Computer Science and Application 计算机科学与应用, 2019, 9(7), 1441-1452
Published Online July 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.97162

Hans汉斯

Study on the Correlation between Germplasm Resource Database and Papers Based on K-Means and CHAID Algorithm

Xinyi Cheng[1], Linna Zhou[1], Junpeng Yuan[2,3*]

[1]School of Management Science and Engineering, Central University of Finance and Economics, Beijing
[2]National Science Library, Chinese Academy of Science, Beijing
[3]Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing
Email: *yuanjp@mail.las.ac.cn

Received: Jul. 10th, 2019; accepted: Jul. 22nd, 2019; published: Jul. 29th, 2019

Abstract

Purpose/Significance: By analyzing the existed three major germplasm resources databases in China and the papers published in the field of germplasm resources, studying the databases and papers on the correlation rules, we give suggestions on how to update the database of the information of the new varieties. Method/Process: The study found that there were much inconsistencies between resource databases and literature, such as the degree of content matching and of updating. We use the maximum similarity method to select germplasm resources database and integrate them, and use word-frequency analysis method to clean the literature. Then the system analysis method based on K-Means algorithm and CHAID algorithm is designed to find out the correlation between the new variety and the existed database varieties in the paper. Results/ Conclusion: Taking the potato "Xing Jia No. 3rd" as an example, using the above system analysis method, it is found that Avon and Ostara two varieties are most closely related to the Chinese Crop Germplasm Resource Information network database, as a suggestion to insert the database.

Keywords

Germplasm Resource, Database, Word-Frequency Analysis Method, K-Means, CHAID

基于K-Means和CHAID算法的种质资源数据库与论文关联研究

程心怡[1]，周琳娜[1]，袁军鹏[2,3*]

*通讯作者。

文章引用: 程心怡, 周琳娜, 袁军鹏. 基于 K-Means 和 CHAID 算法的种质资源数据库与论文关联研究[J]. 计算机科学与应用, 2019, 9(7): 1441-1452. DOI: 10.12677/csa.2019.97162

[1]中央财经大学管理科学与工程学院，北京
[2]中国科学院文献情报中心，北京
[3]中国科学院大学经济与管理学院图书情报与档案管理系，北京
Email: *yuanjp@mail.las.ac.cn

收稿日期：2019年7月10日；录用日期：2019年7月22日；发布日期：2019年7月29日

摘　要

目的/意义：分析中国现有的三个主要种质资源数据库和种质资源领域发表的论文，研究得出数据库和论文之间的相互关联规则，给出如何在数据库中更新文献中新品种信息的相关建议。方法/过程：研究发现种质资源数据库和论文之间的同步更新、内容匹配程度和更新程度存在较多的不一致，采用最大相似度法对种质资源数据库进行选择和信息整合，同时运用词频分析法对文献进行清洗，然后设计了基于K-Means算法和CHAID算法的系统分析方法，找出论文中的新品种与数据库品种间的相关联系。结果/结论：以马铃薯“兴佳3号”为例，应用上述系统分析方法，发现在中国作物种质资源信息网数据库中Avon和Ostara两品种与其联系最紧密，作为对其更新到数据库位置的建议。

关键词

种质资源，数据库，词频分析法，K-Means，CHAID

1. 引言

科技文献和科学数据都是科学研究的重要组成部分，也是学术信息交流的基本单元。根据 2011 年欧盟的 PARSE Insight 研究显示，超过 85%的科学家认为将科学数据与论文进行关联是有用的。种质资源(germplasm resource)是生物体内遗传物质的聚集形式。在遗传育种领域，种质资源的定义是一切能够繁殖的具有一定种质或基因的生物类型的总称[1]。作为重要的遗传物质，种质资源在农作物的繁育方面有着不可或缺的地位。种质资源担任着保持优良作物性状、改良作物品种的角色。种质资源的实物研究成果具体表现为农产品，形成种质资源数据库。另外，学者的研究还产生了大量的学术论文。学术论文一般是基础理论研究和应用研究的产物，是科学研究的代表性成果，农产品则是实际应用的成果，但是，学术论文与种质资源库现在基本是分离的，种质资源数据库与学术论文之间的交互链接不顺畅，农产品与学术论文的更新有脱节现象，主要体现为以下三个方面：

