情报学报 ISSN 1000-0135
第32卷第10期1058-1067,2013年10月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol. 32 No. 10, 1058-1067 October 2013

doi:10.3772/j.issn.1000-0135.2013.10.005

基于CART分类方法的期刊操纵引用行为识别建模研究

孙建军[1] 鞠秀芳[1] 裴 雷[1] 郑彦宁[2] 潘云涛[2]

(1.南京大学信息管理学院,南京 210093; 2.中国科学技术信息研究所,北京 100038)

摘要 当前,一些学术期刊在利益的驱使下,通过大量自引和结成"互引同盟"的方式快速提高被引频次和影响因子等指标,影响了引文分析的公平性。基于此,本文首先利用数据挖掘中的CART分类算法构建期刊操纵引用行为的识别模型,设计了识别操纵引用行为的4个评价指标:自引率、被引年代分布、被引密度比和引用密度比。并采用国内某引文数据库中的50本综合性社会科学期刊作为实验样本,采集该期刊群2009年的引文数据作为训练数据集,2008年的引文数据作为验证数据集。最后,运用2010年的引文数据对期刊操控行为识别模型的有效性进行验证,实验结果证明,本文构建的分类模型可以有效地对期刊引用操纵行为进行识别。

关键词 期刊引用操纵行为 CART算法 自引率 被引年代分布 被引密度比 引用密度比

Research on Model for Recognition of Journal Citation Manipulation Behavior Based on CART

Sun Jianjun[1], Ju Xiufang[1], Pei Lei[1], Zheng Yanning[2] and PanYuntao[2]

(1. *School of information management, Nanjing University, Nanjing* 210093;
2. *Institute of Scientific and Technical Information of China, Beijing* 100038)

Abstract Now some academic journals are driven by the interests in order to improve their cited frequencies and the impact factors of journals quickly by a large number of self-citations or by forming a citation alliance, which affects the fairness of citation analysis. According to the background, the paper first constructs a journal citation manipulation behavior recognition model by CART classification algorithm in data mining, and designs four evaluation indexes for recognizing the manipulation behavior: self-citation rate, cited era distribution, cited density ratio, citation density ratio. Then an experiment was carried out to verify the model with the data collected from a citation database of China. The experiment takes 50 journals in the field of comprehensive social science as its experimental sample, collects the citation data of these journals in 2009 as the training data set and takes the citation data of these journals in 2008 as the validation data set. Finally, the article chooses the citation data of these journals in 2010 to identify the validity of the journal manipulation behavior recognition model. The experiment result showed that the model can effectively recognize the journal citation manipulation behavior.

Keywords journal citation manipulation behavior, CART algorithm, self-citation rate, cited era distribution, cited density ratio, citation density ratio

收稿日期:2013年1月10日

作者简介:**孙建军**,男,1962年生,南京大学信息管理学院院长,博士,教授,博士生导师,主要研究方向:网络信息资源管理与信息经济。**鞠秀芳**,女,1976年生,博士,副研究员,主要研究方向:文献计量、期刊评价。E-mail:jxf@nju.edu.cn。**裴雷**,男,1981年生,南京大学信息管理学院副教授,硕士生导师,主要研究方向:信息政策与信息行为分析。**郑彦宁**,男,1965年生,研究员,博士生导师,主要研究方向:情报技术与方法、竞争情报。**潘云涛**,女,1967年生,研究员,主要研究方向:科学计量学。

1 引 言

科学文献的引证与被引证,是科学发展规律的表现,体现了科学知识和情报内容的积累性、连续性和继承性。文献引证有多方面的原因,一般多为对开拓者表示尊重、对有关著作给予肯定或验证其研究方法等;但也存在一些不良的动机和行为。由于自引可以增加刊物的被引频次[1],互引则可以同时增加刊物的总被引频次和期刊的影响因子,而被引频次和影响因子都是评价期刊影响力的最重要的指标[2]。为了增强自身的竞争力,学术期刊界不恰当的引用,包括自引、互引应运而生。这样的行为干扰了文献间的正常交流,也对引文分析的数据造成了污染,影响了期刊评价的公平性。而对于传统的引文分析而言,主要采用被引频次、影响因子等指标进行期刊评价,这些指标显然无法有效识别和评估期刊操纵引用行为。

在以往探讨期刊引用行为的研究中,文献计量方法与行为研究方法被广泛应用。前者多关注引用类型划分、引证趋势分析等,如 John Mingers 运用一个随机模型发现一个期刊未来的引文与以前引文之间的线性关系随时间的推移递减[3]。后者主要"引用动机"、"引用习惯"角度展开,研究引用行为与期刊特征、文献类型、用户动机以及知识关联等影响因素的相互关系,如 Liming Liang 通过构建出版——引用(p-r)矩阵,以美国信息科学与技术学会杂志(JASIST)和文献学期刊(JDOC)两本杂志的刊出规律进行分析[4]。马凤等从论文引用动机的角度,采用问卷调查形式,分别分析了中国期刊研究界和情报学界的引用动机[5]。刘筱敏等采用的电子期刊全文下载量为对象,运用 Person 相关系数证明了下载行为与引用行为的正向相关关系[6]。杨利军等从引用文献的时间、类型、数量三个维度发现引用习惯对期刊论文被引频次具有显著性影响[7]。李睿等从知识关联揭示差异的角度比较了专利引用行为与期刊论文引用行为的异同[8]。

