DOI:10.13718/j.cnki.xsxb.2019.09.012

基于互联网+数据挖掘的农业数据平台设计①

向模军

成都农业科技职业学院 信息技术中心,成都 611130

摘要:随着农业数据规模日益增加,相应的、有针对性的数据挖掘平台设计成为迫切需要.该文设计了基于互联网+数据挖掘的农业数据平台,包括交互层、功能层、数据层 3 个层次.功能层是平台的核心,负责数据预处理、数据挖掘、挖掘结果展示等任务.针对 Apriori 算法进行了改进,通过频度集合、支持度集合、地址集合的联合使用,避免了重复扫描和频度冗余,进一步提升了挖掘算法的效率.以水稻生长中的二化螟虫害问题为研究对象,展开平台性能的验证性实验.实验结果表明:4 种因素对二化螟爆发影响的强弱顺序为 5 月份降水最大,其次是 5 月份的温度,再次是 2 月份的温度,最后是 2 月份的降水.

关 键 词:互联网+;数据挖掘;农业数据平台;二化螟

中图分类号:TP311 **文献标志码**:A **文章编号**:1000-5471(2019)09-0076-06

随着我国经济的快速发展和各种科学技术的不断升级,农业生产和经营进入了一个前所未有的崭新时代[1].一方面,农业生产和经营中的科技含量不断提高,出现了大量的物联网农业基地[2];另一方面,农业生产和经营中的数据量不断增大,既包括了农作物的生长信息,也包括了各类仪器给出的监测信息[3].在这样的自动化、智能化背景下,如何从海量的数据中为农业生产和经营挖掘出有效的信息,是一个亟待解决的关键问题.借助互联网技术为农业生产和经营构建一个综合数据信息平台,实现对农业数据的多点实时采集、存储和处理,进而采用数据挖掘算法从海量数据中提取出对后续生产和经营的有效数据,是解决此问题的关键所在[4].从全球范围来看,基于数据挖掘技术的平台建设已经有了多年的历史.早在 20 世纪 90 年代,就先后出现了 Salford 系统、DB Miner 系统[5-6],数据挖掘算法当时普遍采用了基于关联规则的算法,用于对数据的整理、过滤、提取和挖掘[7].进入 21 世纪以后,数据挖掘平台开始向第 4 代演进,出现了 SPSS Clementine 系统,挖掘算法也出现了决策树算法、预测算法等更为先进的算法[8-9].基于互联网+数据挖掘的数据平台,在农业领域中也有了一定程度的应用,如农业环境监测、土壤侵蚀监测、农业生产过程监测等等.在农业数据挖掘算法上,关联分析、聚类分析、决策树分析、粗糙集理论被广泛采用[10-12].在本文的研究工作中,将依托互联网进行农业数据平台的体系结构设计,进而设计更具针对性的数据挖掘算法,再通过实验加以验证平台的设计效果.

1 3 层次互联网+数据挖掘的农业数据平台设计

根据网络构成的体系结构,基于网络协议的平台设计大多是层次结构,包含了应用层、表示层、传输层、协议层、数据层等等.从网络构成的层次结构理论出发,为了实现农业数据平台设计,实现平台上的数据采集、数据存储、数据挖掘、各网点交互、用户使用等功能,本文给出了一种 3 层次的框架设计,如图 1 所示.

第 1 个层次为交互层,主要包括了农业数据采集、处理、挖掘的可视化结果展示,对用户功能的响应,各

① 收稿日期:2019-02-22
基金项目:四川省科学技术厅项目(18KPPX0018);成都市科学技术局项目(2018-YF05-00974-SN).
作者简介:向模军(1974-),男,硕士,副教授,主要从事 Web 挖掘研究.

节点之间网络信息传输功能的实现. 这个层次的设计，主要依据 HTML(Hyper Text Markup Language)动态网页技术和数据可视化技术.