1) 该领域的科研成果大多是基于已有的农产品繁育、改良而来，所以，有些学术论文中会存在提及已有的种质资源，但并未没有标注和引用其来源及相关的种质资源库数据；

2) 一般来说，学术论文的成果会领先于实际应用，以种质资源为研究对象的学术论文中，大多数论文的研究成果是创新性的、会提出新的种质资源，这些新种质资源在种质资源库中没有档案记录；

3) 种质资源数据库的范围、存贮格式多样化导致了其与学术论文链接的困难。

因此，十分有必要在种质资源领域，研究论文(学术成果)与种质资源数据库(产品成果)交互链接的关

系，特别是如何将论文提出的新的种质资源可靠的添加到现有资源库，使农作物的繁育、种质资源的更新改良形成良好的闭环。

科学研究中，常通过表达文献核心内容的关键词或主题词的出现频次确定该领域的研究重点和发展动向。由于一篇文献的关键词或主题词是文献核心内容的浓缩和提炼，因此，如果某一关键词在其领域文献中反复出现，则可认为该关键词或主题词所表征的研究主题即为该领域的研究热点[2]。故可将词频分析法应用于种质资源和品种的概念模糊界定中，即通过筛选文献中与二者有关一簇高频词来对一个农作物是否属于品种或种质资源进行判定。K-Means 也称快速聚类，属于覆盖型数值划分聚类算法，一般采用数据间的差异程度作为聚类的测度角度，是一种基于划分的方法，该算法的优点是简单易行，并且适用于处理大规模数据[3]，因此 K-Means 可以应用于本研究中对数据库品种的类别划分。CHAID 模型是决策树中一种基本的分类与回归方法[4]，是最有效率的数据挖掘方法之一，从大量冗杂的数据集中，自动搜索出隐藏在数据集中具有特殊关系的信息，构建成一颗决策树，获得数据中蕴含的知识信息的过程也提取一系列算法规则的过程，可用于本研究中预测新品种将从属于 K-Means 算法已聚成的已有类别中的哪一类，并在此过程中分析新品种与数据库中现有品种之间的相似度(关系度)。因此选择 K-Means 聚类算法和 CHAID 模型展开本研究，并将二者结合构造对农作物新品种分属归类判别的系统分析方法。

本文的主要工作是根据种质资源数据

库和种质资源领域论文的特点，研究文献中的主题词、关键词等对种质资源数据库的判定规则，并尝试利用此规则将新品与原有数据库进行关联。主要工作分为以下三部分：

1) 数据库选取与信息整合。先对各个数据库进行初步介绍，再通过对这些库中的指标对比研究发现异同点，以马铃薯为例列表说明异同。

2) 文献选择与清洗。通过阅读大量文献，发现在目前的学术文献中，缺少很好的阐述种质资源和品种之间区别的文献，相反可以得出品种是包含于种质资源的这一结论；并且就此通过清洗过程找出符合条件的文献进行与数据库建立关联规则。

3) 文献中新品种信息与数据库的关联。在数据库无法及时更新农作物新品种信息的前提下，本文将利用 K-Means 模型和 CHAID 模型间接判断新品种从属于数据库哪一部分。

2. 种质数据库选择与信息整合

2.1. 数据库的选取

种质资源数据库是为了满足育种专家和研究人员对种质资源信息的需求。种质资源数据库，可以实现查询定向培育的有用基因、查找亲本及追踪品种系谱等功能。目前，我国关于种质资源的数据库有中国作物种质信息网、国家农业科学数据共享中心、国家农作物种质资源平台科技资源开放共享目录、中国种业大数据平台、中国农业科学院作物科学研究所、农博数据中心等等。而在本研究中，出于对数据库规模、农业新常态、涵盖范围等三方面的考虑，选取了中国作物种质信息网、中国种业大数据平台以及农博数据中心三个数据库进行研究。