在研究中,期刊的互引现象和操纵引用行为也引起了国内学者的关注。马峥等学者通过将研究对象从特定期刊、某一领域期刊拓展到互引期刊对或互引期刊群,通过计算各期刊之间的互引矩阵,发现中国科技核心期刊分类互引网络示意图可以应用于中国"集团非正常互引"的防范[9],但未针对期刊操作引用行为的甄别进行研究。鉴于此,本文借用数据挖掘建模方法对国内人文社会科学中有操纵引用嫌疑的期刊进行甄别。利用数据挖掘中的 CART 算法来构建能够有效识别期刊操纵行为的模型,通过国内某著名引文数据库中收集大量正常引用的期刊引文数据与有操纵引用行为的期刊引文数据作为样本,提取出有效特征值,形成一个大的矩阵模型,选取 CART 算法作为分类器对样本训练,得出数据模型结果并对其分析,在此基础上选取同组期刊不同年代的特征值对建立的模型的有效性进行验证。利用数据挖掘建模方法来甄别正常引用的期刊与操纵引用的期刊,希望借此控制恶意引用行为的蔓延和扩大的趋势,从而可以督促期刊正常、有序、规范化发展,促进期刊发展步入正轨。

2 研究方法

2.1 分类算法

分类算法是属于预测式数据挖掘的一种数据分析方法,目的是根据样本数据集找出能准确描述并区分数据类或概念的模型。分类就是根据数据集的特点找出类别的概念描述,这个概念描述代表了这类数据的整体信息,也就是该类的内涵描述,并使用这种类的描述对未来的测试数据进行分类。

目前的分类技术有很多,如决策树、贝叶斯网络、神经网络、遗传算法、K-最近邻分类等。其中决策树方法是应用较广泛的算法,其思路是找出最有分辨力的分类属性,把数据库划分为许多子集(每个子集对应树的一个分枝),构成一个分枝过程,然后对每一个子集递归调用分枝过程,直到所有子集包含同一类型的数据。它的每一个树节点可以是叶结点,对应着某一类,也可以对应着一个划分,将该节点对应的样本集划分成若干个子集,每一个子集对应一个节点。其中树的每个节点对应一个非类别属性,每条边对应这个属性的每种可能值,而树的每个叶结点代表一个类别[10]。

2.2 CART 算法

决策树算法的研究发展到现在,学术界已经先后提出了多种不同的算法,常用算法主要包括 ID3 算法、CART 算法、C4.5 算法、CHAID 算法、PUBLIC 算法、SLIQ 算法以及 SPRLNT 算法等。在本文中将使用 CART 算法进行建模,CART(Classification and Regression Trees)算法是由 L. Breiman 等提出的一

种使用非参数方法的二进制递归分类算法，算法采用一种二分递归分割的技术将预测空间递归划分为若干子集，树中的叶节点对应着划分的各个区域，这种划分是由与每个内部节点相关的分支规则(Splitting Rules)确定的[11]。

3 基于CART算法的期刊引文操纵行为建模及实证研究

3.1 指标选取和向量构建

对于分类模型而言，样本的特征选择以及构建样本的空间向量是极为重要的一步。对于期刊引用操纵行为而言，根据笔者多年来从事引文工作的经验，目前各期刊的操纵引用行为主要包括两种，一种是通过大量自引提升总被引频次，第二种是多个期刊之间结成“互引同盟”大量互引。基于这两种操纵行为，本文提出四种指标来描述期刊特征，构建样本向量，识别操纵引用行为，即自引率、被引年代分布、被引密度比和引用密度比。

(1)自引率

期刊自引率指某期刊全部被引次数中，被该刊本身引用的次数所占的比例，其定义为：

$$\text{期刊自引率} = \frac{\text{被某刊自己引用的次数}}{\text{期刊总被引次数}} \quad (1)$$

(2)被引年代分布

普赖斯的研究表明期刊的被引一般在发表之后的第二年达到峰值。一般而言，随着时间的推移，论文的价值将逐渐减弱，直至不被引用。但是对于有操纵行为的期刊而言，由于在短时间内大量自引或互引，其被引将会集中在某个较短的时间段，分布不均匀。据此采用期刊出版后两年的被引次数除以总被引次数作为期刊被引年代集中度的计算公式。

