• 医学信息研究 •

医学信息学领域算法类别使用及影响力研究*

于 琦 马彩珍 邰杨芳 吴胜男 贺培凤

（山西医科大学管理学院 太原 030001）

〔摘要〕 基于 JCR 数据，采用社会调查法、分层抽样法、全文内容分析法筛选 705 篇使用算法论文样本，通过相关标准及专家咨询法获得 16 个算法类别。在此基础上对算法类别使用情况进行统计，基于分类词典对提及次数、提及位置和共现情况进行影响力分析，为相关研究提供参考。

〔关键词〕 医学信息学；算法；使用行为；影响力评估；全文内容分析

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2020.10.004

Study on the Usage and Influence of Algorithmic Categories in Medical Informatics *YU Qi, MA Caizhen, TAI Yangfang, WU Shengnan, HE Peifeng, School of Management, Shanxi Medical University, Taiyuan 030001, China*

〔**Abstract**〕 Based on JCR data, 705 samples of papers using the algorithm are screened by social survey, stratified sampling and full－text content analysis, and 16 algorithmic categories are obtained by relevant standards and expert consultation. On this basis, the usage of the algorithmic categories is counted, and the influence analysis is conducted on the number of mentions, the location of mentions and the co－occurrence based on the classified dictionary, so as to provide references for relevant study.

〔**Keywords**〕 medical informatics; algorithm; behavior of usage; influence assessment; full－text content analysis

1 引言

20 世纪中期，国外生物医学研究者利用计算机处理医学数据，加速计算机科学与生物医学发展，20 世纪 70 年代，“医学信息学”概念在国际信息处理协会会议上正式提出[1-2]。医学信息学是计算机科学技术、现代医学、图书情报学等多学科交叉的应用型新兴学科，是典型的以数据驱动且高度依赖机器学习和深度学习算法的研究领域[1]。随着网络信息技术、计算机科学飞速发展，数据驱动的医学信息学内涵不断丰富，研究领域逐渐广泛。同时文献计量方法逐步应用于医学领域定性热点趋势分析

〔修回日期〕 2020－06－08

〔作者简介〕 于琦，副教授，博士，发表论文 60 余篇，参编专著和教材各 1 部、译著 1 部；通讯作者：贺培凤，教授，博士生导师。

〔基金项目〕 国家自然科学基金面上项目“基于多元分析的科研文献微观实体评价理论与实证研究——以生物医学为例”（项目编号：71573162）；国家自然科学基金面上项目“多维视角下基于文献实体单元共现网络分析的药物关系知识发现研究”（项目编号：71804102）；国家社科基金青年项目“基于框架网络本体的标签系统语义分析研究”（项目编号：13TCQ030）。

及影响力评估等[3]。算法是医学信息学重要工具[4]。学术界对算法在科研领域的调查较少，现有算法影响力评估主要根据同一任务中不同算法的完成效果进行评价[5-6]。这种基于实验效果的评价方法较直接、准确，但存在一定局限性，即实验需要特定数据集、评估者需要较高专业知识水平。本研究基于内容分析对医学信息学领域算法类别的使用情况及影响力进行分析，定量考察不同算法实际应用情况，为深化医学信息学学科认识提供参考。

2 资料与方法

2.1 资料来源

2018 年美国科学情报研究所出版的网络版《期刊引用报告》（Journal Citation Reports，JCR）共收录 26 种医学信息学期刊。本研究基于已有期刊评价研究[7-8]，综合考虑期刊影响因子和特征因子分值，对上述 26 种期刊进行排序，参考专家意见，选择影响因子 > 2.70、特征因子分值 ≥ 0.006 的期刊，最终选取 5 种期刊进行研究，见表 1。

表 1 医学信息学领域 5 种高影响力期刊

序号	刊名	影响因子	特征因子分值
1	《医学互联网研究杂志》（*Journal of Medical Internet Research*）	4.945	0.030 640
2	《美国医学信息学会期刊》（*Journal of the American Medical Informatics Association*）	4.292	0.019 510
3	《生物医学中的计算机方法和程序》（*Computer Methods and Programs in Biomedicine*）	3.424	0.009 340
4	《生物医学信息学杂志》（*Journal of Biomedical Informatics*）	2.950	0.010 300
5	《国际医学信息学杂志》（*International Journal of Medical Informatics*）	2.731	0.006 740