1) 中国作物种质信息网(CGRIS) [5]。主办单位是中国农业科学院作物科学研究所种质信息研究室，CGRIS 是目前世界上最大的植物遗传资源信息系统之一, 拥有粮食、纤维、油料、蔬菜、果树、糖、烟、茶、桑、牧草、绿肥、热作等 340 多种作物、47 万份种质的信息。它包含有 9 个子系统，包括为国家种质库管理和动态监测、青海国家复份库管理等。CGRIS 共存有 700 多个数据库，涵盖 130 万条记录。因此，选取该数据库的原因是其拥有数据多，规模较大。

2) 中国种业大数据平台[6]。该平台由农业部种子管理局、全国农业技术推广服务中心、农业部科技发展中心、中国农业科学院信息研究所共同打造，是在农业发展步入新常态的背景下，为解决作为农业

产业链最上端的种业所面临的机遇与挑战，通过应用大数据理念，不断整合数据资源及建立多部门数据利用协同机制而构建的，并且其与国家农业数据中心和农业信息资源共享服务平台有衔接，是一个会不断更新的数据平台，所以选取这个数据库也是有意义的。

3) 农博数据中心[7]。农博数据中心是农博网成功打造的三大特色平台之一，涵盖企业、县市、品种与产品、实用技术、法规标准、人物等六大类别，有丰富的数据、专业的信息和有效的匹配，因此这种被称作行业内人士的百科全书式数据库[8]的选取也是有必要的。

2.2. 数据库信息对比及整合

以马铃薯这一品种为例，CGRIS 库中有 30 条项目指标描述，中国种业大数据平台有 10 条项目指标，农博数据中心则有 8 条项目指标。在对上述三个具有代表性的数据库进行对比研究后，发现不同的数据库的描述指标有着区别和联系。具体条目不同如下见表 1：

Table 1. Database index comparison results
表 1. 数据库指标对比结果

指标类型	类型简介	种质资源数据库名称	项目指标名称
编号类型	编号一般是为了更好地区别、识别种质资源而存在	CGRIS	库编号
			统一编号
			省编号
			原编号
		中国种业大数据平台	登记编号
		农博数据中心	国家省定编号
名称类型	每个农作物品种都有其唯一的名称	CGRIS	品种名称
			译名
			科名
			属名
			学名
		中国种业大数据平台	品种名称
		农博数据中心	品种名称
			分类名称
来源及保存类型	来源及保存类型中的项指标中包含了种质资源的来源地的地理位置、申请者、登记年份等信息	CGRIS	原产地
			高程
			东经
			北纬
			来源地
			引入时间
			保存单位
		中国种业大数据平台	登记年份
			申请者
		农博数据中心	品种来源
			选育单位
权力类型	权力类型主要是指品种权等	CGRIS	无
		中国种业大数据平台	品种权
		农博数据中心	无

本节通过对中国目前现有的三个主要种质资源数据库中种质资源信息描述条目进行梳理，基于最大相似程度且基于描述描述条目的不同类型对其中相异的部分进行了归纳整理，共分为四个部分：关于编号类型的描述条目，关于名称类型的描述条目，关于来源及保存的描述条目以及关于权力类型的描述条目。每个数据库所侧重的指标类型均不同，例如来源及保存类型中，CGRIS 主要侧重种质资源来源的地理位置，中国种业大数据平台和农博数据中心则更侧重于育种者及育种单位的介绍，描述相同指标有关于性状表现的指标，如幼苗色、株高、芝粗、穗形等等，此处不再赘述。

3. 基于主题词词频分析法的文献选择与清洗

中国知网[9]凭借优质的内容资源、领先的技术和专业的服务，在业界享有极高的声誉为，因此本研究所选取的文献均来源于中国知网。本研究的重点是我国的种质资源，故涉及到的相关学术论文均是中文论文。

3.1. 学术论文主题选取过程

为确定本文高频词阈值选取[10]的方法，本文在中国知网中检索“研究、热点”相关的文献。将摘要=“热点”且主题=“词频”，发表时间为 2018 年和 2017 年作为检索条件，共得到 225 条记录，通过人工筛选去除去部分不合要求的文献，最终得到 170 篇文献。通过提取 170 篇文献中高频词阈值的方法，并以此为代表，整理目前我国学界常用的高频词阈值选取方法，结果如下。

Table 2. Results of threshold selection methods for commonly used high-frequency words
表 2. 常用高频词阈值选取方法结果