$$\text{被引年代集中度} = \frac{\text{期刊出版后两年被引次数}}{\text{总被引次数}} \quad (2)$$

(3)被引密度比

首先引入被引密度的概念，被引密度是指期刊被一群期刊引用的次数与这群期刊的种类数值的比值。

$$\text{期刊被引密度} = \frac{\text{被某群期刊引用次数}}{\text{期刊群期刊种数}} \quad (3)$$

对于有操纵引用行为的期刊而言，其大量被引往往集中分布在几本期刊上，因而定义期刊被引密度比为对某刊的被引做了t%的贡献的核心期刊群的被引密度与该刊总体的被引密度的比值。被引密度比越大，说明期刊被引越是集中，被引密度比越小，说明期刊被引越是均匀。

$$\text{期刊被引密度比} = \frac{t\%\ \text{核心被引密度}}{\text{该刊被引密度}} \quad (4)$$

(4)引用密度比

从数据看，有操纵引用行为的期刊如果被某几本期刊大量引用后，其参考文献的数据中必然会不断出现大量引用的这几本期刊。分析其原因，在引用行为已经成为某种资源的环境下，期刊需要提高他引次数，只有依靠与其他期刊交换引用方可实现，引用密度比的数据可以从另一个角度印证期刊操纵引用的行为。定义引用密度和引用密度比如公式(5)、公式(6)所示：

$$\text{期刊引用密度} = \frac{\text{引用某群期刊总次数}}{\text{期刊群期刊种数}} \quad (5)$$

$$\text{期刊引用密度比} = \frac{t\%\ \text{核心引用密度}}{\text{该刊总引用密度}} \quad (6)$$

使用上述四项指标构建期刊向量为 $X(x_1, x_2, x_3, x_4)$ =(自引率，引文年度集中度，被引密度比，引用密度比)。

3.2 实验样本选取

本文选取该数据库中的50种综合性社会科学期刊作为实验样本，采集该期刊群2009年的引文数据作为训练数据集，2008年的引文数据作为验证数据集，分析各期刊在各年代的自引率、被引年度集中度、被引密度比、引用密度比，并对试验区进行样本选取。为免产生争议，本文不列举期刊具体名称，采用期刊n1－n40来表示正常引用的期刊，m1－m10表示业内普遍认为有操纵引用行为的期刊。根据选定的特征向量，加上是否有操纵引用行为变量judge形成一个50×5的特征值矩阵，如表1所示：

表1 训练样本的特征值矩阵

训练样本	X_1	X_2	X_3	X_4	judge
n1	0.02	0.25	2.182	2.38	0
n2	0.02	0.21	1.892	1.3	0
n3	0.01	0.19	4.536	3.839	0
n4	0.01	0.28	1.624	1.339	0
n5	0.04	0.21	2.75	1.364	0
n6	0.01	0.3	2.2	1	0
n7	0.03	0.26	2.365	2.165	0

续表

训练样本	X_1	X_2	X_3	X_4	judge
n8	0.05	0.38	2.701	2.145	0
n9	0.05	0.38	1.789	1.83	0
n10	0.06	0.39	3.045	1.634	0
m9	0.06	0.41	6.913	3.274	1
n12	0.05	0.37	4.894	2.077	0
n13	0.04	0.4	2.376	2.992	0
n14	0.05	0.34	3.125	2.621	0
m10	0.10	0.4	11.885	6.015	1
n16	0.05	0.39	1.951	1.662	0
m8	0.14	0.59	9.13	8.553	1
n18	0.01	0.28	1.912	2.756	0
n19	0.03	0.25	2.324	2.3	0
n20	0.08	0.33	4.973	2.503	0
n21	0.05	0.38	3.97	6.769	0
n22	0.05	0.27	2.415	2.504	0
n23	0.07	0.26	2.455	2.58	0
n24	0.06	0.32	2.288	2.037	0
m7	0.07	0.71	4.554	2.984	1
n25	0.02	0.26	2.092	2.052	0
m4	0.09	0.4	10.197	3.814	1
m6	0.10	0.49	7.75	2.051	1
n26	0.28	0.38	8.182	5.306	0
n27	0.03	0.39	2.384	1.849	0
n28	0.03	0.27	4.475	2.083	0
n29	0.05	0.39	2.784	2.391	0
m3	0.06	0.44	15.258	3.424	1
m5	0.22	0.47	27.047	3.352	1
n30	0.03	0.26	1.633	2.968	0
n31	0.03	0.29	2.96	2.395	0
n32	0.14	0.26	5.46	5.631	0
n33	0.07	0.3	4.269	2.165	0
n34	0.08	0.38	3.259	2.844	0
n35	0.05	0.35	3.7	4.156	0
n36	0.14	0.26	6.097	1.788	0
m2	0.1	0.65	6.503	12.129	1
n37	0.02	0.35	2.609	1.924	0
n38	0.22	0.22	2.357	2.462	0
n39	0.03	0.3	2.19	2.617	0
m1	0.11	0.48	9.86	7.989	1
n40	0.1	0.38	1.733	1.69	1
n41	0.06	0.35	2.1	1.8	1
n42	0.11	0.38	3.696	2.05	1
n43	0.21	0.51	4.364	1.874	1

