孙姝艺,刘　洛,胡月明. 基于空间数据挖掘的广东省“旱改水”优先区选择[J]. 江苏农业科学,2019,47(4):216 – 222.
doi:10.15889/j.issn.1002 – 1302.2019.04.050

# 基于空间数据挖掘的广东省“旱改水”优先区选择

孙姝艺[1], 刘　洛[2], 胡月明[1]
(1. 国土资源部建设用地再开发重点实验室,广东广州 510640; 2. 华南农业大学资源环境学院,广东广州 510640)

**摘要**:随着新型城镇化水平的不断提高,大量耕地将会被占用。为有效遏制水田面积减少趋势,提升耕地质量,保障粮食安全,各地政府开始实施“旱改水”工程改造,即将同等数量的旱地改造为水田的方式间接开发水田。旱改水优先区域选择是实施改造先后顺序的重要评判标准。借助 WEKA 软件,基于空间数据挖掘的方法来选择“旱改水”优先区:(1)通过属性选择获取相关性较高的 9 个改造因子,实现数据的预处理;(2)通过 K – means 聚类分析将研究对象划分成 5 类簇;(3)通过 Apriori 关联规则挖掘出分区因子属性之间最强关联关系作为决策挖掘出优先改造的簇,并对结果进行分析。应用于我国广东省,试验证明,空间数据挖掘有效地从庞大数据量中提取信息,耦合空间关系,把数据转化为有用的知识,使用空间数据挖掘的方法选择优先区是可行和科学的。

**关键词**:旱改水;优先区;空间数据挖掘;广东省
**中图分类号**:S126　　**文献标志码**:A　　**文章编号**:1002 – 1302(2019)04 – 0216 – 07

耕地不仅是人类劳动的对象,也是重要的生产资料,是人类生存与发展的基础,随着工业化、城市化的不断发展,人地矛盾的日益尖锐,粮食需求压力的逐步增加,迫使人们更加关注我国有限的耕地资源[1]。在严格的土地管理制度背景下,为保证国家粮食安全而制定的耕地保护政策——占补平衡制度将在未来会不断地强化和完善。为贯彻党中央、国务院关于最严格耕地保护制度的总体要求,明确关于建设占用耕地要“占优补优、占水田补水田”的规定,有效遏制水田面积减少趋势,提升耕地质量进一步挖掘耕地后备资源,解决新增建设用地占用水田补充平衡问题,各地政府开始实施“旱改水”工程改造。“旱改水”是保护耕地提升耕地质量的重要手段,有利于农业增产、农民增收,能够有效提高土地的产出效益,形成稳定的生产能力,做到藏粮于田[2-3]。

受地形地貌水源条件等因素的限制,补充耕地中水田比例不高,实现占优补优难度很大,因此选择区位条件较好、有灌溉水源、土壤肥力较高的部分旱地改造成水田。但又受资金、社会环境、自然环境、生态等因素影响,各级政府在一定时期内投资改造水田的规模是有限的,因此改造水田的选择在空间和时间上存在着先后顺序,而“旱改水”优先区选择是改造区域先后顺序的重要评判标准。

现阶段国内关于“旱改水”分区的研究较少,与此相关的研究主要集中在通过传统的“旱改水”适宜性评价来进行潜力分区[4-6]。传统的“旱改水”适宜性评价存在专家打分不确定性强、强调分值为分区依据而非空间关系、不能挖掘其区域分布规律及因子内在联系等缺点,因此缺乏一种科学的、定量的方法。广东省“旱改水”分区包括大量的空间数据分析,为了有效耦合空间位置和属性相关性,笔者提出一种结合空间聚类和关联规则的数据挖掘方法,选择优先区,并进一步发现优先区域分布规律和分区因子之间的关联关系。

空间聚类作为聚类分析的一个研究方向,是指根据空间异质性将空间对象分成由相似对象组成的类[7]。目前,已有许多研究提出了针对不同数据类型的基于多种空间聚类算法的土地分区,由于空间聚类方法能根据空间对象的属性对空间对象进行分类划分,因而空间聚类方法也是土地分区的一种重要方法。迄今为止,人们已提出了大量的空间聚类算法,本研究在试验分析的基础上,选择适用于数值型大数据集的 K – means 算法对试验数据进行聚类,典型的 K – means 算法在空间聚类各算法中一直处于核心地位,该算法以平方误差准则较好地实现了空间聚类,对于大数据集的处理效率较高[6]。

关联规则挖掘是数据挖掘研究领域中的一个重要任务,旨在挖掘事务数据库中有意义的关联,找出隐藏在数据间的相互关系。随着大量数据被不停地收集和存储,从数据库中挖掘关联规则显得越来越有必要性,本研究采用最经典的 Apriori 算法,求出最强关联的分区因子,进而推出他们的关系,然后将这些规则转换来选择优先区,为决策提供重要的依据。