第 2 个层次为功能层，是与农业数据处理相关的各种功能设计. 结构功能包括数据库的功能配置、数据库的控制引擎、数据库的建模重构. 具体功能包括农业数据选取抽样、农业数据预处理、农业数据变换、农业数据挖掘、挖掘结果评估决策及决策采纳.

第 3 个层次为数据层，主要负责对数据库的管理. 这里涉及到的数据库包括农业数据库、相关数据库. 其中，农业数据库负责存放农业生产和经营的相关数据，相关数据库则存放各种相关的仪器设备信息及其他有关材料的信息等.

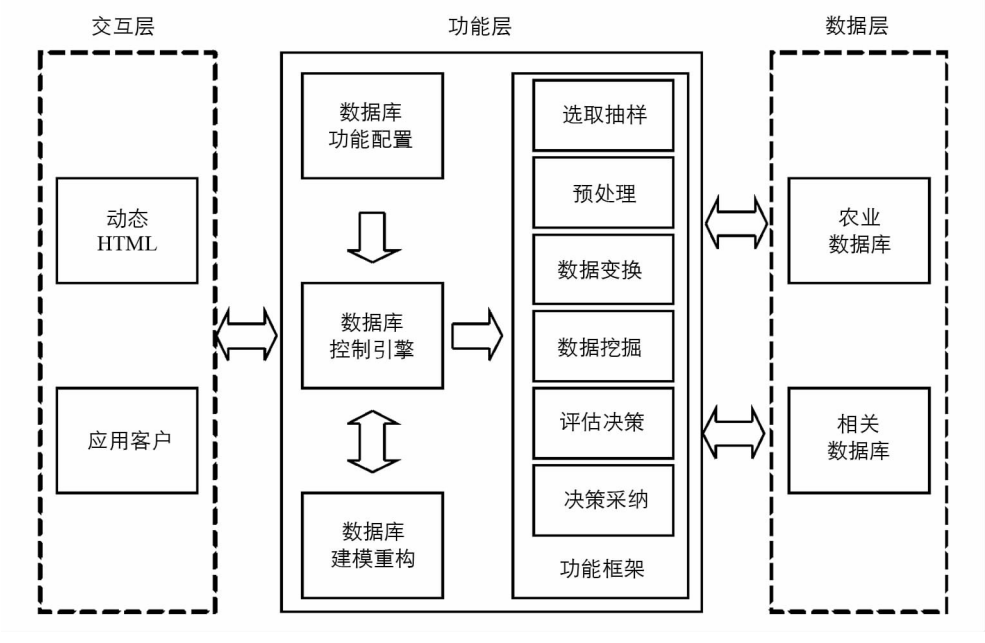

图 1　本文的 3 层次农业数据平台

在这个平台下，整个数据挖掘工作的运行机理如图 2 所示.