注：X_1，X_2，X_3，X_4 分别表示自引率，被引年度集中度，被引密度比，引用密度比，judge = 1 表示期刊受操纵引用，judge = 0 表示期刊正常引用。

3.3 模型训练

根据所选取的训练样本，实验采用 CART 算法，选择测试变量为期刊的 4 个特征属性（x_1、x_2、x_3、x_4），设定目标变量为期刊是否有操纵引用行为 judge：judge = 0 时，判定期刊无操纵引用行为，judge = 1 时，判定期刊有操纵引用行为。算法规则设置如下：

（1）树的最大高度为 5；

（2）最大代理树 5，最小杂质改变为 0.0001，也就是说进行新的划分后，父节点和子节点之间的纯度改变量低于 0.0001 则不再划分；

（3）以 Gini 系数作为分类目标的杂质度量，杂质是指树所定义的子群的输出字段的变化范围；

（4）停止规则：当前节点的样本数少于总样本数的 2%，即不再进行分割；或是分割生成的子节点的样本数少于总样本数的 1%，即不再进行分割；

（5）修剪树的标准错误法则：修剪前的正确率与修剪后的正确率比值为 1；

（6）先验概率：制定不同 judge 值的先验概率为 50%。

实验利用测试变量和目标变量构成的 50 个训练样本，根据上述规则使用 CART 算法生成分类器。由于在确定期刊被引密度比和期刊引用密度比时对核心期刊的 t 值的不确定性，我们对于 $t = 0.2$，$t = 0.3$，$t = 0.4$ 利用 CART 算法生成不同的分类器。由图 1 至图 6 可以看到，t 取不同值时，各变量在模型识别过程中占有不同的权重，生成的分类器和决策树也不相同。其中，$t = 0.3$ 时，自引率、被引年度集中度、被引密度比、引用密度比四个指标的对识别模型的贡献度（权重）为：0.084、0.5、0.334、0.082。

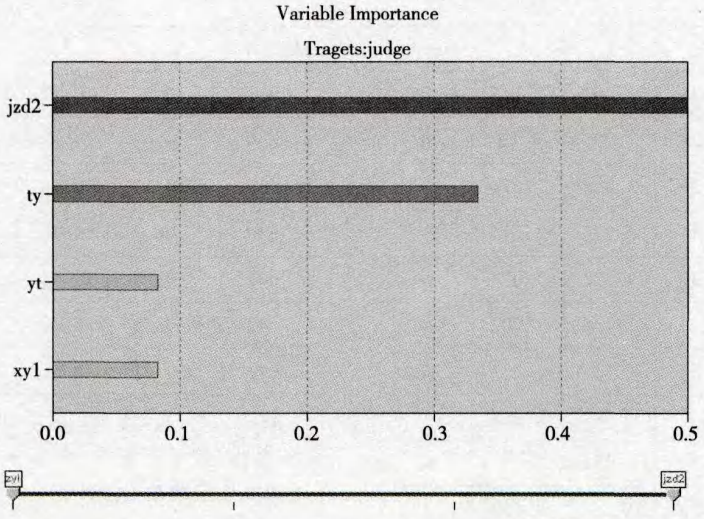

图1 基于 $t=0.2$ 分类器及其变量重要性

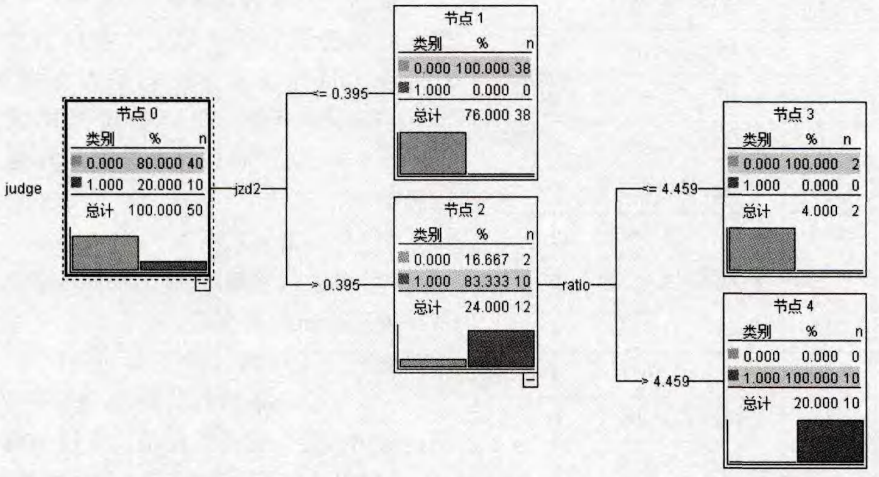

图2 分类器生成的决策树($t=0.2$)