## 1　材料与方法

### 1.1　研究区与数据来源

1.1.1　研究区概况　广东省全境位于 20°09′ ~ 25°31′N 和 109°45′ ~ 117°20′E 之间。地处我国大陆最南部,东邻福建省,北接江西省、湖南省,西连广西省,南临南海,珠江口东西 2 侧分别与香港、澳门特别行政区接壤,西南部雷州半岛隔琼州海峡与海南省相望。全省拥有丰富的耕地资源,目前全省已储备位列全国第 2 的可用于占补平衡的耕地指标 11.3 万 $hm^2$。然而,由于快速的城镇化和受可开发资源所限,全省储

收稿日期:2017 – 10 – 24
基金项目:广东省省级科技计划(编号:2013A040600002)。
作者简介:孙姝艺(1993—),女,山东日照人,硕士,主要从事土地信息工程研究。E – mail:292467636@qq.comton。
通信作者:胡月明,博士,教授,主要从事土地利用研究。E – mail:ymhu163@163.com。

备的耕地指标中水田、水浇地较少，难以满足“占优补优、占水田补水田”的新要求，因此“旱改水”优先区选择对于提高改造效率是迫切需要的。

1.1.2　数据来源　本研究的对象为广东省旱地、可调整地类、未利用地以及全省现有补充耕地项目（历年园地山坡地开发补充耕地项目，项目红线在空间上已和上述 3 种地类进行叠加分析处理，空间位置不重叠），以包含 13 个地级市 123 个县的 1 683 403 块图斑作为分区单元。基础数据来源于广东省 2013 年土地利用变更调查数据地类图斑；广东省 1∶50 万 DEM 数字高程模型，2013 年广东省各县（市、区）耕地质量年度更新成果用于图层属性的获取；道路、居民点、水系等基础地理数据，第 2 次土壤普查数据、2010—2015 年各年度水资源年报数据等用于相关指标数据获取。

1.2　研究方法

1.2.1　属性选择　数据预处理是数据挖掘的重要一环，要使挖掘内核更有效地挖掘出知识，就必须为它提供干净、准确、简洁的数据。属性选择通常作为数据挖掘的一个预处理步骤，在数据选择和为数据挖掘做准备的过程中起着重要的作用[8]。这个过程就是通过搜索数据中所有可能的属性组合，删除不相关和（或）冗余的属性选出 1 个有 $m(m<N)$ 个属性的子集，也就是预测效果最好的属性子集。

一般来说，属性选择算法由 4 个基本步骤组成：子集产生、子集评估、停止准则和结果有效性验证[9]。属性选择基本步骤见图 1。

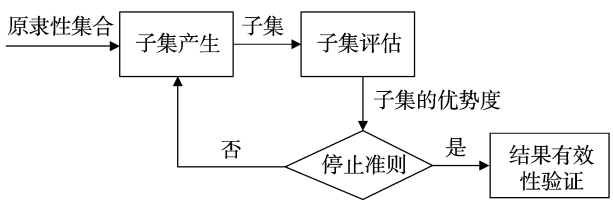


**图1　属性选择基本步骤**

子集产生是一个搜索过程，它产生用于评估的属性子集。子集产生过程所生成的每个子集都需要用事先确定的评估准则进行评估，并且与先前符合准则最好的子集进行比较，如果它更好一些，那么就用它替换前一个最优的子集。属性选择过程可以在满足以下的条件之一时停止：（1）一个预先定义所要选择的属性数；（2）预先定义的迭代次数；（3）是否增加（或删除）任何属性都不产生更好的子集；（4）已经根据确定的评估标准获得最优的子集。选择的最优子集需要通过在所选子集和原属性集间进行不同的测试和比较，使用人工和现实世界的数据集所产生的结果进行有效性验证。

手工选择属性既烦琐又容易出错，自动的属性选择方法通常更快更好。为了实现属性自动化，本研究借助 WEKA 软件设立的 CfSubsetEval 属性评估器和 BestFirst 搜索方法进行属性选择。（1）CfSubsetEval 评估器评估每个属性的预测能力及其相互之间的冗余度，倾向于选择与类别属性相关度高，但相互之间相关度低的属性。选项迭代添加与类别属性相关度最高的属性，只要子集中不包含与当前属性相关度更高的属性。评估器将缺失值视为单独值，也可以将缺失值记为与出现频率成正比的其他值。（2）BestFirst 搜索方法执行带回溯的贪婪爬山法，用户可以指定在系统回溯之前，必须连续遇到多少个无法改善的节点。它可以从空属性集开始向前搜索，也可以从全集开始向后搜索，还可以从中间点（通过属性索引列表指定）开始双向搜索并考虑所有可能的单个属性的增删操作。

1.2.2　数据标准化　为了统一量化各个因子之间的关系和对总目标的贡献，首先需要对分区指标进行一致性处理与指标无量纲化，将数据表进行极差标准化和文本标准化处理。

1.2.2.1　极差标准化　由于有效土层厚度、有机质含量、障碍层深度要求“越大越好”，采用上限效果测度，即：

$$A_{ij}=\frac{X_{ij}-\min(X_{ij})}{\max(X_{ij})-\min(X_{ij})}(X_{ij}\text{为正指标})；$$

