

基于文本分类方法识别《史记》的伪作

赵建明¹ 李春晖² 姚念民² 姚念军³

(福建师范大学福清分校电子与信息工程学院 福清 350300)¹
(大连理工大学 大连 116024)² (大庆油田热电厂 大庆 163711)³

摘 要 使用基于机器学习的文本分类方法对《史记》的伪作识别进行了研究。《史记》是我国第一部纪传体通史,其伪作的识别历来是其研究中的重点问题。但传统的研究方法较为主观,不能定量,且多种结论互相矛盾。文中提出一种基于文本分类技术的伪作鉴定方法,对《史记》的伪作识别给出定量的研究。该方法具有通用性,可以应用于多个文史问题的研究中。

关键词 机器学习,自然语言处理,史记,伪作

中图分类号 TP391.1 文献标识码 A

Fake Chapters Recognition in Shiji Based on Text Classification Methods

ZHAO Jian-ming¹ LI Chun-hui² YAO Nian-min² YAO Nian-jun³

(School of Electronic and Information Engineering, Fuqing Branch of Fujian Normal University, Fuqing 350300, China)¹

(Dalian University of Technology, Dalian 116024, China)²

(Daqing Oil Field Thermal Power Plant, Daqing 163711, China)³

Abstract The text classification methods based on the machine learning are used to study how to distinguish the fake chapters in Shiji. Shiji is the first general history book in our country and distinguishing fake chapters is always one of the main problems in its study. But the traditional methods are subjective and can not apply quantitative analysis, and even many famous results are contradictive. In this paper, a method of distinguishing fake chapters was presented which can do quantitative research on this subject. This method can also be used in many other studies of the history.

Keywords Machine learning, Natural language processing, Shiji, Fake

1 引言

《史记》是我国第一部纪传体通史,被称为“二十四史”之首,记载了从上古传说中的黄帝时代到汉武帝元狩元年的共3000多年的历史。其记录历史的严谨性和叙事的文学性一直为历代研究者所称道,鲁迅称之为“史家之绝唱,无韵之离骚”。但《史记》流传久远,其中一些篇章可能是后人加入或篡改的,如何识别其中的伪作历来是史学家和古代文学研究者的热点问题。目前自然语言处理方法方兴未艾,出现了以机器学习理论为基础的多种有效处理现代语言的方法。但利用这些方法处理古代汉语还是一个新领域,本文试图应用文本分类方法研究《史记》的伪作识别,也为使用最新的文本处理方法来处理古汉语探索新思路。

2 相关工作

国内外已有大量关于《史记》的研究。文献[1]评述了百年间《史记》的研究状况。《史记》中的真伪之辩历来是研究的重点。1917年,王国维根据甲骨文卜辞纠正了《史记》中的上甲以后几位殷先王先公的次序,崔适在《史记探源》中将《史记》中涉及古文经的内容一律删除,认为这些是刘歆所续。廖季平在民国初年提出过“屈原并没有这个人”的说法。因此

《史记·屈原列传》是后人伪造的,附和这个观点的有胡适、何天行、卫聚贤等人。文献[2]反驳了这个说法。国外学者对《史记》的研究也蔚为壮观。文献[3]梳理了百年来美国学者从文本、史学、文学和哲学视角研究《史记》的具体状况,评说他们在研究《史记》时提出的新见解和新观点。对《史记》伪作的辨识也是其中的重要内容,如马丁·克恩(Martin Kern)在《关于〈史记〉卷24乐书的真实性及意识形态的注释》中指出乐书这一卷并非出自司马迁之手。2003年,他又发表《司马迁〈史记〉中的司马相如传与赋的问题》,认为现在《史记》中的这一卷并不是出自汉武帝时期。倪豪士(William H. Nienhauser)则在其领衔的译著《史记·汉本纪》序言中对此观点进行了驳斥。日本学者泷川资言在《史记会注考证·史记总论》“史记附益条”中写道:涉及《史记》补窜的篇目有34篇。

以上研究的结论很多,有的互相矛盾。其研究方法大都是由研究者找出文中的一些例子进行引证,然后进行主观的分析,找出矛盾的地方,最后得出结论。但这种传统研究方法非常主观,结论不一定具有说服力,这从多位学者的互相矛盾的结论中可见一斑。传统研究方法的缺陷在于其例证只是少量的孤证,一般不具有代表性和普遍性,据此得出的结论也就缺乏可信度且不能被量化。本文使用机器学习技术中的文本分类方法对《史记》进行分析,得出的结论有一定的数据作为