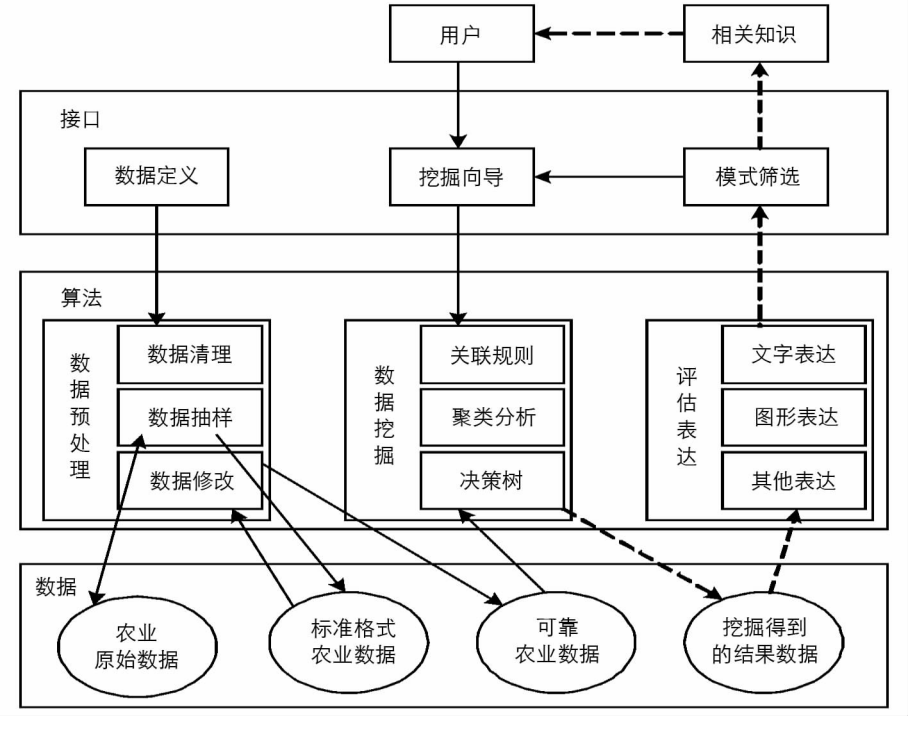

图 2　本文数据挖掘平台的运行机理

首先，做好数据准备工作，包括农业数据库构建、对数据库中数据进行抽样修改和定期清理更新. 其次，用户向平台发起数据挖掘的具体需求，通过挖掘向导关联到具体的挖掘模块，平台开始执行相关的挖掘算法. 再次，平台根据挖掘算法得到的结果向用户提供评估结果，同时根据不同的表达形式在平台上完成相应的列表展示或者图形展示.

2 基于改进 Apriori 算法的农业数据挖掘方法设计

数据挖掘是整个农业数据平台设计的核心功能，也是本文设计中的一个关键工作. 从已经出现的数据挖掘方法来看，Apriori 算法是一种非常常见的方法，不仅原理简单，而且易于编程实现，同时具有较高的挖掘效率.

但是，在实际应用中 Apriori 算法也表现出一些问题，主要有：①Apriori 算法的挖掘过程中会形成一个较大规模的候选集合；②Apriori 算法需要对被处理的数据执行多次扫描；③Apriori 算法中的部分关联规则存在冗余.

为此，本文在 Apriori 算法的基础上改进以提升其性能，改进后的 Apriori 算法步骤如下：

第 1 步，对农业数据库中的数据进行全扫描处理，从而确定出不含重复数据的候选数据集合. 为了达成比传统 Apriori 算法更好的效果，此处配置一个累加器，累加器的计数结果作为对候选数据集中候选元素的挖掘频度的支持，进而得到频度集合 $L1$ 和对应支持度集合 $C1$.

第 2 步，对频度集合执行自链接处理，从而形成二阶候选频度集合 $L2$ 和对应支持度集合 $C2$. 在这个操作过程中，根据支持度最小剪枝处理原则，所有支持度小于 2 的数据都将被剔除出考虑范围. 为了便于后续查找，为 $L2$ 配置地址集合 $A2$，如图 3 所示.