地形坡度要求“越小越好”，采用下限效果测度，即：

$$A_{ij}=\frac{\max X_{ij}-X_{ij}}{\max(X_{ij})-\min(X_{ij})}(X_{ij}\text{为负指标})。$$

式中：$A_{ij}$ 为标准化后的指标；$X_{ij}$ 为原数据，代表第 $t$ 个单元第 $j$ 个指标。标准化之后，各要素的最大值为 1，最小值为 0，其余数值都在 0 和 1 之间，这样就对所有单元的数据进行了标准化。

对于中向指标 pH 值，以 6.0～7.9 左右 2 边分别进行正向标准化和负向标准化。

1.2.2.2　文本标准化　以表层土壤质地、剖面构型、灌溉保证率、排水条件作为文本数据，需要根据因子等级进行赋分，然后转换为数值型数据再进行正向的极差标准化处理。

1.2.3　K－means 空间聚类　空间聚类（spatial clustering）是要在一个较大的多维数据集中采用距离度量以标志出聚类，使得同一聚类中的对象有较高的相似度，而不同聚类中的对象彼此不同，是空间数据挖掘的一个重要组成部分[10-11]。

K－means 算法是很典型的基于距离的聚类算法，采用距离作为相似性的评价指标，根据给定的 $k$ 值随机产生 $k$ 个分组中心，将所有实例分为围绕这些中心的 $k$ 个分组，然后通过反复迭代不断改进分组，直至分组效果最佳，即组内实例距离最近，组间实例距离最远，最后形成 $k$ 个簇。该算法对于数值型属性的聚类效果较好，并且对于大数据集具有快速、简单、效率高的优点，算法基本思想和一般步骤如下：（1）设数据集合 $D=\{x_1,x_2,\cdots,x_n\}$，从 $n$ 个地块中随机选取其中的 $k$ 个地块作为初始聚类中心 $M_i(i=1,2,\cdots,k)$。（2）分别计算各地块 $x_m(m=1,2,\cdots,n)$ 到 $k$ 个初始聚类中心 $M_i$ 的距离，根据最小距离划分数据集合，将各个元素归到与其距离最小的类中，形成 $k$ 个类簇。（3）计算各类簇中元素平均值作为新的聚类中心。（4）相似度的计算采用欧氏距离，即 2 点之间的欧式空间直线距离。考虑邻近度为欧氏距离的数据，通常使用聚类的平方误差作为度量聚类质量的目标函数。聚类平方误差 $E$ 定义如下：

$$E=\sum_{i=1}^{k}\sum_{j=1}^{k_i}(x_j-M_i)^2。\qquad(4)$$

式中：$k_i$ 为第 $i$ 个类簇中包含地块的数量，重复步骤（2）、步骤（3），直至平方误差 $E$ 稳定在最小值，直到簇不再发生变化，最后获得 $k$ 个聚类具有各聚类内部紧凑、聚类间相异的特点。

本研究在试验分析的基础上，选择适用于数值型大数据集的 K－means 算法对数据进行聚类，在聚类中心的选取上，通过不断调节参数使误差平方和最小，来确定最优 $k$。聚类过程借助 WEKA 3.8 软件和 ArcGIS 10.2 实现。

1.2.4　Apriori 关联规则　数据关联是数据库中存在的一类重要的可被发现的知识。若 2 个或多个变量的取值之间存在某种规律性,就称为关联。关联分析的目的是找出数据库中隐藏的关联网。有时并不知道数据库中数据的关联函数,即使知道也是不确定的,因此关联分析生成的规则带有可信度。

关联规则是数据挖掘技术中的一种挖掘信息的技术,是从海量的数据中找到项与项之间有用的关联关系,关联规则挖掘就是从大量的数据中挖掘出有价值描述数据项之间相互联系的有关知识,有助于发现数据库中不同属性之间的联系[12-13]。

Apriori 算法是关联分析中应用最广泛的一种算法,是基于关联规则的基础,首先找到频繁集,再由频繁集推出关联的规则。算法的核心思想是基于频繁集理论的一种递推方法,其目的是根据最小支持度阈值和最小置信度阈值从给定的数据集中挖掘出期望的关联规则[9]。

该算法利用了一个层次顺序搜索的循环方法来完成频繁项集的挖掘工作。基本过程如下:

(1)首先计算所有的 C1;(2)扫描数据库,删除其中的非频繁子集,生成 L1(1 - 频繁项集);(3)将 L1 与自己连接生成 C2(候选 2 - 项集);(4)扫描数据库,删除 C2 中的非频繁子集,生成 L2(2 - 频繁项集);(5)依此类推,通过 Lk - 1(k - 1 - 频繁项集)与自己连接生成 Ck(候选 k - 项集),然后扫描数据库,生成 Lk(频繁 k - 项集),直到不再有频繁项集产生为止。


注:最小支持数为2

**图2　候选项集和频繁项集的产生过程**

满足最小支持度阈值和最小信任度阈值的关联规则称为强规则。通过使用数据挖掘软件 WEKA 及关联算法 Apriori 可以分析数据库中一些项集的关联性关系,找出某些强关联规则。