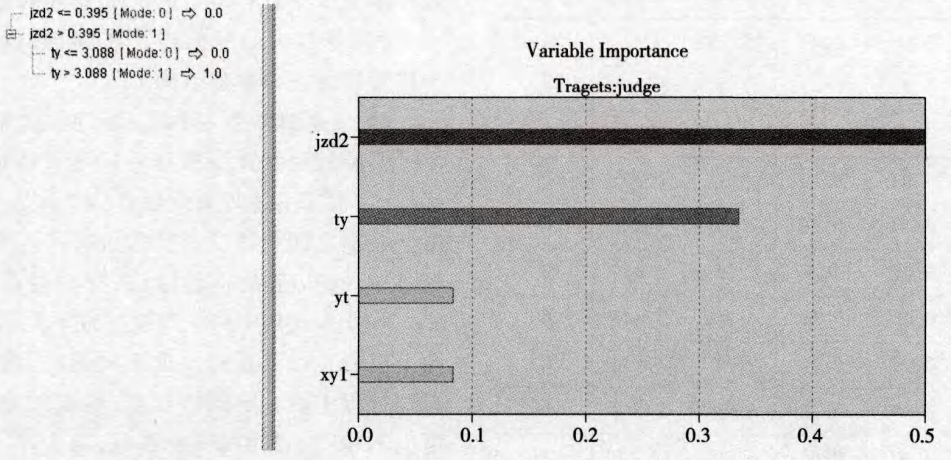

图3 基于 $t=0.3$ 分类器及其变量重要性

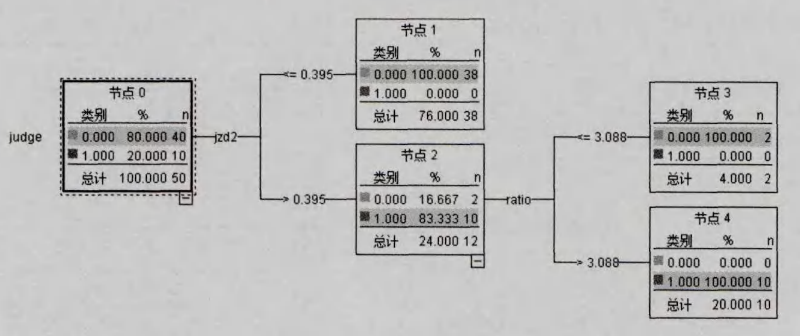

图 4　分类器生成的决策树($t=0.3$)

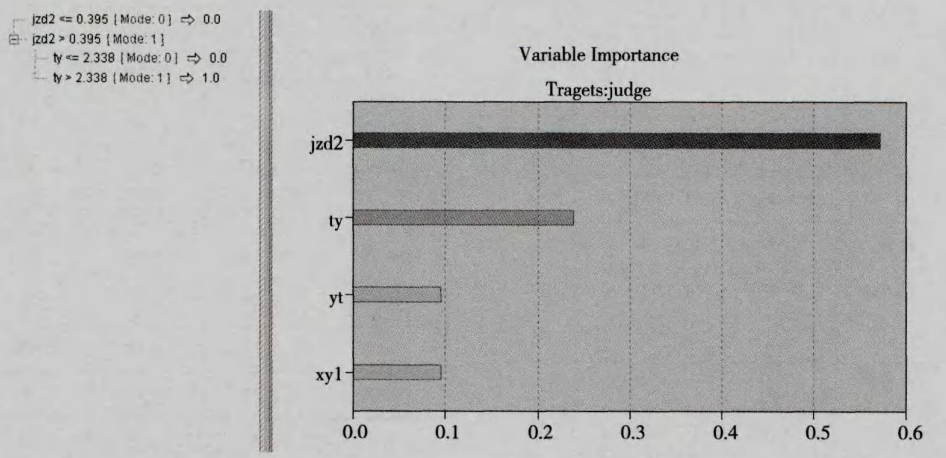

图 5　基于 $t=0.4$ 分类器及其变量重要性

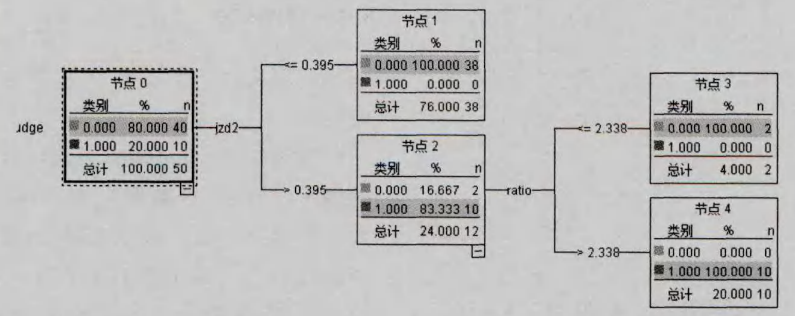

图 6　分类器生成的决策树($t=0.4$)