注：影响因子和特征因子分值为 2018 年数据。

在 Web of Science 数据库中，以 5 种期刊名称进行检索，限定检索年限为 2009 – 2018 年，文章类型选择"Article"，检索结果为 7 948 篇。采用社会调查法中总样本容量公式（1）进行样本规模确定，其中 t 为置信度所对应的临界值，e 为抽样误差。允许抽样误差为 2%、置信度为 95% 计算得到总样本容量为 2 401，采用分层抽样方法确定每种期刊样本量，见表 2。

$$n = \frac{t^2}{4e^2} \quad (1)$$

表 2 2009 – 2018 年 5 种期刊样本量汇总（篇）

刊名	检索结果	样本量
《医学互联网研究杂志》	2 033	615
《美国医学信息学会期刊》	1 578	477
《国际医学信息学杂志》	1 114	337
《生物医学信息学杂志》	1 343	406
《生物医学中的计算机方法和程序》	1 880	566
合计	7 948	2 401

2.2 数据标注与处理

采用全文内容分析法对 2 401 篇论文算法使用情况进行深入分析。首先，根据已有研究中提出的算法句标注类目[9]，建立本研究标注信息，见表 3。其次，依据标注信息类目进行标注。共有 705 篇论文使用算法，共涉及 170 种算法，在此基础上根据《数据挖掘 10 大算法》（*The Top* 10 *Algorithms in Data Mining*）一书标准[10]及专家咨询进行算法名标准化及算法分类，在该书 10 大数据挖掘算法类别基础上新增回归算法、人工神经网络、文本分析、降维、模型、时频分析、检测等 7 种算法类别，最终得到 16 种算法类别（序列模式算法在本研究样本未使用，故不做分析），见表 4。最后基于算法分类词典方法对标注结果进行统计。

表 3 算法句标注信息

ID	论文号
Article – title	篇名
Key – words	关键词
No	算法句编号
Section – type	算法句所在章节类型
Algorithm	算法名称
Content	算法句具体内容

表 4 算法分类词典

序号	算法类别	算法标准名	别名
1	分类算法	CART	Classification and Regression Tree, Classification and Regression
2	回归算法	Logistic Regression	LR, Logistic, Multinomial Logistic Regression, Logistic – regression – binary
3	聚类算法	k – means	k means, SimpleKMeans, KMC, KM
4	统计学习	Support Vector Machine	SVM, support vector machines, svm, svms, SCAD – SVM, LIBSVM, 1C – SVMs, 2C – SVMs, MSVM, SVM + Linear Kernel, SVM + RBF Kernel, OP – SVM, OC – SVM, TC – SVM, SA – SVM, Nu – Support Vector Machine, nuSVC, GA – SVM
5	关联分析	Apriori	—
6	链接挖掘	PageRank	—
7	集成学习	AdaBoost	GBM, GB, GBDT, Adaptive Boosting, Gradient Boosting, Gradient Boosted Feature Selection, Boosted Trees, Gradient Boosting Machine, Boosting, LogitBoost, ADB, TrAdaBoost
8	遗传算法	GA	Genetic Algorithm
9	降维算法	PCA	Principal Components Analysis
10	人工神经网络	ANN	Artificial Neural Network
11	模型算法	CRF	Conditional Random Field Algorithm
12	时频分析算法	STFT	Scale – invariant Feature Transform
13	文本分析算法	TF – IDF	—
14	图形挖掘	LBD	—
15	检测算法	ACF	—
16	粗糙集	LEM2	—

2.3 算法使用评价指标

2.3.1 提及次数 即算法在文章中出现的次数，将提及次数分为 3 个指标。（1）提及论文数。借鉴学术论文影响力评价 Count One 方法[11]，即某种算法类别属下的某种算法无论在一篇文章中出现多少次只记为 1 次，对其提及次数进行累加。例如一篇文章中提及算法类别 A 中的算法 a 和算法 b 则该篇文章算法类别提及次数记为 2。（2）提及总次数。借鉴 Ding 等提出的 Count X 方法[11]，考虑算法反复提及情况对算法类别影响力进行评估，即记录一篇论文中某种算法类别属下的所有算法出现次数。（3）平均提及次数，即算法类别提及总次数与提及论文数比值。

2.3.2 提及位置 即算法类别所在章节类型。学术论文各章节重要性不同[12]，因此不同章节提及算法的重要性不同，导致算法类别在不同章节类型中影响力不同。结合实证型研究论文 IMRDC（Introduction – Material and methods – Results – Discussion – Conclusion）结构[13]将章节划分为 5 种类型，见表 5。因部分算法可能只出现在摘要中故将 Abstract 也作为一种章节类型进行研究。