## 2　结果与分析

### 2.1　分区指标体系

作为广东省旱地改水田区域划分的分析研究,是基于土地适宜性评价的基础上进行筛选分区指标。结合广东省的实际情况,分别从地形、气候、水源、土壤、交通、社会经济条件等影响因素出发,参考农用地质量分等规程、农用地定级规程等资料来确定,选定具有典型性和稳定性的主导因子,包括高程、地形坡度、田面坡度、水源保障程度、排水条件、交通通达度、连片性、有效土层厚度、表层土壤质地、土壤剖面构型、pH 值、障碍层次、地表岩石露头度、土壤盐渍化程度等可选分区因子。

通过自动属性选择,按照相关度的大小,选取地形坡度、有效土层厚度、表层土壤质地、剖面构型、有机质含量、pH 值、障碍层深度、灌溉保证率、排水条件 9 个属性因子,广东省旱改水分区指标体系见表 1。

土壤是影响耕地质量的最基本要素[14-15]。在众多影响土壤质量的因素中,表土质地、剖面构型、pH 值直接影响土壤结构、土壤耕性、土壤阳离子交换量、土壤容重、土壤空隙状况等其他土壤理化性状;有效土层厚度和土壤有机质含量是最能反映土壤自然生产潜力的指标。

**表 1　广东省旱改水分区指标体系**

| 因子类别 | 分区因子 | 等级 | 指标 |
|---|---|---|---|
| 地形条件 | 地形坡度 | 1 | <2° |
| | | 2 | 2° ~ <5° |
| | | 3 | 5° ~ <8° |
| | | 4 | 8° ~ <15° |
| | | 5 | 15° ~25° |
| | 有效土层厚度 | 1 | ≥100 cm |
| | | 2 | 60 ~ <100 cm |
| | | 3 | 30 ~ <60 cm |
| | | 4 | <30 cm |
| | 表层土壤质地 | 1 | 壤土 |
| | | 2 | 黏土 |
| | | 3 | 沙土 |
| | | 4 | 砾质土 |
| | 剖面构型 | 1 | 通体壤、壤/沙/壤 |
| | | 2 | 壤/黏/壤 |
| | | 3 | 沙/黏/沙、壤/黏/黏、壤/沙/沙 |
| | | 4 | 沙/黏/黏 |
| | | 5 | 黏/沙/黏、通体黏、黏/沙/沙 |
| | | 6 | 通体沙、通体砾 |
| 土壤条件 | 有机质含量 | 1 | ≥4.0% |
| | | 2 | >3.0% ~4.0% |
| | | 3 | >2.0% ~3.0% |
| | | 4 | >1.0% ~2.0% |
| | | 5 | 0.6% ~1.0% |
| | | 6 | <0.6% |
| | pH 值 | 1 | 6.0 ~ <7.9 |
| | | 2 | 5.5 ~ <6.0 或 7.9 ~ <8.5 |
| | | 3 | 5.0 ~ <5.5 或 8.5 ~ <9.0 |
| | | 4 | 4.5 ~ <5.0 |
| | | 5 | <4.5 或 9.0 ~ <9.5 |
| | | 6 | ≥9.5 |
| | 障碍层深度 | 1 | 60 ~90 cm |
| | | 2 | 30 ~ <60 cm |
| | | 3 | <30 cm |
| | 灌溉保证率 | 1 | 充分满足 |
| | | 2 | 基本满足 |
| | | 3 | 一般满足 |
| | | 4 | 无灌溉条件 |
| 补充水田潜力 | 排水条件 | 1 | 排水体系健全 |
| | | 2 | 排水体系基本健全 |
| | | 3 | 排水体系一般 |
| | | 4 | 无排水条件 |

地形是影响耕地质量的重要因素。根据国家规定,25°以上坡地不能开发,小于 25°的范围,坡度越小,越适宜改造。地表起伏越大坡度越陡,土壤侵蚀作用越强,水土流失越严重;地形起伏越小,对农田水利化越有利。

补充水田潜力是改造水田的关键条件,灌溉保证率是灌溉用水量的保证程度,农田排水是改善农业生产条件,保证作物高产稳产的重要措施之一。灌溉保证率和排水条件越好,说明越适宜改造。

## 2.2　聚类分区结果分析

在 WEKA 中选定 SimpleK means 算法,随机选择聚类数 numClusters,设定参数为“SimpleK means N 5 – S 10”,Cluster mode 选取“Use training set”。经过多次不同 $k$ 值和样本种子值运算比较误差大小,最终选择 $k=5$,seed = 20 时误差平方和为最小。

聚类运算的结果中,数据集中的区域被分为 5 个相似的分组,聚类中心 $k$ 的各因子属性和标准差见表 2,在省级行政图上表示出 5 类区域的空间分布见图 3。

各项标准差均不超过 0.5,说明同一分组的实例间距离较近,分组有效,无须剔除异常值。

从整体分布来看,5 类区域在广东省各县区分布比较均匀,但由于地理条件的差异,西部地区种类相对密集。从各分类来看,聚类 0 中限制性因子障碍层深度远远高于其他几类,