3.4　模型验证

分类需要将数据集分为训练数据集和检验数据集,在训练数据集上建立模型,然后在检验数据集上评估其质量。在确定衡量期刊被引密度比和引用密度比的核心期刊群时,t 取不同的数值生成的模型不同,但利用训练数据集分类的结果都相同,因此我们不能确定识别度最高的模型中 t 的最优值,因此接下来将在 t 取不同值时,依据前文提及的算法规则,对训练数据集生成的模型进行验证以及评价。

由于期刊发表具有持续性和滞后性,期刊一般会在同一选刊周期内连续操纵期刊引用行为,因此我们取该数据库同一选刊周期内的 2008 年的数据对建立的期刊引用操控行为识别模型进行验证。我们依然以前述 50 种综合性社会科学期刊为实验样本,采集该期刊群在 2008 年的引文数据作为检验数据集(表 2)。按照上文所述建模步骤,分别把正常类期刊和异常类期刊的特征值矩阵放入到已建立的模型中,利用学习生成的模型对期刊进行分类,并将分类结果与已知期刊的分类进行对照,从而判定之

前利用机器学习建成模型的准确性和稳健性。

表2 验证样本的特征值矩阵

验证样本	X_1	X_2	X_3	X_4
n1	0.02	0.25	2.182	2.38
n2	0.02	0.21	1.892	1.3
n3	0.01	0.19	4.536	3.839
n4	0.01	0.28	1.624	1.339
n5	0.04	0.21	2.75	1.364
n6	0.01	0.3	2.2	1
n7	0.03	0.26	2.365	2.165
n8	0.05	0.38	2.701	2.145
n9	0.05	0.38	1.789	1.83
n10	0.06	0.39	3.045	1.634
m9	0.06	0.41	6.913	3.274
n12	0.05	0.37	4.894	2.077
n13	0.04	0.4	2.376	2.992
n14	0.05	0.34	3.125	2.621
m10	0.1	0.4	11.885	6.015
n16	0.05	0.39	1.951	1.662
m8	0.14	0.59	9.13	8.553
n18	0.01	0.28	1.912	2.756
n19	0.03	0.25	2.324	2.3
n20	0.08	0.33	4.973	2.503
n21	0.05	0.38	3.97	6.769
n22	0.05	0.27	2.415	2.504
n23	0.07	0.26	2.455	2.58
n24	0.06	0.32	2.288	2.037
m7	0.07	0.71	4.554	2.984
n25	0.02	0.26	2.092	2.052
m4	0.09	0.4	10.197	3.814
m6	0.10	0.49	7.75	2.051
n26	0.28	0.38	8.182	5.306
n27	0.03	0.39	2.384	1.849
n28	0.03	0.27	4.475	2.083
n29	0.05	0.39	2.784	2.391
m3	0.06	0.44	15.258	3.424
m5	0.22	0.47	27.047	3.352
n30	0.03	0.26	1.633	2.968
n31	0.03	0.29	2.96	2.395
n32	0.14	0.26	5.46	5.631
n33	0.07	0.3	4.269	2.165
n34	0.08	0.38	3.259	2.844
n35	0.05	0.35	3.7	4.156
n36	0.14	0.26	6.097	1.788
m2	0.1	0.65	6.503	12.129
n37	0.02	0.35	2.609	1.924
n38	0.22	0.22	2.357	2.462
n39	0.03	0.3	2.19	2.617
m1	0.11	0.48	9.86	7.989
n40	0.10	0.38	1.733	1.69
n41	0.06	0.35	2.1	1.8
n42	0.11	0.38	3.696	2.05
n43	0.21	0.51	4.364	1.874

续表

注：X_1、X_2、X_3、X_4 分别表示自引率、被引年度集中度、被引密度比、引用密度比。

3.5 模型评价

一般来说，最佳模型是产生最小损失的那个模型。

混淆矩阵(confusion matrix)作为分类规则特征的表示，是区分分类器识别不同样本情况的有用工具。它包括了每一类的样本个数，包括正确的和错误的分类。主对角线给出了每一类正确分类的样本的个数，非对角线上的元素则表示未被正确分类的样本个数。

图7为 $t=0.2$、$t=0.3$、$t=0.4$ 时生成的分类器模型对于测试样本的混淆矩阵，图8、图9、图10分别为 $t=0.2$、$t=0.3$、$t=0.4$ 时的分类器对于验证样本的混淆矩阵。

为了度量分类器的预测精度，假设每个被错分的数据会产生相同的成本，在数据挖掘模型评价中引入误差率和准确率作为性能指标对建立的模型进行评估。

误差率 R 为检验集中误差数目 E 和样本数 S 的比值：

$$R = E/S \tag{6}$$

judge		0.0	1.0
0.0	Count	40	0
	Row %	100.000	0.000
	Column %	100.000	0.000
	Total %	80.000	0.000
1.0	Count	0	10
	Row %	0.000	100.000
	Column %	0.000	100.000
	Total %	0.000	20.000