支持,可以进行量化分析,具有一定的客观性。虽然已有很多学者将文本分类技术应用于自然语言处理,但将该技术应用于古代汉语的分析还很少见的。

使用机器学习的方法对文本进行分类的研究是近年来的热点,如文献[5-6],但将该技术用于古文特别是《史记》的研究还较少见。

3 文本分类工具介绍

Scikit-learn 是基于 Python 的机器学习开源工具包。基本功能主要被分为 6 个部分:分类、回归、聚类、数据降维、模型选择、数据转换。以下简单介绍本文用到的数据转换、模型选择和分类功能。

Scikit-learn 提供了 TfidfVectorizer 类用于文档向量化,它支持把字符串形式的文档转换成词袋(Bag of Words)形式的向量。当文档以词袋形式向量化时,不保存文档中词语之间的顺序信息。通常将一个给定的词语在文档中出现的次数作为词袋向量中该词语的权重,称之为词频(Term Frequency, TF)权重表示。如式(1)所示,式中分子是该词在文件中出现的次数,而分母则是在文件中所有字词出现的次数之和。

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (1)$$

词频通常会被正规化,以防止它偏向长的文件。正规化有两种方式,分别称为 L1 正规化和 L2 正规化。L1 正规化是用词语出现次数除以文档中所有词语出现总数。L2 正规化是用词语出现次数除以文档中各词语出现次数的平方和的平方根。

如果把词频向量中所有非零元素设置为 1,即只考虑词语在该文档中是否出现,则称为布尔权重表示。

在词频基础上乘以 IDF 反文档频率(Inverse Document Frequency),则称为 TF-IDF 权重表示,如式(2)所示。某一特定词语的 IDF,可以由总文件数目除以包含该词语之文件的数目,再将得到的商取对数得到。IDF 的主要思想是:如果包含词条 t_i 的文档越少, IDF 越大,则说明词条 t 具有很好的类别区分能力。

$$idf_i = \log \frac{|D|}{|\{j: t_i \in d_j\}|} \quad (2)$$

其中, $|D|$ 是语料库中的文件总数,分母是包含词条 t_i 的文档个数。

Scikit-learn 提供了 GridSearchCV 类用于模型的参数搜索, GridSearchCV 在数据集上产生 K 对训练/确认集。对给定参数,在 K 对训练集上分别训练并在确认集上进行检验,最后取 K 次实验的平均得分作为该模型在该参数上的得分估计。在完成参数空间搜索后,给出该模型在该数据集上的最优参数估计。

Scikit-learn 提供了众多的分类模型供研究者解决问题所用,每个模型都有其适合的应用领域。分类模型可以划分为线性模型,支持向量机、最近邻分类、朴素贝叶斯、决策树、集成方法(ensemble methods)等几大类。

本文选择以下模型作为候选模型:线性模型中的随机梯度下降分类(SGDClassifier),逻辑回归分类(LogisticRegression),支持向量机中的 LinearSVC 和 SVC,朴素贝叶斯中的多项式朴素贝叶斯(MultinomialNB),最近邻分类中的 K 近

邻分类(KNeighborsClassifier),集成方法中的随机森林分类器(RandomForestClassifier)。

4 实验设计

文中所提的伪作识别方法基于一个假设:由于《史记》的主体由司马迁撰写,其写作风格和用词规律在全书会保持一定的一致性。如果是伪作,则其写作风格与用词规律会与原书有一定的偏离。基于此假设,如果使用文本分类方法对《史记》与疑似伪作进行分类处理,则伪作很可能会与原作分为不同的类。基于该思路,设计实验如下。

挑选《史记》、《战国策》、《汉书》和《吴越春秋》4 部古籍作为样本数据,其中《史记》成书时间约为公元前 93 年,《战国策》成书时间约为公元 6 年,《汉书》和《吴越春秋》的成书时间约为公元 83 年。由于成书时间大致相同,把《汉书》和《吴越春秋》作为一类,这样一共有 3 个类别。