表 2　聚类中心属性

| 类别 | 指标 | 地形坡度 | 有效土层厚度 | 表土质地 | 剖面构型 | 有机质含量 | pH 值 | 灌溉保证率 | 排水条件 | 障碍层深度 |
|---|---|---|---|---|---|---|---|---|---|---|
| 聚类 0 | 均值 | 0.991 | 0.186 | 0.951 | 0.832 | 0.057 | 0.797 | 0.710 | 0.758 | 0.999 |
| | 标准差 | 0.009 8 | 0.068 3 | 0.144 4 | 0.294 4 | 0.024 8 | 0.104 1 | 0.341 9 | 0.279 7 | 0.021 3 |
| 聚类 1 | 均值 | 0.994 | 0.215 | 0.835 | 0.330 | 0.059 | 0.795 | 0.904 | 0.795 | 0.001 |
| | 标准差 | 0.008 5 | 0.045 4 | 0.214 6 | 0.177 1 | 0.023 4 | 0.096 1 | 0.160 6 | 0.253 0 | 0.013 0 |
| 聚类 2 | 均值 | 0.992 | 0.198 | 0.931 | 0.861 | 0.052 | 0.773 | 0.202 | 0.244 | 0.048 |
| | 标准差 | 0.007 8 | 0.060 5 | 0.173 9 | 0.252 9 | 0.022 5 | 0.097 4 | 0.293 7 | 0.195 1 | 0.212 8 |
| 聚类 3 | 均值 | 0.990 | 0.209 | 0.931 | 0.760 | 0.048 | 0.793 | 0 | 0.918 | 0 |
| | 标准差 | 0.009 9 | 0.058 2 | 0.175 9 | 0.314 4 | 0.022 2 | 0.107 1 | 0.006 0 | 0.098 3 | 0.007 0 |
| 聚类 4 | 均值 | 0.992 | 0.206 | 0.991 | 0.993 | 0.060 | 0.804 | 0.924 | 0.779 | 0.001 |
| | 标准差 | 0.009 4 | 0.065 1 | 0.066 1 | 0.033 3 | 0.019 8 | 0.104 2 | 0.137 7 | 0.274 5 | 0.018 1 |

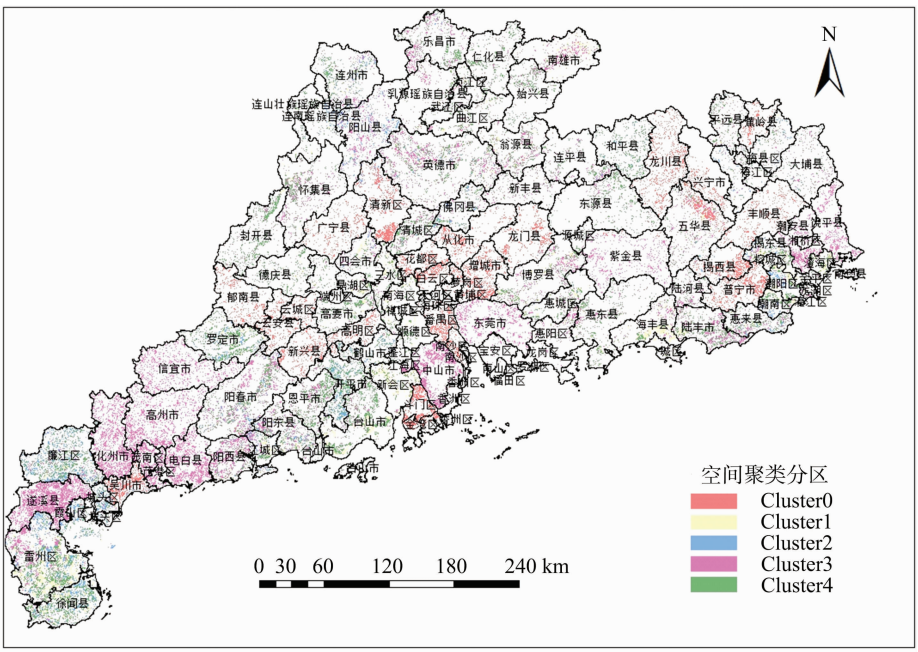


图3　广东省“旱改水”空间聚类分区

地形坡度、有效土层厚度值相对较低，主要分布在珠三角平原中部，粤东沿海区西北部一些坡度陡、土壤条件稍差的地区；聚类 1、聚类 2 是分布最少的 2 类，聚类 1 中地形坡度、有效土层厚度最高，但表土质地最低，零星分布在雷州半岛和珠三角南部、粤东沿海地带一些地形平缓、土壤自然生产潜力较高的地区，但由于海水的作用产生土壤盐渍化使得土壤肥力差；而聚类 2 中 pH 值、排水条件最低，其他属性值较均匀，处于中等位置，与聚类 1 分布相似，只是多为内陆地区，可能由于该地区土壤呈酸性；聚类 3 中地形坡度、有机质含量、灌溉保证率、障碍层深度最低，尤其是灌溉保证率、障碍层深度为 0 而排水条件却最高，均匀分布在全省各个地区，以粤西沿海区东北部、粤西北山区北部、粤东沿海区东部较为集中，该地区坡度较大，土壤自然生产潜力较低，水源灌溉条件差，可能不适宜进行改造；聚类 4 分布也较为广泛，集中在粤西北山区大部、粤西沿海区，该类地区表土质地、剖面构型、有机质含量、pH 值、灌溉保证率都是最高，其他几个属性值也都不低，该地区地势平坦、土壤肥沃、水源丰富，比较适宜改造。