方法	数量/篇数	占比
频次选取法	82	48.24%
前 N 位选取法	42	24.71%
高低频词界定公式选取法	14	8.24%
普赖斯公式选取法	6	3.53%
频次选取法+前 N 位选取法	6	3.53%
中心度选取法	5	2.94%
高低频词界定公式选取法+前 N 位选取法	5	2.94%
普赖斯公式选取法+前 N 位选取法	4	2.35%
普赖斯公式选取法+频次选取法	4	2.35%
高低频词界定公式选取法+普赖斯公式选取法	2	1.17%

由表 2 可以看出频次选取法[11]最为常用，其普遍性最高，因此将此作为本文的高频词阈值选取方法。

为确定与种质资源和品种概念相关的关键词，本文在中国知网中检索“种质资源、品种”相关的文献。将主题=“种质资源”或者主题=“品种”，并且主题=“农作物”作为检索条件，按照主题排序，共得到 7285 条记录，通过主题贴近的标准选取前 2000 条文献，然后通过人工筛选去除去部分不合要求的文献，最终得到 1955 篇文献。应用词频分析法[12]，通过 python 中的结巴词包对这 1955 篇文献的关键词和摘要进行词频统计。最终发现与农作物种质资源和品种相关的文献研究主题可以分为三类(不同主题热点词见表 3)：

1) 对作物种质资源中遗传物质的相关研究(生物基因方面)；

2) 对农作物品种的某性状的相关研究；

3) 对农作物品种的介绍。

Table 3. Different topic hot words
表 3. 不同主题热点词

研究主题	热点词
对作物种质资源中遗传物质的相关研究	SSR ISSR 标记 基因 保护 遗传 多样性 图谱 指纹
对农作物品种的某性状的相关研究	抗性 品质 形状 抗寒性 指标
对农作物品种的介绍	筛选 育成 育种 选育 变异 杂交 新品种

经筛选可知第三类研究主题才是本文后续研究所需要的文章，结合前面的高频词，本文在中国知网中，以篇名=“筛选”或篇名=“育成”，或篇名=“育种”，或篇名=“选育”，或篇名=“变异”，或篇名=“杂交”，或者篇名=“新品种”，并且主题=“农作物”为检索条件，共得到 2267 篇文献。进而利用词频分析法对这 2267 篇文献的摘要进行高频词统计，其中与数据库指标相关且出现频率最高的前十个词如下表 4：

Table 4. The top ten most frequent words
表 4. 频率最高的前十个词

词语	出现频数
审定	2082
育种	1420
产量	345
母本	227
父本	219
含量	212
形状	170
编号	162
生育期	119
抗性	87

3.2. 相关文献主题分析总结

由 3.1 小节中的文献主题分析过程可知，种质资源研究的学术论文主要分为三种类型：对作物种质资源中遗传物质的相关研究；对农作物品种的某性状的相关研究；对农作物品种的介绍。其中主题是农作物品种的介绍的学术论文是本研究所需的论文，继续应用词频分析后得出结论：文献中出现频率最高、与数据库指标关联性最强的十个词中，关于品种的介绍性词汇与数据库中重合度很小，除却几个确定性的指标，如名称和编号等，其余描述种质资源信息的指标在摘要中均无法与库中对应上，因此有必要建立文献与数据库的关联规则，以便新品种出现时将其及时补录进种质资源数据库中。

4. 文献品种信息与数据库信息的关联研究

种质资源数据库与学术论文二者的关联是双向的，但是由于在种质资源的研究中农产品的产出才是

目的，所以研究人员一般更关注文章中的信息在数据库中是否更新。现实生活中，从农作物新品种的创造到最终在数据库中添加相关品种条目，往往需要一个漫长的审定过程，而期间往往是相关介绍论文的出现早于数据库的更新。故当一篇有关品种介绍的新文献出现时，通过查询审定编号，人们可以判断数据库中是否存在相关的品种，而目前而言，人们往往可能会在数据库中找不到相关的信息，所以有必要提供一种规则来链接文献和数据库。本文接下来将提供对一个新品种出现时，其应该如何更新入数据库的建议。