表 5 章节类型划分

类型	Abstract	Introduction	Method	Results	Conclusion
功能	摘要	引言、相关工作概述	研究方法描述、评价方法、具体的实验过程	实验结果与结果的深入讨论	结论与未来工作说明

2.3.3 共现情况 即一篇论文同时涉及两种或两种以上算法，共现次数越多算法间关系越密切。共现情况次数经计算提及论文数得到。

3 结果分析

3.1 使用趋势

3.1.1 年代变化 705 篇提及算法类别论文数量呈整体上升趋势，其中 2015 年期刊刊载论文数量相对较少致使算法使用论文刊载量较少。分类算法、统计学习、人工神经网络算法使用论文数量较高且逐年递增，尤其在 2015 年后增幅明显，而图形挖掘、检测算法、粗糙集 3 种算法数量较少但呈逐年上升趋势，见图 1。说明医学信息学领域对分类算法等 3 类算法依赖程度较强；其他算法发挥越来越重要作用。

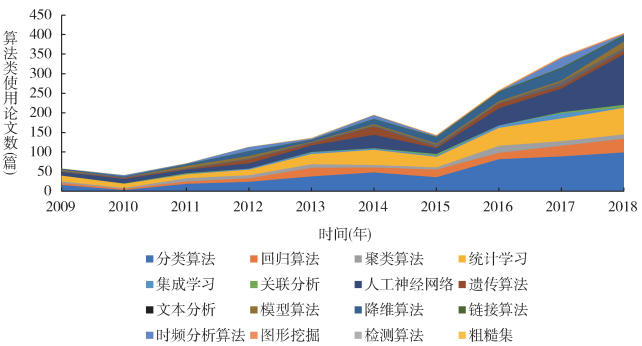

图 1 算法类别使用变化趋势

3.1.2 算法类别使用的期刊变化趋势 16 个算法类别在 5 种期刊的使用各不相同，在《生物医学信息学杂志》和《生物医学中的计算机方法和程序》期刊论文中都有提及且提及论文数较多，在其他 3 种期刊使用较少，其中《医学互联网研究杂志》提及算法类别最少，仅为 7 类。说明生物医学中的计算机方法与程序及生物医学信息计量对算法依赖程度较高。分类算法、统计算法、人工神经网络算法在 5 种期刊中提及论文数较多，图形挖掘、检测算法、粗糙集算法提及论文数较少，见图 2。

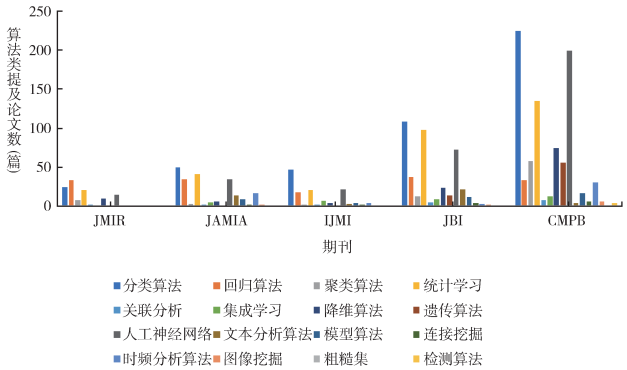

图 2 算法类别期刊使用变化趋势

3.2 使用影响力

3.2.1 提及次数（表 6） 算法类别提及论文数越多，则该算法类别使用越多、影响力越大；当两种算法类别提及论文数相同时，提及总次数高则影响力大；平均提及次数反映算法类别在单篇论文中的使用情况。其中分类算法提及论文数最高，约占 65%，有研究者指出构建分类器系统是数据挖掘最常用工具之一[10]，因此使用率较高。人工神经网络算法排名第 2，第 3 为统计学习算法。随着人工智能发展，人工神经网络在医学信息学领域应用广泛，如在预测与估计、模式识别、生物医学等方面取得较大进展；统计学习是基于概率的算法，能更好地实现预测，从而提高科研效率[14-15]。回归、聚类、降维、遗传算法提及论文数较多，原因在于：回归算法原理简单易实现；聚类算法可从新视角把握数据资源价值；降维算法可去除数据噪声和不重要特征，提高数据处理速度；遗传算法为近年理论和应用研究热点等。粗糙集算法类别仅有 1 篇论文提及，排名最低。算法类别提及论文数与总提及次数排名结果基本一致，而平均提及次数排名发生变化。这表明提及论文数、提及总次数和平均提及次数间不成正比。平均提及次数在 2 ~ 17 间浮动，遗传、模型和文本分析算法分别从提及论文数结果中的第 7、9、10 位升至平均提及次数结果前 3