## 2.3　关联规则挖掘结果分析

WEKA 平台中的 Filtered Associator 增加了数据过滤器，将经过筛选转换的数据在 WEKA 所支持的任意基本关联算法中进行分析。根据关联规则，将研究数据设置为这样的项目集：I = {地形坡度、有效土层厚度、表土质地、剖面构型、有机质含量、pH 值、障碍层深度、灌溉保证率、排水条件}。由于 Apriori 算法无法对数值型数据进行操作，在做关联分析前将原数据类型转换为 9 个类别的分类型数据，数据转换完成后，设置参数“Apriori - N 10 - T 0 - C 0.9 - D 0.05 - U 1.0 - M 0.1 - S - 1.0 - c - 1”进行关联分析。不断扫描数据样本后产生的关联规则如下：

Best rules found：

(1) PMGX = TTR 1 003 787 = = > BTZD = R 972 990 < conf：(0.97) > lift：(1.15) lev：(0.08) [85 355] conv：(5.23)；

(2) PMGX = TTR YJZHL = 2 - 3 482 938 = = > BTZD = R 475 644 < conf：(0.98) > lift：(1.17) lev：(0.05) [47 915]

conv:(10.3);

(3) PMGX = TTR GGBZL = 1j 467 383 = = > BTZD = R 458 865 < conf:(0.98) > lift:(1.16) lev:(0.04) [45 035] conv:(9.07);

(4) PMGX = TTR PSTJ = 1j 393 078 = = > BTZD = R 384 780 < conf:(0.98) > lift:(1.16) lev:(0.04) [38 708] conv:(8.13);

(5) PMGX = TTR PHZ = 6 - 7.9 381 444 = = > BTZD = R 372 374 < conf:(0.98) > lift:(1.16) lev:(0.05) [49 997] conv:(6.51);

(6) PMGX = TTR ZACSD = < 30 724 958 = = > BTZD = R 700 451 < conf:(0.97) > lift:(1.15) lev:(0.06) [59 086] conv:(4.81);

(7) YXTCHD = > = 100 PMGX = TTR ZACSD = < 30 509 355 = = > BTZD = R 492 916 < conf:(0.96) > lift:(1.12) lev:(0.03) [13 953] conv:(3.25);

(8) PMGX = TTR YJZHL = 2 - 3 ZACSD = < 30 104 102 = = > BTZD = R 102 718 < conf:(0.98) > lift:(1.15) lev:(0.03) [13 745] conv:(10.92);

(9) PMGX = TTR GGBZL = 1j ZACSD = < 30 110 655 = = > BTZD = R 108 576 < conf:(0.98) > lift:(1.15) lev:(0.03) [14 002] conv:(7.73);

(10) YXTCHD = > = 100 PMGX = TTR ZACSD = < 30 138 634 = = > BTZD = R 132 439 < conf:(0.96) > lift:(1.12) lev:(0.03) [13 953] conv:(3.25)。

观察关联规则,每条规则的置信度均很高(表 3 至表 5),从中可以获取改造因子不同属性之间的联系,筛选置信度最高(0.98)的关联规则得出:(1)由剖面构型是通体壤、有机质含量 2 ~ 3、障碍层深度 < 30 与表土质地为壤土的关联规则可知,剖面构型是通体壤、有机质含量在 2 ~ 3 且限制性因子障碍层深度满 足 < 30 的地块,它们的表土质地也相对应是最优的壤土。(2)同样,剖面构型是通体壤,灌溉保证率是 1 级或排水条件是 1 级或者 pH 值在 6.0 ~ 7.9 的地块通常表土质地也是最优的。

**表 3　2 项因子关联规律**

| 左项因子 | 频次 | | 右项因子 | 频次 | 置信度 |
|---|---|---|---|---|---|
| 剖面构型 = 通体壤 | 1 003 787 | = = > | 表土质地 = 壤土 | 972 990 | 0.97 |
| 有机质含量 = 1 ~ 2 | 289 441 | = = > | 表土质地 = 壤土 | 264 726 | 0.91 |
| 地形坡度 = 5 ~ 8 | 280 302 | = = > | 表土质地 = 壤土 | 253 064 | 0.90 |