我们需要构建用于文本分类的系统,其工作流程如下。

(1)预处理:需要对下载的 txt 格式文件进行预处理,清除掉与训练数据无关的内容,并对书本按卷进行切分。经过预处理,得到《史记》130 卷(篇),《战国策》33 卷,《汉书》120 卷,《吴越春秋》10 卷。

(2)数据切割:把样本数据分为 3 个部分,其中《史记》中疑为伪作的 10 卷作为保留集,在剩余各 283 卷中随机选择 80% 作为训练集用于模型选择及训练,其余 20% 作为测试集用于验证模型的有效性。

(3)繁简转换:原始数据中存在部分繁体字,如果不进行处理,将成为分类的重要特征,对分类结果产生重大影响,如‘於’、‘後’、‘餘’等,将其统一转换成简体字进行处理。

(4)分词:由于没有标注好的古汉语语料,使用基于现代汉语语料进行训练得到的分词模型在古汉语中效果不理想。考虑到古汉语单字词出现频次高的情况,采用了基于单字的切分方案,即把一个汉字作为一个词。

(5)文档向量化:要使机器学习算法能够处理文档,必需把文档表示成向量形式。文中采用常用的词袋(Bag of Words)^[2]表示法进行文档表示。

(6)模型选择及参数确定:基于 Scikit-learn 对在多个模型上对向量化表示的文本数据进行交叉确认,以确定适用于该样本集的分类模型及该模型的参数。

(7)模型验证:使用测试集验证模型的有效性,估计模型的精确度。

(8)预测:对保留集数据进行预测,输出分类结果。

5 实验结果

基于 Linux 系统用 Python 实现了文本分类系统,分别使用布尔模型^[3]、词频(TF)、词频-逆文档词频(TF-IDF)^[4]对经过预处理后的数据进行文档向量化,使用 Scikit-learn 的网格搜索功能在 7 个分类模型上对不同参数进行五折交叉确认,以确定最优模型及相关参数,实验结果如表 1 所列。图 1 以柱状图的形式展示了不同分类器在 3 种文档向量化方式下的五折交叉验证的结果,从图中可以很清楚地看到,在不同向量化方式下,逻辑回归模型和随机森林分类器都取得了比较好的结果。其中逻辑回归模型在参数设置为{'penalty': 'l1', 'C': 100}时达到了最高的分类精确度 93.7%。

表 1 网格搜索结果

模型名称	参数搜索设置	向量化表示	最优参数	得分
SGD	{ 'alpha': (0.0001, 0.00001, 0.000001), 'penalty': ('l1', 'l2', 'elasticnet') }	BOOL	{ 'penalty': 'l2', 'alpha': 1e-05 }	0.865
		TF	{ 'penalty': 'l1', 'alpha': 1e-06 }	0.743
		TF-IDF	{ 'penalty': 'l1', 'alpha': 0.0001 }	0.711
LR	{ 'C': (1, 10, 100, 1000), 'penalty': ('l1', 'l2') }	BOOL	{ 'penalty': 'l1', 'C': 100 }	0.937
		TF	{ 'penalty': 'l1', 'C': 1000 }	0.838
		TF-IDF	{ 'penalty': 'l1', 'C': 100 }	0.895
LinearSVC	{ 'C': (1, 10, 100, 1000), 'loss': ('hinge', 'squared_hinge') }	BOOL	{ 'loss': 'hinge', 'C': 1 }	0.822
		TF	{ 'loss': 'hinge', 'C': 10 }	0.779
		TF-IDF	{ 'loss': 'hinge', 'C': 100 }	0.789
SVC	{ 'kernel': ['rbf', 'linear'], 'C': [1, 10, 100, 1000] }	BOOL	{ 'kernel': 'linear', 'C': 1 }	0.854
		TF	{ 'kernel': 'linear', 'C': 100 }	0.701
		TF-IDF	{ 'kernel': 'linear', 'C': 100 }	0.763
MultinomialNB	{ 'alpha': (1.0, 0.1, 0.01, 1E-3, 1E-4, 1E-5) }	BOOL	{ 'alpha': 1.0 }	0.818
		TF	{ 'alpha': 1e-05 }	0.512
		TF-IDF	{ 'alpha': 0.01 }	0.737
Kneighbors	{ 'n_neighbors': (5, 10, 15, 20), 'weights': ('uniform', 'distance') }	BOOL	{ 'n_neighbors': 20, 'weights': 'uniform' }	0.807
		TF	{ 'n_neighbors': 20, 'weights': 'distance' }	0.728
		TF-IDF	{ 'n_neighbors': 20, 'weights': 'uniform' }	0.672
RandomForest	{ 'n_estimators': (10, 50, 100) }	BOOL	{ 'n_estimators': 50 }	0.847
		TF	{ 'n_estimators': 100 }	0.864
		TF-IDF	{ 'n_estimators': 100 }	0.856