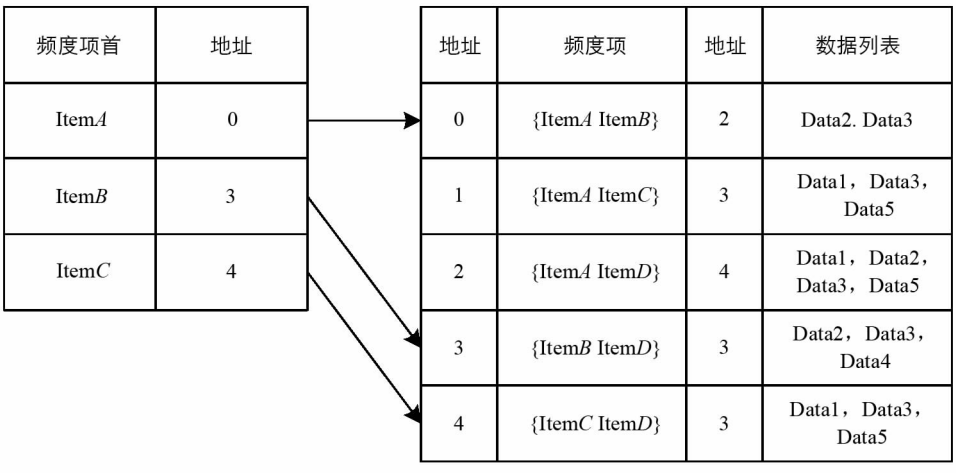

图 3 频集合 $L2$ 对应的地址集合 $A2$

第 3 步，不断拓展第 2 步的处理，进行到第 k 次时，已经得到频度集合 $L(k-1)$、对应支持度集合 $C(k-1)$、对应地址集合 $A(k-1)$. 对于本次支持度集合 $C(k)$，如果某条数据不满足关联条件，将不再将其纳入 $L(k)$.

不断执行上述处理，当 $C(k)$被清空时，挖掘分类执行完毕.

3 实验结果与分析

为了验证本文设计的农业数据平台和改进的 Apriori 算法的可用性和有效性，展开如下实验研究.

研究的主要对象为农业作物病虫害问题. 对于水稻、棉花等农作物，二化螟是非常常见的虫害. 如果能理清二化螟和温度、降水等气候条件的关系，就可以预判出二化螟的爆发节点和规模，从而进行有效的预防，提升水稻产量.

根据二化螟的爆发规律，2 月份和 5 月份的温度和降水条件是非常重要的影响因素，也是本文要分析的主要数据. 本文以 2001－2018 年 18 年间这 2 个月份的温度和降水相关数据为研究对象，通过改进

Apriori算法挖掘出二化螟与气候条件之间的规律(表 1). 上述数据来源于中国农业大数据平台和中国农业信息网，对于数据的统计采用 Eviews 软件完成.

表 1　本文研究的相关数据(2001－2018 年度)

年份	2 月份温度/℃	2 月份降水/mm³	5 月份温度/℃	5 月份降水/mm³	二化螟总量/万例
2001	1.6	29.9	17.1	88.9	102.4
2002	－1.2	31.4	16.9	140.6	133.5
2003	－1.1	67.2	15.2	166.8	241.7
2004	－1.8	32.3	15.3	241.4	268.6
2005	－0.6	101.5	16.9	299.5	300.2
2006	1.5	29.4	17.4	228.3	128.7
2007	3.2	28.1	18.1	75.4	42.3
2008	2.5	3.3	16.5	14.2	200.5
2009	4.0	18.5	16.2	100.6	105.4
2010	－0.3	21.2	18.8	149.2	388.7
2011	－0.6	2.4	17.9	88.6	128.9
2012	3.6	1.6	16.5	83.5	151.2
2013	2.5	40.8	14.2	131.6	217.6
2014	1.6	78.3	15.6	252.5	309.3
2015	0.4	19.2	19.2	244.7	266.6
2016	－0.8	87.5	18.3	100.3	104.5
2017	－1.2	42.1	17.6	66.7	108.4
2018	0.6	5.5	20.6	108.5	221.7

将上述表格中的数据整理到本文的农业数据平台之下，进而登陆平台进行数据挖掘处理. 因为本文的平台尚处于内网应用阶段，因此平台的访问网址为 http：//192.168.0.1/Agricluture Platform/Data Mining/.

登陆平台后，进入数据挖掘功能，将相关数据进行图形展示的效果如图 4 所示.

