5 份, 每次使用其中 3 份训练, 在第 4 份上调整参数, 在最后 1 份上进行测试.

为了进行对比我们分别引入了两个 baseline 系统, 其中 baseline1 系统是 kaggle 在该比赛中用的 baseline, 使用文章的单词数和文章的字符个数对文章的得分进行预测. 对于 baseline2 系统我们提取了一些目前系统中常用的特征, 文章的字符长度, 文章的单词长度, 文章中疑问句和感叹句个数, 高级词汇个数, 拼写错误个数, 停用词个数, n-gram 和 POS n-gram 等特征, 并且结合 Hongbo Chen 于 2012 年发表在 IEEE 上的文章[16]中使用的特征, 来进行评分.

5.2 特征效果对比

我们首先使用最简单的方法来对比不同特征对于系统的影响, 直接用支持向量机(SVM)对于提取的特征进行回归[14]. 具体使用的是 libsvm[15].

表 5 实验结果

数据集	系统	r	$\bar{d}$	σ^2
子集 1	baseline1	0.791	0.679	0.818
	baseline2	0.833	0.680	0.800
	语言学特征	0.845	0.649	0.674
子集 2	baseline1	0.679	0.489	0.317
	baseline2	0.716	0.420	0.287
	语言学特征	0.725	0.416	0.284

© 中国科学院软件研究所 http://www.c-s-a.org.cn

子集 3	baseline1	0.732	0.392	0.341
	baseline2	0.733	0.445	0.328
	语言学特征	0.714	0.438	0.328
子集 4	baseline1	0.757	0.457	0.409
	baseline2	0.758	0.471	0.360
	语言学特征	0.772	0.468	0.356
子集 5	baseline1	0.812	0.395	0.321
	baseline2	0.818	0.455	0.320
	语言学特征	0.818	0.445	0.311
子集 6	baseline1	0.703	0.487	0.499
	baseline2	0.734	0.489	0.376
	语言学特征	0.782	0.479	0.366
子集 7	baseline1	0.657	2.568	10.926
	baseline2	0.787	2.239	8.203
	语言学特征	0.803	2.167	7.457
子集 8	baseline1	0.539	3.888	24.7
	baseline2	0.721	3.150	16.242
	语言学特征	0.723	3.154	16.136

如表 5 中所示, 我们的语言学特征在这 8 个子集中的 7 个子集上取得了最高的人机评测相关系数. 下面我们看一下这 8 个子集上的整体评测效果, 因为每个子集的评分区间不同, 所以我们先对得分区间进行归一化, 其公式如式(17)所示.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{17}$$

其中, x 表示文章的得分, $\max(x)$表示 x 所在子集的最高分, $\min(x)$表示 x 所在子集的最低分.

归一化之后, 我们再来看 8 个子集上的整体效果. 从全部数据结果来看, 相比于 baseline1 系统和 baseline2 系统, 语言学特征系统评分在相关系数方面分别取得了 14.1%和 5.4%的性能提升.

表 6 8 个子集整体效果

系统	r	$\bar{d}$	σ^2
baseline1	0.667	0.161	0.058
baseline2	0.722	0.153	0.053
语言学特征	0.761	0.146	0.052

5.3 自编码器

下面我们使用编码器对于提取到的特征进行重构, 语言学特征系统提取到的原始特征总共 194 维, 我们分别进行压缩编码和稀疏编码, 实验效果如表 7 所示.

表 7 自编码器特征重构

重构维数	r	$\bar{d}$	σ^2
20 维	0.29	0.366	0.107
40 维	0.303	0.358	0.105
80 维	0.669	0.159	0.058
100 维	0.787	0.142	0.049
140 维	0.463	0.203	0.078
200 维	0.37	0.278	0.085
400 维	0.453	0.233	0.081
600 维	0.298	0.363	0.106
1000 维	0.284	0.371	0.109
2000 维	0.224	0.387	0.109

表 7 中 20 维~140 维是使用自编码器进行特征的压缩编码, 200 维~2000 维是使用自编码器进行特征的稀疏编码. 其中特征压缩到 100 维时, 此时的相关系数达到 0.787, 比直接使用支持向量机进行回归效果提升了 3.4%.