4.1. 系统分析算法设计

本节将 K-Means 聚类算法与 CHAID 分类算法结合在一起，形成一个系统分析算法，将用于对给出新品种信息插入数据库的建议，具体算法设计见下图 1。

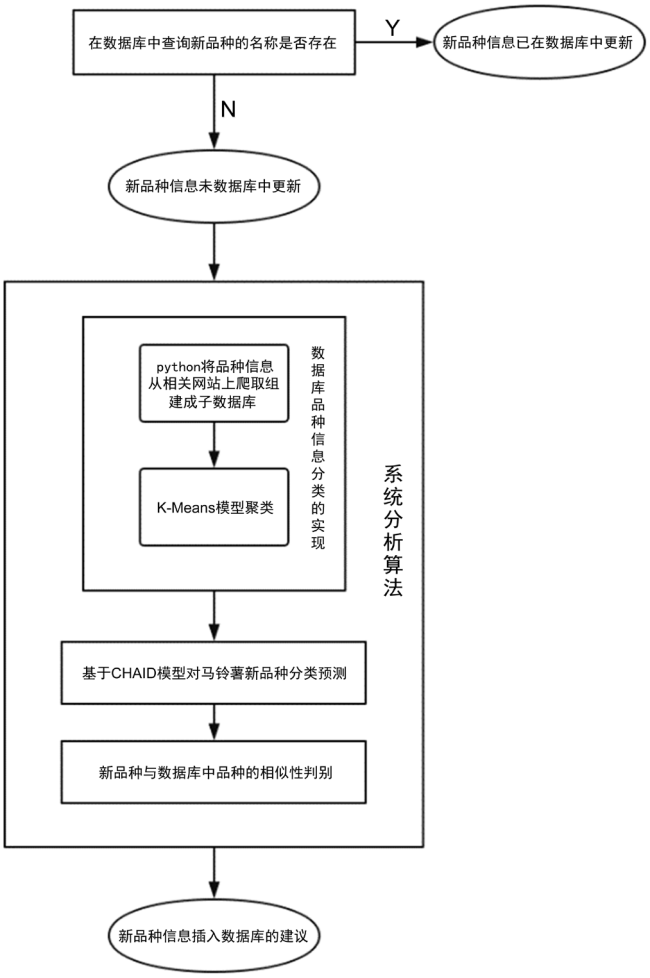

Figure 1. Algorithm flow diagram
图 1. 算法流程展示图

首先判断新品种信息是否已在数据库中进行封信，若无，则进行下一步。该系统分析算法首先运用 K-Means 聚类算法对种质资源数据库信息进行聚类，然后通过 CHAID 分类算法对文献中新品种信息与数据库中原有种质资源信息进行分类预测，并以此论述种质资源数据库和文献的关联规则，最后给出关于文献中新品种如何更新反映在数据库中的建议。

4.2. 更新数据库——以马铃薯一类为例

马铃薯一直是我国人民的主要食物，其种植面积大，研究范围广，研究时间长，故本文以中国作物种质资源信息网中农作物马铃薯子数据库中所有的品种信息条目作为测试样本集，进行研究，

4.2.1. 基于 K-Means 模型的马铃薯数据库品种信息分类的实现

首先利用 python 爬去中国作物种质资源信息网中马铃薯品种的全部数据条目，共得到 714 条信息记录。其次将这 714 条数据作为样本集，利用 clementine 软件中的 K-Means 模型对其进行聚类，最终聚成四类，具体流程图见图 2。

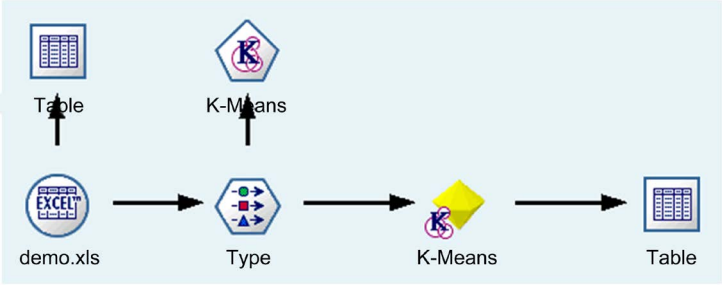

Figure 2. Data flow presentation diagram
图 2. 数据流展示图

在 714 条数据记录中，每条数据均有 30 个字段属性，其中由于统一编号、品种名称、译名和病毒抗感性无法作为判断品种的依据，故不予以输入，最终以其他的 26 个字段作为输入变量用以进行 K-Means 聚类。最终将 714 数据共聚为 4 类，结果如下图 3。