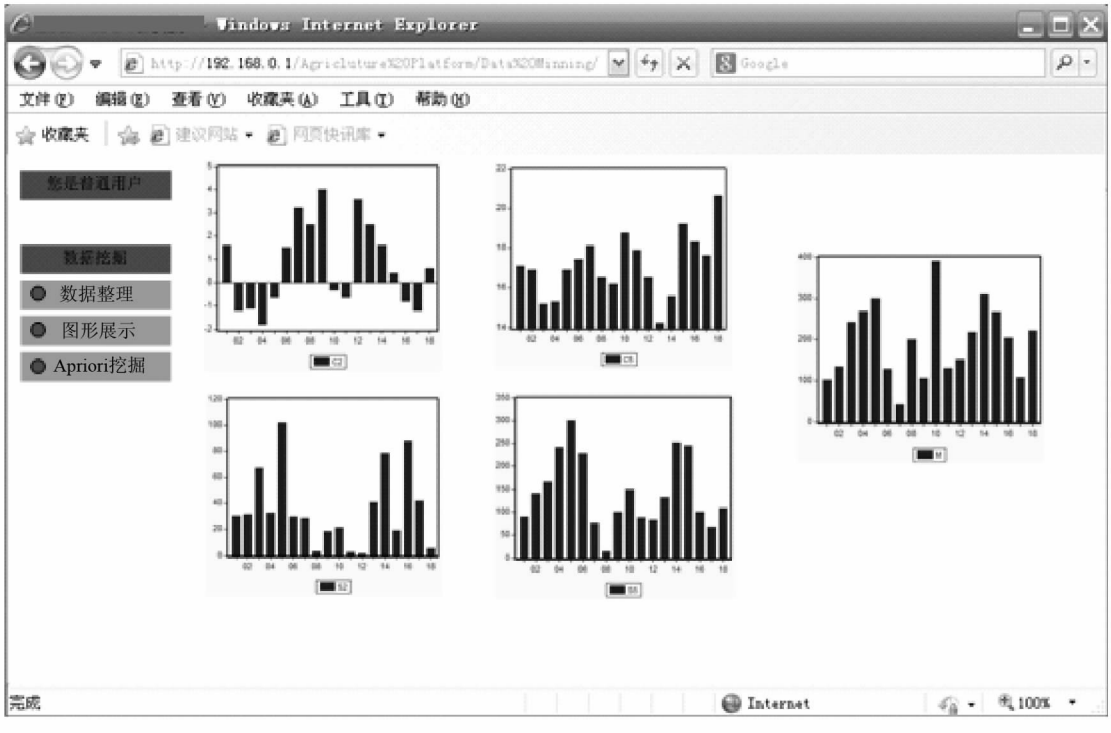

图 4　本文设计的基于互联网＋数据挖掘的农业数据平台

在图4中，基于互联网+数据挖掘的农业数据平台可以通过任意浏览器操作，在浏览器上方输入网址即可进入. 进入到这个Web界面以后，主视图左侧是一系列的功能菜单，本界面下就包含了用户信息和数据挖掘功能2个菜单. 数据挖掘功能菜单下，又包含了数据整理、图形展示、Apriori挖掘3个功能按钮，点击后可以激活各自后台封装的功能函数. 主视图的中间区域是数据整理的图形结果和数据挖掘的信息展示区域.

进一步点击Apriori挖掘功能，执行本文的改进Apriori算法，可以得到如下结论：

1) 二化螟爆发规模确实受到气候因素的影响，温度和降水都直接影响二化螟爆发.

2) 从表1中的四个要素来看，它们对二化螟爆发影响的强弱顺序为5月份的降水最大，其次是5月份的温度，再次是2月份的温度，最后是2月份的降水.

3) 对于5月份降水量而言，当降水量介于75.4～100.6 mm^3 时，二化螟的爆发量小于151.2万例. 这表明降水量越少，二化螟爆发的规模越小. 相比之下，2月份的降水对于二化螟爆发的影响，要弱于5月份的降水. 当然，这与综合情况有关，因为2月份的温度不适合二化螟爆发.

4) 对于5月份的温度而言，相对温度低时二化螟爆发规模较大，相对温度高时二化螟爆发规模较小. 虽然2月份的温度不适合二化螟爆发，但也并非温度越高就越好.

在本文设计的农业数据平台下，借助改进Apriori算法，挖掘出了二化螟爆发于气候条件的关系.