5.4 特征值离散化

我们再结合特征值的离散化, 先使用基于于信息增益的方法, 将连续特征离散化到高维的二值特征. 再使用自编码器来进行特征压缩. 其实验效果如表 8 所示. 使用特征值离散化后, 原始特征 194 维扩展到了 15800 维二值的 0,1 特征. 我们再使用自编码器对这 15800 维特征进行压缩重编码. 如表 8 所示, 当自编码器将特征维数压缩到 1000 维时相关系数达到 0.803, 相比于未经过特征值离散化效果提升了 2.0%.

表 8 特征离散化后的特征重编码

重构维数	r	$\bar{d}$	σ^2
500 维	0.678	0.159	0.054
1000 维	0.803	0.139	0.046
2000 维	0.609	0.172	0.062
4000 维	0.604	0.173	0.062
6000 维	0.553	0.181	0.067
8000 维	0.474	0.192	0.075

5.5 分层多项模型

考虑到我们使用支持向量机进行回归的输出结果是连续性的值, 而人工评分给出的是离散化的得分值, 因此我们可以尝试使用分类的方法进行自动评分. 然而一般的分类方式其类别和类别之间没有嵌套或者大小的关系, 这和我们的任务相违背. 这里我们使用分层多项模型(Hierarchical Multinomial Model)来进行分类, 在该模型中, 类别和类别之间有嵌套包含的关系, 这和我们自动评分任务中得分和得分之间的关系非常吻合. 具体使用的是 matlab 实现的机器学习工具

© 中国科学院软件研究所 http://www.c-s-a.org.cn

包[1]. 为了进行对比, 我们同样将支持向量机的评分结

表 9 回归和分类结果对比

模型	r	$\overline{d}$	σ^2
支持向量机规整	0.774	0.143	0.051
分层多项模型	0.792	0.141	0.048

如表 9 所示, 支持向量机回归输出的是离散的值, 其对作文评分的结果在规整到人工评测的边界之后, 人机相关系数从 0.803 下降到 0.774. 相比之下, 分层多项模型虽然给出的人机相关系数是 0.792, 但是因为是分类的结果所以不需要进一步规整, 相比于支持向量机的结果显然更优.

我们对于 baseline1 和 baseline2 同样加入了自编码器, 特征离散化, 分层多项模型进行测试. 结合语言学特征模型, 这三组系统的实验效果如图 2 所示. 纵向比较来看, 无论哪一组实验, 我们的语言学特征系统和两个 baseline 比较, 均能取得最优的效果. 横向来看, 相比于最原始的支持向量机回归, 我们的自编码器, 特征值离散化的使用均能使得系统的性能得到进一步提高. 因为回归得出的结果是连续性数值, 输出得分在规整之后系统性能必然会有所下降. 最后我们使用分层多项模型进行分类, 直接给与一篇作文输出离散的得分结果, 这相比于回归之后再规整的结果人机相关系数更高.

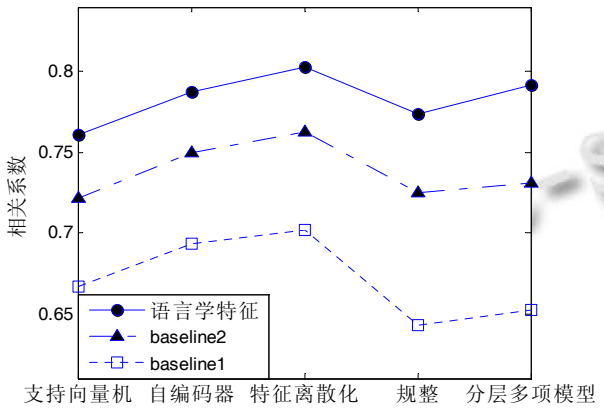

图 2 三组系统实验效果

5.6 主题无关性

最后我们从主题依赖性的角度出发来考察这 3 组系统. 因为数据集包含 8 个子集, 因此我们将全部数据按主题的不同分成 5 份进行交叉验证, 使得训练用的作文和测试作文之间没有主题交叉. 其实验效果如下表所示. 可以看出, 两个 baseline 系统, 特别是 baseline2 系统中引入了大量 n-gram 等和文章主题相关的特征. 这直接导致了在预测其他主题作文时系统性能的下降. 而我们的语言学特征系统使用的都是主题无关特征, 在面对不同主题的测试样本时, 依然能保持很好的鲁棒性.