**表 4　3 项因子关联规律**

| 左项因子 | 频次 | | 右项因子 | 频次 | 置信度 |
|---|---|---|---|---|---|
| 剖面构型 = 通体壤;有机质含量 = 2 ~ 3 | 482 938 | = = > | 表土质地 = 壤土 | 475 644 | 0.98 |
| 剖面构型 = 通体壤;灌溉保证率 = 1 级 | 467 383 | = = > | 表土质地 = 壤土 | 458 865 | 0.98 |
| 剖面构型 = 通体壤;排水条件 = 1 级 | 393 078 | = = > | 表土质地 = 壤土 | 384 780 | 0.98 |
| 剖面构型 = 通体壤;pH 值 = 6.0 ~ 7.9 | 381 444 | = = > | 表土质地 = 壤土 | 372 374 | 0.98 |
| 剖面构型 = 通体壤;障碍层深度 < 30 | 724 958 | = = > | 表土质地 = 壤土 | 700 451 | 0.97 |
| 剖面构型 = 通体壤;有效土层厚度 ≥ 100 | 509 355 | = = > | 表土质地 = 壤土 | 492 916 | 0.96 |
| 地形坡度 = 2 ~ 5;有机质含量 = 2 ~ 3 | 172 754 | = = > | 表土质地 = 壤土 | 159 759 | 0.92 |
| 有机质含量 = 2 ~ 3;障碍层深度 < 30 | 324 371 | = = > | 表土质地 = 壤土 | 295 487 | 0.91 |
| 有效土层厚度 ≥ 100;障碍层深度 < 30 | 399 502 | = = > | 表土质地 = 壤土 | 360 346 | 0.90 |

**表 5　4 项因子关联规律**

| 左项因子 | 频次 | | 右项因子 | 频次 | 置信度 |
|---|---|---|---|---|---|
| 剖面构型 = 通体壤;有机质含量 = 2 ~ 3;障碍层深度 < 30 | 104 102 | = = > | 表土质地 = 壤土 | 102 718 | 0.99 |
| 剖面构型 = 通体壤;灌溉保证率 = 1 级;障碍层深度 < 30 | 110 655 | = = > | 表土质地 = 壤土 | 108 576 | 0.98 |
| 有效土层厚度 ≥ 100;剖面构型 = 通体壤;障碍层深度 < 30 | 138 634 | = = > | 表土质地 = 壤土 | 132 439 | 0.96 |
| 有机质含量 = 2 ~ 3;灌溉保证率 = 1 级;障碍层深度 < 30 | 257 775 | = = > | 表土质地 = 壤土 | 233 467 | 0.91 |

### 2.4　优先区的选择结果分析

由于空间聚类分区结果并不能直观地表明优先区为哪一类,为了挖掘出到底哪一类是最优先区,将 9 个指标属性的所有分区单元进行关联规则挖掘。土壤是影响耕地质量的最基本要素,而表土质地是壤土作为最频繁项集,本身就是土壤的最优条件,因此和与它形成最强关联的因子属性的组合作为选择优先区的决策标准。

将置信度最高的 6 条关联规则作为选择优先区的标准,基于已经划分好的聚类区,发现符合最强关联规则的聚类区是聚类 4,广东省旱改水改造最优先区分布见图 4。

(1)从优先区地块数量来看,共涉及 13 个市,62 个县区,408 841 个地块,共计 975 628.35 $hm^2$,占总数的 28.36%。(2)从图 3 可以看出,从整体分布来看,优先区主要分布在地势平坦、水资源充足的粤西北山区和粤西沿海区,其中粤西沿海区的优先改造区多集中在雷州半岛。这些地区的旱地及耕地后备的灌溉保证率、地形坡度等条件相对优越,并且广东省旱地和望天田主要分布在湛江市、清远市、阳江市、韶关市等地,尤以湛江市为最,55% 的耕地为旱地和望天田,较为适宜于“旱改水”工程土地整治。珠三角平原区仅在江门市、肇庆市、惠州市分布,这是因为此地区河流水网密布,灌溉条件优越,水田占据较大比例,开发水田的潜力已经基本用尽。(3)从各县区潜力来看,雷州县、廉江县、台山县、英德县优先区地块占较高比例,潜力较大。这些区域距水源相对较近,地面坡度较小,有效土层相对较厚,稍经整治便会具有较好的农业生