位，原因在于其作为新兴算法在医学信息学领域使用较少，在使用时需较多篇幅描述解释原理而反复提及。在提及论文数中位列第 1 的分类算法跌至第 9 位，可能由于该算法类别原理较简单而解释较少。其他算法类别的 3 种排序结果差距较小。

表 6 提及论文数、提及总次数、平均提及次数结果

序号	算法类别	提及论文数	提及总次数	平均提及次数
1	分类算法	454(1)	3 459(3)	7.62(9)
2	人工神经网络	341(2)	4 169(1)	12.23(5)
3	统计学习	314(3)	3 915(2)	12.47(4)
4	回归算法	155(4)	1 011(5)	6.52(10)
5	降维算法	117(5)	973(6)	8.32(8)
6	聚类算法	84(6)	854(7)	10.17(6)
7	遗传算法	70(7)	1 127(4)	16.10(1)
8	时频分析算法	54(8)	154(11)	2.85(14)
9	模型算法	49(9)	713(8)	14.55(2)
10	文本分析算法	43(10)	598(9)	13.91(3)
11	集成学习	34(11)	322(10)	9.47(7)
12	关联分析	19(12)	45(14)	2.37(15)
13	链接挖掘	14(13)	73(13)	5.21(11)
14	图形挖掘	12(14)	126(12)	4.85(12)
15	检测算法	4(15)	13(15)	3.25(13)
16	粗糙集	1(16)	2(16)	2.00(16)

3.2.2 提及位置（图 3）

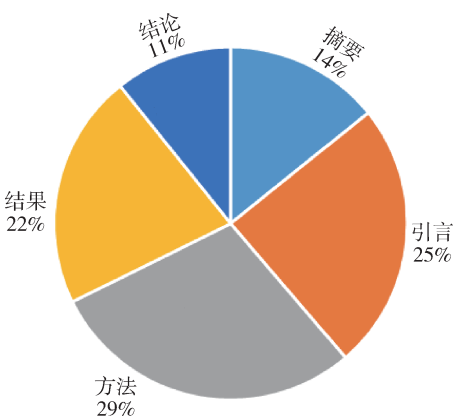

图 3 算法类别各章节分布情况

算法类别在不同章节类型中的提及论文数不同，其中提及论文数最多的是“方法”部分，“结论”部分最少。“摘要”部分是对全文的简要概括，部分期刊要求其包含“目的”、“方法”、“结果”、“结论”4 部分，导致此章节类型中算法提及论文数较低；“引言”部分需对文章所用算法做简单背景介绍，因此会有一定频次的算法提及；“方法”部分是全文描述方法核心部分，算法在该章节类型提及论文数显著增加；“结果”部分对实验所得结果进行分析，不需要对算法相关内容进行详细阐述，因此算法提及论文数相对下降；“结论”部分对全文大致流程和结果做简要总结但不会大量描述，此章节类型算法提及论文数较少。综上，在不同章节类型提及算法其作用不同、影响力不同。本文重点针对“方法”和“结果”章节，分析不同位置各算法类别共现情况，见表 7。算法类别提及次数在“方法”与“结果”部分一般高于其他章节类型，其次是“引言”部分，在“摘要”和“结论”部分提及较少。说明医学信息学研究领域算法主要作为具体实验方法使用。根据“方法”和“结果”章节类型统计结果，排前 3 位的为分类、统计、人工神经网络算法，其提及论文数远高于其他算法；回归、聚类、降维、遗传算法提及论文数较高；检测和粗糙集算法提及较少。与前文研究结果一致。

表 7 各个算法类别在各章节类型中分布情况

算法类别	摘要	引言	方法	结果	结论
分类算法	159	272	376	318	118
回归算法	52	75	121	77	27
聚类算法	33	59	72	54	24
统计学习	131	229	260	205	98
关联分析	5	6	8	1	2
集成学习	20	24	34	20	10
降维算法	32	84	72	60	22
遗传算法	45	61	52	52	39
人工神经网络	123	252	250	189	140
文本分析算法	23	34	57	32	20
模型算法	15	34	33	25	17
链接挖掘	5	8	9	7	3
时频分析算法	9	18	14	8	6
图像挖掘	3	4	10	5	4
粗糙集	0	1	0	0	1
检测算法	1	1	1	1	0