表 10 8 个子集间相互进行交叉验证

系统	r	$\overline{d}$	σ^2
baseline1	0.622	0.165	0.066
baseline2	0.636	0.165	0.065
语言学特征	0.745	0.151	0.045

6 总结

本文依据英文写作的技巧, 提取了大量的主题无关特征. 然后通过特征离散化减少异常样本对系统的干扰, 自编码器对特征进一步重构以提高特征表达能力. 最后我们分析了作文评分任务的特点使用分层多项模型来输出文章的最终得分. 实验表明, 一方面我们的模型和特征要显著优于传统的方法, 另一方面我们的系统在测试不同主题的作文时显示出了良好的主题无关性.

参考文献

1 梁茂成,文秋芳.国外作文自动评分系统评述及启示.外语电化教学,1997:18–24.

2 Attali Y, Burstein J. Automated essay scoring with e-rater®V. 2. The Journal of Technology, Learning and Assessment, 2006, 4(3): 3–30.

3 Daigon A. Computer grading of English composition. The English Journal, 1966, 55(1): 46–52.

4 Landauer TK. Automatic essay assessment. Assessment in education: Principles, policy & practice, 2003, 10(3): 295–308.

5 Dale R, Anisimoff I, Narroway G. HOO 2012: A report on the preposition and determiner error correction shared task. The 7th Workshop on the Innovative Use of NLP for Building Educational Applications. June 3-8, 2012. 54–62.

6 Ng HT, Wu SM, Wu Y, et al. The CoNLL-2013 shared task on grammatical error correction. Proc. of the Seventeenth

[1] http://cn.mathworks.com/help/stats/index.html

© 中国科学院软件研究所 http://www.c-s-a.org.cn

Conference on Computational Natural Language Learning. August 8–9, 2013.1–12.

7 Larkey LS. Automatic essay grading using text categorization techniques. Proc. of the 21st annual international ACM SIGIR conference on Research and development in information retrival. 1998. 90–95.

8 Persing I, Davis A, Ng V. Modeling organization in student essays. Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing. 2010. 229–239.

9 Persing I, Ng V. Modeling prompt adherence in student essays. Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics(ACL). June 2014. 1534–1543.

10 Persing I, Ng V. Modeling thesis clarity in student essays. Proc. of the 51st Annual Meeting of the Association for Computational Linguistics. August4-9, 2013. 260–269.

11 Chen H, He B. Automatic essay scoring by maximizing human-machine agreement. Proc. of the 2013 conference on Empirical Methods in Natural Language Processing. 2013. 1741–1752.

12 Hinkel E. Second language writers’ text: Linguistic and rhetorical features. Routledge, 2002.

13 Marneffe MCD, Cartney BM, Manning CD. Generating typed dependency parses from phrase structure parses. Proc. of Language Resources and Evaluation Conference. 2006.

14 Burges CJC. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 1998, 2(2): 121–167.

15 Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Trans. on Intelligent Systems and Technology. April, 2011.

16 Chen H, He B, Luo TJ, et al. A ranked-based learning approach to automated essay scoring. Second International Conference on Cloud and Green Computing. 2012.

17 Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science, 2006, 313(5786): 504–507.

18 Dougherty J, Kohavi R, Sahami M. Supervised and unsupervised discretization of continuous features. Machine learning: Proc. of the Twelfth International Conference. 1995. 12. 194–202.

19 Dougherty J, Kohavi R, Sahami M. Supervised and unsupervised discretization of continuous features. Machine Learning: Proc. of the 12th International Conference. San Mateo. Morgan Kaufmann Publishers. 1995. 194–202.

20 葛诗利.面向大学英语教学的通用计算机作文评分和反馈方法研究[博士学位论文].北京:北京语言大学,2008.

www.c-s-a.org.cn

© 中国科学院软件研究所 http://www.c-s-a.org.cn