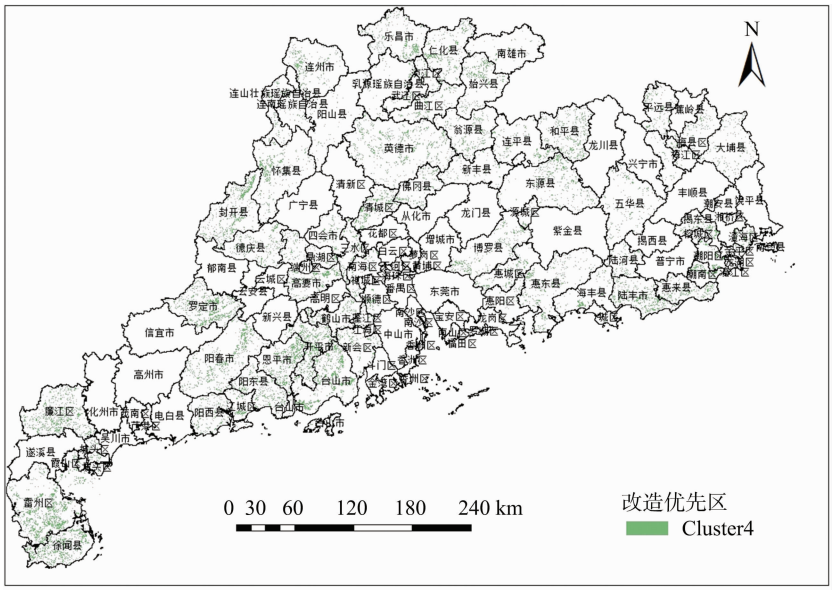


图4　广东省旱改水优先区空间分布示意

产条件。

## 3　讨论与结论

笔者是以农用地定级和分等规程划分的指标为依据,基于空间数据挖掘的方法开展的“旱改水”优先区选择研究尝试,在实践中难免存在一些不足之处,并与其他分区结果有一定的差别。需要说明的是:(1)分区指标体系中指标没有考虑权重,主要是因为指标的筛选不是通过专家打分,而是数据的自动属性选择的相关性大的指标,通过数据标准化转换为数值型数据来统一运算。(2)分区结果中同一县区可能呈现不同类别,这是因为研究以图斑作为分区单元而不是县级行政区,这样可以避免同一县区雷同性。(3)关联规则主要作为分区结果的进一步决策,更科学地表明聚类 4 是优先区,不作为主要的研究结果。(4)分区结果与其他学者的分区方案存在一定的差异,这主要归因于研究角度、指标体系和分区依据等的差别。(5)本研究的目的在于为广东省“旱改水”改造顺序提供一定的科学依据,对于各地详细的改造顺序,还需要在实践中结合县(市)的具体社会经济特点进一步探索研究。

通过研究获得如下结论:(1)以广东省 1 683 403 块旱地为分区单元,结合研究区域的地形、土壤、补充水田潜力等条件,从可选的 14 个指标中通过属性选择相关性较高的地形条件、土壤条件和补充水田潜力条件 9 个指标,建立了“旱改水”分区指标体系。(2)通过 K - means 聚类分析得到 5 类区域,并以 Apriori 关联规则挖掘出的最强关联作为决策选择出优先区,主要分布在粤西北山区和粤西沿海区,并分析了该区域的实际条件和作为优先区的原因。(3)研究结果表明,采用空间聚类和关联规则相结合的数据挖掘方法,科学地挖掘出“旱改水”优先区,为各级政府“旱改水”改造工作提供重要的科学依据,有利于节省资金和人力物力资源。

**参考文献:**

[1]齐艳红,潘　旭,赵映慧. 浙江省江山市旱改水适宜性评价[J]. 安徽农业科学,2016(35):202 - 204.

[2]刘正国,游振波,黄　俊. 江西省旱地改水田土地整治研究——以永丰县瑶田镇湖西村旱改水项目为例[J]. 安徽农业科学,2015(36):185 - 187,229.

[3]王　君. 旱地改水田项目中新增水田的适宜性评价方法研究——以湖南省华容县梅田湖镇北剅口村金鸡村旱地改水田项目为例[J]. 农业与技术,2015(24):55 - 56.

[4]陈印军,肖碧林,陈京香. 我国耕地“占补平衡”与土地开发整理效果分析与建议[J]. 中国农业资源与区划,2010(1):1 - 6.

[5]张　琳,张凤荣,薛永森,等. 中国各省耕地数量占补平衡趋势预测[J]. 资源科学,2007(6):114 - 119.

[6]胡　科,石培基. 甘肃省耕地质量评价研究[J]. 中国土地科学,2008(11):38 - 43.

[7]王大龙,秦　琦. 关于数据挖掘原理与算法的浅析[J]. 科技创新导报,2010(2):193 - 193.

[8]Dash M,Liu H. Feature selection for classification[J]. Intelligent Data Analysis,1997,1(1/2/3/4):131 - 156.

[9]Liao Z Y,Fu X F,Wang Y G. The research of improved apriori algorithm[J]. Applied Mechanics and Materials,2013,263/264/265/266:2179 - 2184.

[10]Frank E,Hall M,Trigg L,et al. Data mining in bioinformatics using Weka[J]. Bioinformatics,2004,20(15):2479 - 2481.

[11]Hall M,Frank E,Holmes G,et al. The WEKA data mining software[J]. ACM Sigkdd Explorations Newsletter,2009,11(1):10 - 18.

[12]Shahrivari S,Jalili S. Single - pass and linear - time k - means clustering based on MapReduce[J]. Information Systems,2016,60:1 - 12.

[13]Guo Y,Wang M X,Li X. Application of an improved Apriori algorithm in a mobile e - commerce recommendation system[J]. Industrial Management & Data Systems,2017,117(2):287 - 303.

[14]郭　敏,李淑杰. 基于局部空间自相关的耕地质量空间集聚性和保护分区——以吉林省九台市为例[J]. 江苏农业科学,2017,45(3):206 - 210.

[15]杜会石,张　爽,王　柏,等. 自清式土壤研磨机转速对土壤元素分析值的影响[J]. 江苏农业科学,2017,45(10):170 - 173.