3.2.3 基于共现情况的算法类别影响力分析 利

用 VOSviewer 软件依据提及论文数分析 170 种算法共现情况，见图 4，其中节点越大表示算法被提及次数越多，即重要性越大、影响力越大；连线表示两种算法在同一篇文中被共同提及次数，次数越多连线越粗。在收集到的 705 篇文章中提及两种或两种以上算法的 512 篇，约占 73%。统计学习算法类别中支持向量机（Support Vector Machine，SVM）算法节点最大且与其他 71 种算法均有连线，说明 SVM 算法是医学信息学领域常用算法。研究发现 SVM 算法主要受统计学理论支持，是一种非线性机器学习算法，能够对数据进行高精度处理[14]，是最稳定、最精确的算法之一[10]。分类算法类别中的 Naive Bayes 节点与其他 60 种算法均有连线，其中与人工神经网络算法中的近似最近邻（Approximate Nearest Neighbors，ANN）算法、回归算法中的逻辑回归（Logistic Regression，LR）算法、分类算法中的决策树（Decision Tree，DT）和随机森林（Random Forest，RF）算法共现次数较高，超过 40 次。这可能由于其原理简单，易应用于大量数据集。K－近邻算法（K－Nearest Neighbor，KNN）排名第 3，可能由于其精度高且适用数据范围为数值型和标符型，处理数据较方便。同时分类算法中的 RF、DT、C4.5 算法及人工神经网络算法类别中的 ANN、卷积神经网络（Convolutional Neural Networks，CNN）等算法也具有较高共现次数。这可能由于 RF 算法能有效运行于大数据集，评估各特征在分类问题上的重要性，在预测疾病风险和患者诊断方面应用前景广阔；ANN 是一种类似于生物神经网络的非线性算法，可模拟人脑某些智能行为，为近年研究热点[16]。LR 算法、降维算法类别中的主成分分析（Principal Component Analysis，PCA）算法、遗传算法等节点较大，聚类算法类别各算法节点较小且连线强度较低。说明分类、统计、人工神经网络算法共现使用较为频繁，而聚类算法多为单独使用。

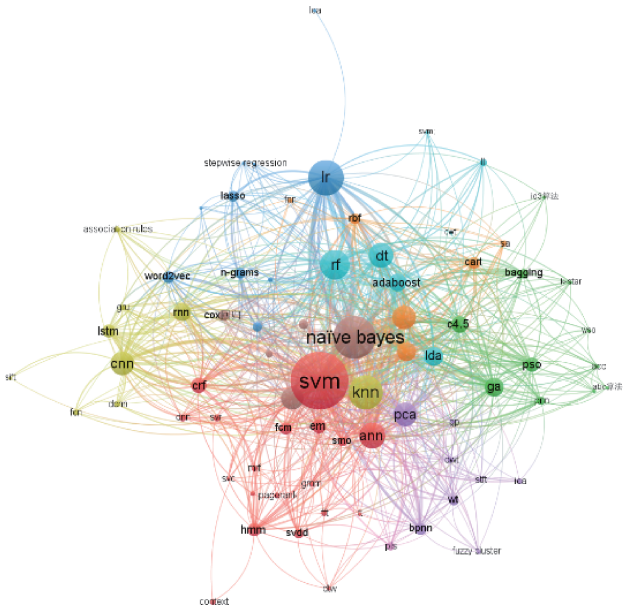

图 4 算法共现情况网络图谱

4 讨论

4.1 算法类别使用总体趋势

除 2015 年外，期刊中使用算法类别的论文比例呈逐年上升趋势；各期刊对算法类别依赖程度不同。医学信息学领域算法使用类文章不足 30%，与软件使用类文章占比接近[16]。说明该领域研究对算法和软件依赖性较低，但呈逐年上升趋势；不同期刊对算法类别依赖程度不同，16 种算法类别在《生物医学中的计算机方法和程序》期刊均有涉及。

4.2 算法类别使用影响力

分类、统计、人工神经网络等算法类别提及次数较多、提及位置较集中、共现次数较多，具有较高影响力。首先，算法类别提及论文数和提及总次数指标对算法类别影响力评估几乎没有差别。可以假定算法类别影响力范围越广、提及论文数越多，相应提及总次数越高。就提及次数来看，提及论文数和提及总次数可反映算法类别影响力广度。而平