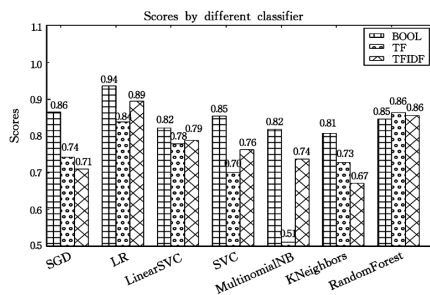


图 1 不同分类模型在 3 种文档向量化方式下的得分

选择具有最高得分的逻辑回归模型及参数设置{ 'penalty': 'l1', 'C': 100 }作为分类模型,使用布尔模型作为文档向量化方式。用训练集数据重新训练模型,之后在测试集数据上检测模型的有效性。表 2 列出了在测试集上的实验结果。

表 2 逻辑回归模型在测试集上的结果

	precision	recall	f1-score	support
《汉书》《吴越春秋》	0.96	0.96	0.96	26
《史记》	0.96	1.00	0.98	24
《战国策》	1.00	0.86	0.92	7
avg / total	0.97	0.96	0.96	57

从实验数据可知,逻辑回归模型在测试集上取得了比较好的效果,模型的平均精确度、召回率和 F1 值都在 0.96 以上。

用该模型对保留集中疑似伪作的文章进行预测,表 3 列出了逻辑回归模型的输出结果。

表 3 逻辑回归模型在保留集上的预测输出

卷名	预测结果
孝景本纪	《史记》
孝武本纪	《汉书》《吴越春秋》
律书	《史记》
礼书	《史记》
傅靳蒯成列传	《汉书》《吴越春秋》
龟策列传	《史记》
汉兴以来将相名臣年表	《汉书》《吴越春秋》
三王世家	《史记》
乐书	《史记》
日者列传	《史记》

从表 3 中可以看出,《孝武本纪》、《傅靳蒯成列传》和《汉兴以来将相名臣年表》是伪作的可能性最大。

结束语 本文使用较新的文本分类技术对《史记》的伪作识别进行了研究,其结论与目前流行结论基本一致。但该结论具有更高的客观性,与传统方法相比,该结果是对原文本进行大量统计分析得出的结果,而不是使用孤证进行分析后的片面结论,因此该结论更具可信度。

但本文所使用的方法还较初步。由于古汉语的标注语料库较少,文中在一些处理中只能进行简化处理,如果能对古汉语进行正确分词,其预期结果将会更精确。由于缺乏已标注的古汉语语料,且在实验中试验了使用现代汉语分词工具对古汉语进行分词时的实验结果较差,因此选择了使用单字分词方法,从最后的结果看,其实验效果较好,但尚有改进空间。

总体上,本文方法具有通用性,可以对多个古汉语研究的问题进行数量认证;我们也正在对该方法进行改进,以使其能应用于更多的文史课题。

参 考 文 献

- [1] 陈桐生. 百年《史记》研究的回顾与前瞻[J]. 文学遗产, 2001(1): 120-128.
- [2] 王开富. 《史记·屈原列传》非伪作辨[J]. 重庆师院学报(哲学社会科学版), 1984(2): 53-57.
- [3] 吴原元. 百年来美国学者的《史记》研究述略[J]. 史学集刊, 2012(4): 59-68.
- [4] 泷川资言. 《史记会注考证附校补·史记总论》“史记附益条”[M]. 上海: 上海古籍出版社, 1986.
- [5] 黄伟. 基于多分类器投票集成的半监督情感分类方法研究[D]. 上海: 上海交通大学, 2015.
- [6] 黄仁, 张卫. 基于 word2vec 的互联网商品评论情感倾向研究[J]. 计算机科学, 2016, 43(S1): 387-389.