4 结　论

针对农业数据便捷的智能化处理问题，本文研究了基于互联网+数据挖掘的农业数据平台. 设计了3层次的平台框架，供用户进行最后的决策. 针对数据挖掘这一核心工作，对Apriori算法进行了改进，进一步提升了挖掘算法的效率. 针对水稻生长中的二化螟虫害问题，对平台的可视化效果和数据挖掘性能进行了展示. 改进Apriori算法清晰地找出了二化螟爆发规模和气候条件之间的关系，为用户的防虫害工作提供了判断依据.

参考文献：

[1] 陈祺琪，张俊飚，程琳琳，等. 农业科技资源配置能力区域差异分析及驱动因子分解 [J]. 科研管理，2016，37(3)：110-123.

[2] 杜克明，褚金翔，孙忠富，et al. WebGIS在农业环境物联网监测系统中的设计与实现 [J]. 农业工程学报，2016，25(4)：171-178.

[3] DUVEILLER G，DEFOURNY P. A Conceptual Framework to Define the Spatial Resolution Requirements for Agricultural Monitoring Using Remote Sensing [J]. Remote Sensing of Environment，2010，114(11)：2637-2650.

[4] 江　明. 基于数据挖掘的休闲农业交易平台的设计与研究 [D]. 杭州：浙江理工大学，2015.

[5] ANTLE J M，BASSO B，CONANT R T，et al. Towards a New Generation of Agricultural System Data，Models and Knowledge Products：Design and Improvement [J]. Agricultural Systems，2017，155：255-268.

[6] ANTLE J M，JONES J W，ROSENZWEIG C. Next Generation Agricultural System Models and Knowledge Products：Synthesis and Strategy [J]. Agricultural Systems，2017，22(1)：155，179.

[7] 佚　名. 北方设施农业气象灾害监测预警智能服务系统设计与实现 [J]. 农业工程学报，2018，34(23)：149-156.

[8] SALI G，MONACO F，MAZZOCCHI C，et al. Exploring Land Use Scenarios in Metropolitan Areas：Food Balance in a Local Agricultural System by Using a Multi-objective Optimization Model [J]. Agriculture & Agricultural Science Procedia，2016，8：211-221.

[9] 赵立安，李修华，周永华，等. 基于农业物联网的火龙果生长环境大数据分析 [J]. 节水灌溉，2018(3)：58-62.

[10] MAJUMDAR J，NARASEEYAPPA S，ANKALAKI S. Analysis of Agriculture Data Using Data Mining Techniques：Application of Big Data [J]. Journal of Big Data，2017，4(1)：11-20.

[11] MOROTA G, VENTURA R V, SILVA F F, et al. Machine Learning and Data Mining Advance Predictive Big Data Analysis in Precision Animal Agriculture [J]. Journal of Animal Science, 2018, 96(4): 1540-1550.

[12] THORP K R, WANG G, BRONSON K F, et al. Hyperspectral Data Mining to Identify Relevant Canopy Spectral Features for Estimating Durum Wheat Growth, Nitrogen Status, and Grain Yield [J]. Computers & Electronics in Agriculture, 2017, 136: 1-12.

Design of Agricultural Data Platform Based on Internet ＋ Data Mining

XIANG Mo-jun

Information Technology Center of Chengdu Agricultural College, Chengdu 611130, China

Abstract: With the increasing scale of agricultural data, the design of corresponding data mining platform has become an urgent problem. An agricultural data platform has been designed in this paper based on Internet ＋ data mining, which consists of three levels: interaction layer, function layer and data layer. Functional layer is the core of the platform, responsible for data preprocessing, data mining, mining results display and other tasks. The Apriori algorithm is improved. By using frequency set, support set and address set together, it avoids repeated scanning and frequency redundancy, and further improves the efficiency of mining algorithm. Taking the problem of rice stem borer pests in rice growth as the research object, the validation experiment of platform performance was carried out. The results show that the order of four factors influencing the outbreak of stem borer was the precipitation in May, followed by the temperature in May, February and February.

Key words: Internet ＋, data mining, agricultural data platform, Chilo suppressali

责任编辑　夏　娟