均提及次数相较前两项指标对算法类别影响力评估有一定变化，排在前 3 位的算法类别均排名下跌但仍居前列。可以认为算法类别对论文影响力程度越深平均提及次数越高。其次，提及位置影响力反映算法类别在论文不同位置的集中程度，提及位置影响力越高算法类别在“方法”和“结果”部分的占比越大。最后，共现情况影响力越高，算法影响范围越大，集中程度越高。综上，分类、统计、人工神经网络算法影响力广度和深度均高于其他算法；回归、聚类、降维、遗传算法影响力广度和深度次之；检测和粗糙集算法类别影响力广度和深度最低。此外，模型和文本分析算法影响力广度不足，但有较强深度，说明其在少数论文中反复使用，在“方法”和“结果”位置的集中程度、共现情况影响力均排在中后位置。

5 结语

基于内容分析的量化评估可相对全面地统计算法类别在特定领域使用情况，有助于了解算法类别价值并根据科研任务类型选择算法类别及决策算法。未来可考虑获取多种期刊全部论文进行研究；在现有研究基础上可基于年代、使用国家等更多方面进行影响力评估；可区分算法提及和算法使用概念，研究算法类别在文中不同使用身份的影响力差异。

参考文献

1 代涛. 医学信息学的发展与思考 [J]. 医学信息学杂志, 2011, 32 (6): 2 - 16.

2 曹高芳, 于微微, 李继宏, 等. 国内外医学信息教育研究比较 [J]. 预防医学情报杂志, 2013, 29 (1): 62 - 65.

3 宁鹏飞, 聂馥玲. 交叉学科背景下医学信息学发展特征的文献计量分析——基于计算机科学视角 (1980 - 2017 年) [J]. 内蒙古大学学报 (自然科学版), 2019, 50 (1): 50 - 58.

4 王玉琢, 章成志. 考虑全文本内容的算法学术影响力分析研究 [J]. 图书情报工作, 2017, 61 (23): 6 - 14.

5 Wilbanks E G, Facciotti M T, Veenstra G J C. Evaluation of Algorithm Performance in ChIP - Seq Peak Detection [J]. PLoS One, 2010, 5 (7): 1 - 12.

6 Settouti N, Bechar M E A, Chikh M A. Statistical Comparisons of the Top 10 Algorithms in Data Mining for Classication Task [J]. International Journal of Interactive Multimedia and Artificial Intelligence, 2016, 4 (1): 46 - 51.

7 刘宇, 叶继元, 袁曦临. 图书情报学期刊的分层结构: 基于同行评议的实证研究 [J]. 中国图书馆学报, 2011, 37 (2): 105 - 114.

8 苏芳荔, 孙建军. 期刊引用认同指标在期刊评价中的适用性分析 [J]. 中国图书馆学报, 2012, 38 (1): 96 - 104.

9 章成志, 丁睿祎, 王玉琢. 基于学术论文全文内容的算法使用行为及其影响力研究 [J]. 情报学报, 2018, 37 (12): 1175 - 1187.

10 Wu X, Kumar V, Ross Quinlan J, et al. Top 10 Algorithms in Data Mining [J]. Knowledge and Information Systems, 2007, 14 (1): 1 - 37.

11 Ding Y, Liu X Z, Guo C, et al. The Distribution of References Across Texts: some implications for citation analysis [J]. Journal of Informetrics, 2013, 7 (3): 583 - 592.

12 Mccain K W, Turner K. Citation Context Analysis and Aging Patterns of Journal Articles in Molecular Genetics [J]. Scientometrics, 1989, 17 (1 - 2): 127 - 163.

13 Lin L, Evans S. Structural Patterns in Empirical Research Articles: a cross - disciplinary study [J]. English for Specific Purposes, 2012, 31 (3): 150 - 160.

14 葛恭豪. 机器学习算法原理及效率分析 [J]. 电子世界, 2018 (1): 65 - 66.

15 姜娜, 杨海燕, 顾庆传, 等. 机器学习及其算法和发展分析 [J]. 信息与电脑 (理论版), 2019 (1): 83 - 84, 87.

16 杨波, 王雪, 佘曾溧. 生物信息学文献中的科学软件利用行为研究 [J]. 情报学报, 2016, 35 (11): 1140 - 1147.