Figure 3. Clustering results
图 3. 聚类结果

在所有 26 个输入指标中，K-Means 模型最终选择了上述 19 个变量作为聚类的依据。其中 proximities 表格中表示的是 cluster-1 与其他三类的距离。其中$KMD-K-Means 表示的是各样本与本类中心的距离。

4.2.2. 基于 CHAID 模型对马铃薯新品种分类预测

本文从中国知网选取了 15 篇 2018 年最新发布的关于马铃薯新品种的文献，确保其在数据库没有记录后，将文献中的信息按数据库的指标形式整理好后，输入 clementine 软件中作为测试集，以第一步中分好类的马铃薯数据库条目作为分类样本集，利用 CHAID 模型对其进行分类预测。

在测试集中共输入十五条数据，每条数据共有 17 个输入字段，分别是：大小，皮色，肉色，芽眼深浅，皮光滑度，整齐度，丰产性，淀粉含量，病毒抗感性，晚疫病抗性，pvx，pvy，蒸食品质，耐储性，送留单位，其它，省；以及类别作为预测的输出字段，用于判断新品种应从属于第一步中四类中的哪一类。

从 clementine 软件中可以看出，由卡方检验和 F 检验可以验证出在 17 个输入变量中，最重要的 6 个变量是 pvx，pvy，大小，形状，丰产性和皮光滑度，其重要程度见表 5。并且最终以此 6 个变量作为分类的节点。

Table 5. Classification node selection
表 5. 分类节点选择

变量	重要程度
pvx	0.397
pvy	0.329
大小	0.159
形状	0.084
丰产性	0.026
皮光滑度	0.005

	整齐度	丰产性	淀粉...	病毒抗感性	晚疫病抗性	pvx	pvy	蒸食品质	耐储性	送留单位	其它	省	类别	距离	$R-类别	$RC-类别
1	$null$	高	12.4...	中抗轻花叶和重花叶	中抗	$...	$...	$null$	$null$	$null$	$n...	黑龙江	$n...	$n...	cluster-1	0.994
2	$null$	高	15.1...	抗Y病毒，中抗X病毒	$null$	$...	$...	$null$	$null$	$null$	$n...	内蒙古	$n...	$n...	cluster-1	0.583
3	$null$	$null$	19.4...	中抗X、Y病毒	$null$	$...	$...	$null$	$null$	$null$	$n...	山西	$n...	$n...	cluster-1	0.994
4	$null$	$null$	$null$	抗X病毒，对Y病毒耐病性，不抗晚疫病	$null$	$...	$...	$null$	$null$	$null$	$n...	山西	$n...	$n...	cluster-1	0.583
5	$null$	$null$	12.4...	中感晚疫病，中抗轻花叶病毒病和重花叶病毒病	$null$	$...	$...	$null$	$null$	$null$	$n...	$null$	$n...	$n...	cluster-1	0.994
6	$null$	$null$	15.8...	抗晚疫病，中抗轻花叶病毒病和重花叶病毒病	$null$	$...	$...	$null$	$null$	$null$	$n...	云南	$n...	$n...	cluster-1	0.583
7	$null$	$null$	15.3...	$null$	较抗病	$...	$...	$null$	$null$	$null$	$n...	$null$	$n...	$n...	cluster-1	0.583
8	$null$	$null$	15.5...	$null$	较强	$...	$...	$null$	$null$	$null$	$n...	湖北	$n...	$n...	cluster-1	0.583
9	$null$	$null$	15.9...	$null$	中抗	$...	$...	$null$	$null$	$null$	$n...	湖北	$n...	$n...	cluster-1	0.583
10	$null$	$null$	11.0...	$null$	抗	$...	$...	$null$	$null$	$null$	$n...	河南	$n...	$n...	cluster-1	0.994
11	整齐	中	12.9...	$null$	感	无	无	上	差	$null$	$n...	国外	$n...	$n...	cluster-2	0.921
12	整齐	中	16.0...	$null$	感	无	无	上	优	$null$	$n...	国外	$n...	$n...	cluster-2	0.921
13	整齐	中	14.4...	$null$	感	无	有	下	差	$null$	$n...	国外	$n...	$n...	cluster-3	0.867
14	不整齐	低	12.4...	$null$	感	有	有	上	差	$null$	$n...	国外	$n...	$n...	cluster-3	0.898
15	整齐	低	13.7...	$null$	感	有	无	上	差	$null$	$n...	国外	$n...	$n...	cluster-3	0.552