Results for output field judge
Comparing $R-judge with judge

Correct	50	100%
Wrong	0	0%
Total	50	

图 7　$t=0.2$、$t=0.3$、$t=0.4$ 分类器基于测试样本的的混淆矩阵

judge		0.0	1.0
0.0	Count	36	4
	Row %	90.000	10.000
	Column %	90.000	40.000
	Total %	72.000	8.000
1.0	Count	4	6
	Row %	40.000	60.000
	Column %	10.000	60.000
	Total %	8.000	12.000

Comparing $R-judge with judge

Correct	42	84%
Wrong	8	16%
Total	50	

图 8　$t=0.2$ 分类器基于验证样本的分类结果

judge		0.0	1.0
0.0	Count	37	3
	Row %	92.500	7.500
	Column %	90.244	33.333
	Total %	74.000	6.000
1.0	Count	4	6
	Row %	40.000	60.000
	Column %	9.756	66.667
	Total %	8.000	12.000

Comparing $R-judge with judge

Correct	43	86%
Wrong	7	14%
Total	50	

图 9　$t=0.3$ 分类器基于验证样本的分类结果

judge		0.0	1.0
0.0	Count	36	4
	Row %	90.000	10.000
	Column %	90.000	40.000
	Total %	72.000	8.000
1.0	Count	4	6
	Row %	40.000	60.000
	Column %	10.000	60.000
	Total %	8.000	12.000

Comparing $R-judge with judge

Correct	42	84%
Wrong	8	16%
Total	50	

图 10　$t=0.4$ 分类器基于验证样本的分类结果

准确率 A 为检验集中正确分类数和样本数 S 的比值：

$$A=1-R=(S-E)/S \tag{7}$$

各性能评价结果如下：

表 3　基于不同 t 值的性能评价结果

t 值	准确率	误差率
0.2	84%	16%
0.3	86%	14%
0.4	84%	16%

在许多数据挖掘应用中，用度量全面误差率的一个数来描述模型的性能是不合适的。要描述模型的质量，必须有更加复杂和全局性的度量。为了进一步评论我们建立的模型，我们引入模型的敏感性（sensitivity），特异性（specificity），精度（precision），错误正例（false positives），错误负例（false negative）几个度量指标[12]，分别表示为：

敏感性（sensitivity）= t_pos/pos　(8)

特异性（specificity）= t_neg/pos　(9)

精度（precision）= t_pos/（t_pos + f_pos）　(10)

错误正例（false positives）= 1 -（t - neg）/net　(11)

错误负例（false negative）= 1 - t_pos/pos　(12)

其中，t_pos 是真正的样本个数，pos 是正样本个

数,t_neg是真负的样本个数,neg是负样本的个数,f_pos是假正的样本个数。

最终模型的准确率为公式(13):

$$A = \frac{t\_pos}{pos} \times \frac{pos}{(pos+neg)} + \frac{t\_neg}{neg} \times \frac{neg}{(pos+neg)} \quad (13)$$

基于上述模型的混淆矩阵,我们对建立的模型进行评估:

由以上分析可以看出,利用决策树生成的模型在t取不同值时均有良好的分类效果,$t=0.3$时分类效果最佳。

3.6 期刊操控引用行为识别模型实证研究

前文利用CART算法提出了一种期刊引用操控行为的识别模型,在此基础上,笔者对该数据库2010年的人文社科总论的期刊进行识别实证研究。期刊引用操控行为识别模型计算期刊引用集中度时,使用了期刊的引文年代数据和期刊被引及引用数据,因而能综合各方面信息,比较全面地、动态地对期刊引用操控行为做出评价。

3.6.1 数据来源与实验结果

该数据库2010年收录了包含经济学、管理学、社会学等25个大类的527种期刊,年度引用期刊论文563 241篇,中文期刊论文3 331 141篇,其中引用该数据库来源刊论文190 061篇。2010年的人文社科总论的期刊依然为50种,但由于2010年和2009年及2008年不属于一个选刊周期,因此,2010年的期刊样本和2009年的期刊样本有所重叠也有部分变化,2009年有操控引用行为的部分期刊在2010年未被选为来源刊。对2010年的所有期刊引文进行统计,得到实证样本期刊的自引率、引用密度比、被引密度比及年度集中度。

3.6.2 实验结果分析

根据前文,$t=0.3$时模型的分类效果最好,我们使用$t=0.3$的模型对2010年的人文社科总论期刊引文数据进行分类计算,可以得出如下结论:

(1)2010年的522种期刊中,有165个期刊判断为有异常引用行为,置信度为0.916 667。

(2)2010年人文社科期刊所属的25个学科中有19个学科的期刊有异常引用行为。从2010年的数据来看,期刊所属学科越大,该学科期刊越多,有异常引用行为的期刊也越多。期刊所属学科越小,该学科期刊越少,有异常引用行为的期刊也越少或是没有。