Figure 4. Predicted results
图 4. 预测结果

图 4 说明，在十五个马铃薯新品种中有 10 个被预测为属于第一类马铃薯，2 个预测为属于第二类，3 个被预测为第三类。其中$RC-类别一列表示该样本从属该类的概率。以上为对一个新品种应该如何插入数据库的建议，其中无法具体判断新品种应该所属的位置，只能模糊判断它从属于哪一类别，和哪一群马铃薯品种的特性更为接近，从而间接为其如何插入数据库提供初步建议。

4.2.3. 新品种与数据库中品种的相似性判别

在前一部分中，利用 CHAID 模型可以判断出新品种从属于数据库中的哪一类别，接下来以兴佳 3 号为例(具体指标见表 6)，以变量重要性为序对 cluster-1 中的品种信息进行排序，从而找出数据库中与兴佳 3 号最为相似的品种。

Table 6. Xingjia No. 3's index
表 6. 兴佳 3 号指标情况

品种名称	pvx	pvy	大小	形状	丰产性	皮光滑度
兴佳 3 号	-	-	-	椭圆	高	-

按照上述条件对 cluster-1 中的数据进行筛选，最终得到 46 条马铃薯品种信息，其上述 6 个指标描述与兴佳 3 号完全一致，在综合考虑淀粉含量等其他因素，最终找出两个马铃薯品种和兴佳 3 号联系最为紧密，分别是 Avon 和 Ostara。至此，以上为对兴佳 3 号如何插入数据库的全部建议。

4.3. 模型效果分析

分类模型中，C5.0 模型[13]、Discriminant 模型[14]、CHAID 模型等一直为研究人员大量应用于各种研究中。本文以中国作物种质信息网马铃薯子数据库的 714 条数据记录作为训练集上，分别测试了 C5.0、Discriminant、CHAID 三个模型，三个模型结果对比如下图 5、图 6、图 7：