(3)从综合类人文社科期刊2010年的数据来看,之前有异常引用行为的期刊在2010年的数据依然保持异常,并且有异常引用行为的期刊数量有所上涨。由此,可以推断,由于没有有效措施遏制期刊的异常引用行为,期刊界片面追求引用数据的异常引用行为还在持续增加,并呈扩散态势。对于期刊界异常引用行为的控制已经刻不容缓。

4 结 语

本文应用决策树数据挖掘方法对国内综合性人文社会科学期刊进行期刊引用是否受操纵进行识别,通过选取一定数量的正常和异常期刊作为样本并建立识别模型,采用CART分类方法完成了对期刊操纵行为的识别,借助SPSS CLEMNENTINE工具完成数据处理、模型构建和分类工作,最后验证了此种方法的可行性。对于本文选用的构建空间向量的特征值以及CART方法,通过对模型的验证和应用分析可以看出,CART算法对样本中的正常期刊和异常期刊识别的正确率均值在85%以上,分类效果比较理想,符合决策树分类要求。为了进一步验证特征值与本模型方法的有效性和可靠性,本文利用该数据库2010年度的数据进行验证,对验证结果的分析显示了此模型方法的有效性,证明了这种分类器判断的灵敏性和可靠性。

表4 基于不同t值的评价结果

t值	敏感性	特异性	精度	错误正例	错误负例	准确率
0.2	60%	90%	70%	50%	40%	84%
0.3	60%	92.5%	74.2%	42.9%	40%	86%
0.4	60%	90%	70%	50%	40%	84%

采用基于 CART 方法的期刊操控引用行为识别模型,通过智能计算,能有效识别出期刊发展的异常行为,节省了从众多期刊中检测可疑期刊的人力资源,降低期刊检测的成本。数据挖掘建模方法的应用可以有效控制期刊操纵行为的产生和蔓延,并在一定程度上起到监督管理期刊正常发展的作用。当然,对于少量一些特殊期刊,其识别结果概率较低的,我们可以收集期刊数据,辅助以人工判断。

引文分析技术日趋完善,其应用不断扩大,已发展成为文献计量学的重要方法之一。当前引文分析所面临的首要问题是各种形式、各种动机的操纵期刊引用的现象日益增多,引文分析赖以生存的数据基础呈现受污染状态。如何从海量的、复杂的期刊引用数据中提取有效的、正常引用的数据,使之服务于核定核心期刊、人才评价、研究学科结构是学术工作者面临的重要问题。而借用科学工具、采用科学思维方式对其处理,正是为了辅助期刊规范化发展。最终目的是使科学发展成果呈现出其真实面貌,期刊发展在正常轨道上发展,而不受人为因素的影响。

参 考 文 献

[1] 郭建顺,张学东,李文红,等. 我国科技期刊的高自引率及其不合理自引的甄别[J]. 中国科技期刊研究,2010,21 (4): 455-458.

[2] 苏成,潘云涛,马峥,等. 权威因子:一个新的期刊评价指标[J]. 编辑学报,2010,22(4):369-373.

[3] John Mingers, Quentin L. Burrell. Modeling citation behavior in Management Science journals [J]. Information Processing & Management, 2006, 42(6): 1451-1464

[4] Liming Liang, Ronald Rousseau. Measuring a journal's input rhythm based on its publication-reference matrix [J]. Journal of Informetrics, 2010, 4(2): 201-209.

[5] 马凤, 武夷山. 关于论文引用动机的问卷调查研究——以中国期刊研究界和情报学界为例[J]. 情报杂志,2009,28(6):9-14.

[6] 刘筱敏,张建勇. 数字资源获取对科学研究的影响——电子期刊全文下载与引用分析[J]. 大学图书馆学报,2009 (1):60-62.

[7] 杨利军,万小渝. 引用习惯对我国期刊论文被引频次的影响分析——以情报学为例[J]. 情报科学,2012,30(7):1093-1096.

[8] 李睿,孟连生. 论专利引用行为与期刊论文引用行为在揭示知识关联方面的差异[J]. 情报学报,2010, 29 (3): 474-478.

[9] 马峥,王娜,周国臻,等. 中国科技核心期刊分类互引网络模式研究[J]. 科学学研究, 2012, 30 (7): 983- 991.

[10] 宋广玲,郝忠孝. 一种基于 CART 的决策树改进算法[J]. 哈尔滨理工大学学报,2009,14(2):17-20.

[11] Breiman L, Friedman J H, Olshen R A, et al. Classification and Regression Trees[M]. Belmont, CA: Wadsworth, 1984.

[12] 秦锋,杨波,程泽凯. 分类器性能评价标准研究[J]. 计算机技术与发展,2006,16(10):85-88.

(责任编辑 化柏林)