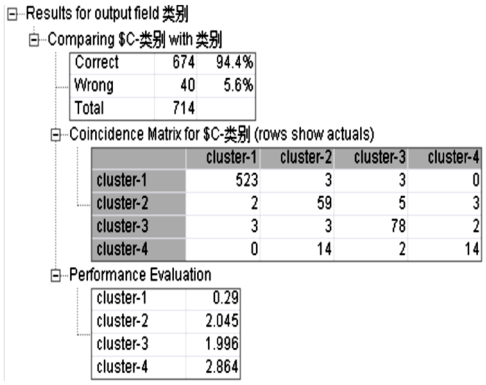

Figure 5. C5.0 model result diagram
图 5. C5.0 模型结果图

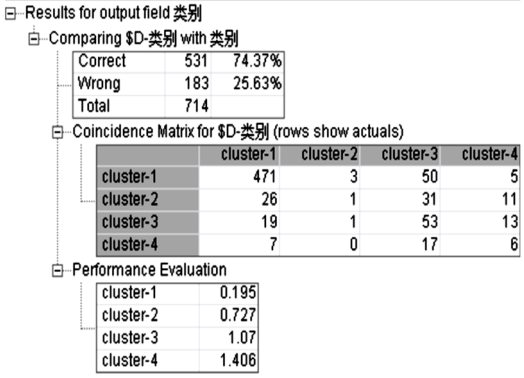

Figure 6. Discriminant model result diagram
图 6. Discriminant 模型结果图

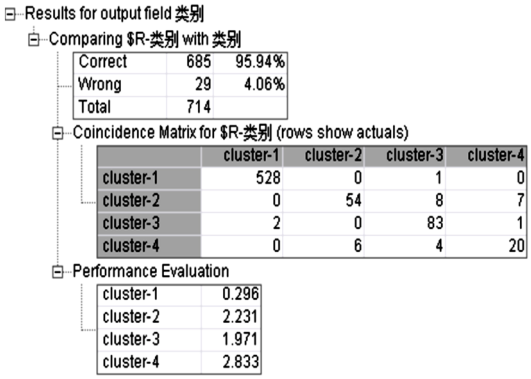

Figure 7. CHAID model result diagram
图 7. CHAID 模型结果图

从三个图结果我们可以直观的看到 C5.0 预测准确率为 94.4%，Discriminant 预测准确率为 74.37%，CHAID 预测准确率为 95.94%，因此，从准确率而言，我们应该选择 CHAID 模型；从模型运行速度角度考虑，三个模型都能较快运行，差别不显著。

而将 C5.0 模型和 CHAID 模型应用到测试集上，发现 C5.0 模型无法在给出一个未知类别的新样本情况下运行，而 CHAID 能成功的预测新样本的类别。

因此综合考虑模型准确率、可行性及运行速度，CHAID 是作为预测论文中出现的品种与原本数据库中哪类样本最为接近的模型。

5. 结论

本文主要通过对种质数据库信息的整合，运用词频分析法对相关文献进行清洗筛选后，运用 K-Means 算法和 CHAID 算法对文献新品种和种质资源数据库给出其关联规则的建议，且系统分析算法模型在准确率、可行性和运行速度方面均令人满意。

但其中的不足主要体现为本研究中输入变量无法量化，其中所有的输入变量不均是数值型变量，故无法对其量化处理，故可能存在偶然性误差。由于本例中输入指标中任何一个都不足以单独作为判断品种类型的指标，故取得输入变量个数不足，导致分类预测仍会存在错误。故今后可以尝试对 K-Means 模型的算法进行改进，或者利用其他模型例如 K 中心店模型对聚类结果进行改进；或尝试增加输入指标的个数，因为取得变量个数不足导致判断新品种从属于哪一类还是会有错误，如果将能增加输入变量个数应该能进一步提高准确率，从而进一步完善本文的研究工作。

基金项目

本论文系 2018 年“中国科学院文献情报领域引进优秀人才计划”择优支持项目、中国科学院文献情报中心重点任务专项：科学文献与科学数据的开放链接服务研究成果。

参考文献

[1] 潘家驹. 作物育种学总论[M]. 北京: 中国农业出版社, 1994.

[2] 马费成, 张勤. 国内外知识管理研究热点——基于词频的统计分析[J]. 情报学报, 2006, 25(2): 163-171.

[3] 金微, 陈慧萍. 基于分层聚类的 k-means 算法[J]. 河海大学常州分校学报, 2007, 21(1): 7-10.

[4] 张超, 陈平雁, 张小远. CHAID 方法及其在高校教师职业倦怠感影响因素分析中的应用[J]. 第一军医大学学报, 2003, 23(12): 1352-1354.

[5] 中国作物种质信息网(CGRIS). http://www.cgris.net/query/croplist.php

[6] 中国种业大数据平台. http://202.127.42.47:6010/index.aspx

[7] 农博数据中心. http://shuju.aweb.com.cn/breed/breed-14-1.shtml

[8] 余红. 玉溪生态农产品农博会上受欢迎[J]. 云南农业, 2018, 358(11): 27.

[9] 中国知网. http://www.cnki.net

[10] 刘小敏, 王昊, 李心蕾, 等. 不同特征粒度在微博短文本分类中作用的比较研究[J]. 情报科学, 2018, 36(12): 128-135.

[11] 唐晓波. 基于本体和 Word2Vec 的文本知识片段语义标引[J]. 情报科学, 2019, 37(4): 97-102.

[12] 赵蓉英, 余波. 中外近五年智库研究演进脉络与热点主题对比分析[J]. 情报科学, 2018, 36(11): 5-11+18.

[13] 庞素琳, 巩吉璋. C5.0 分类算法及在银行个人信用评级中的应用[J]. 系统工程理论与实践, 2009, 29(12): 94-104.

[14] Chen, B.-W. (2018) Incomplete Data Classification—Fisher Discriminant Ratios versus Welch Discriminant Ratios. *Future Generation Computer Systems*.