一种基于 Web 的农业数据挖掘平台①

罗凤娥

(湖南人文科技学院 通信与控制工程系，娄底 417000)

摘 要：针对农业数据具有季节性、时效性、多层次性和求新性等复杂特点，以及当前数据挖掘软件中普遍存在的功能耦合过于紧密、数据挖掘算法难以灵活替用等问题，提出了一种基于 Web 的农业数据挖掘平台的原理框架，并基于责任链式模式实现了原型平台 agri-Miner，能保证体系结构的可扩展性以及数据挖掘算法的灵活可替换性。最后，成功应用到农业病虫害以验证平台的有效性。

关键词：农业数据挖掘；原理框架；责任链模式；可扩展性

Web-Based Agricultural Data Mining Platform

LUO Feng-E

(Department of Communication & Control Engineering, Hunan Institute of Humanities, Science&Technology, Loudi 417000, China)

Abstract: As the agricultural data has such complex characteristics as seasonality, timeliness, multi-level nature as well as novelty, furthermore, the present data mining software is closely coupled, data mining algorithms is hard to reused and inflexible replaced, this paper presents a theoretic framework for Web-based agricultural data mining platform. Then, it implements prototyping platform called agri-Miner with Responsibility Chain Pattern, thus to ensure this architecture to be scalable and data mining arithmetic to be replaced. Finally, this platform is successfully applied in data analysis as plant diseases and insect pests, so to verify its effectiveness of Web-based agricultural data mining platform.

Key words: agricultural data mining; theoretic framework; responsibility chain pattern; scalability

随着农业数据的快速积累和增长，如何利用海量的农业数据以获取科学的农业知识、规律和决策支持信息成为非常重要的课题。数据挖掘技术定义为提取隐含在数据集合中新颖的、人们事先不知道的、潜在有用的信息和知识。将数据挖掘技术引入到农业领域数据处理，可以有效地从农业数据中找出潜在的和有用的农业知识，从而为农业领域提供规划、预警、决策等服务，提高农业环境与农产品质量水平，促进农业生产高效、协调和可持续发展。

我国将数据挖掘技术应用到农业领域数据挖掘比较典型的工作有：天津大学郑向群基于数据仓库的农业环境信息决策系统，包括农业环境信息数据仓库、联机分析处理和数据挖掘等功能模块，基于数据仓库和工作流挖掘技术实现了土壤环境监测全流程优化问题[1,2]，从土壤养分的数据仓库中挖掘出土壤肥力评价规则，指导农业生产科学施肥。江西庐山区水务局查骏雄提出了一种土壤侵蚀分析系统中的数据挖掘方案，通过对大量的水土保持信息数据的分析，从中抽取潜在的土壤侵蚀的变化规律与变化模式，为用户提供土壤侵蚀数据规律性的分析服务[3]。天津大学赵怿甦采用模糊评价算法在综合考虑农业环境各污染物因子的基础上按地区实际条件评测环境，并采用 Apriori 算法找出污染物因子之间的相关性，从而挖掘出导致环境差异的具体原因[4]东北农业大学孟军将数据挖掘方法应用到农业生产决策支持系统上，给出了基于数据仓库的农业生产决策支持系统结构[5]等。目前数据挖掘技术在农业数据处理方面基本处于经典挖掘算法在实例应用层次，没形成一套系统的、成体系的研究

① 收稿时间:2011-03-16;收到修改稿时间:2011-04-26

与开发方法。特别是，系统采用紧耦合设计，数据挖掘算法与农业数据是紧耦合的，用户只能使用现有的算法，无法按农业应用需求灵活增添、改造算法。因此，如何设计一个松散耦合的、支持功能模块的重用和算法灵活替换已成为一个值得研究的问题。

由于农业数据具有季节性、地域性、综合性、时效性、多层次性和求新性等复杂特点。本文通过给出一种基于 Web 的农业数据挖掘平台的原理框架，为构建基于 Web 的农业数据挖掘平台提供理论指导；采用责任链模式实现一个基于 Web 的农业数据挖掘平台 agri-Miner，并将其应用到具体的农业领域数据处理取得较好效果。

1 基于Web的农业数据挖掘平台的原理框架

数据挖掘平台的通用性、灵活性和可扩展性是制约其解决应用问题能力的重要因素。现有面向农业领域的数据挖掘平台大多采用较为固定的处理流程，针对不同的应用数据和数据挖掘任务，需要调整和修改应用模块，部署较为困难。因此，寻求最小更改代码对提高面向特定应用领域的数据挖掘平台的灵活性至关重要。针对上述需求，给出了基于 Web 的农业数据挖掘平台的原理框架（如图 1 所示），具体包括“左中右”三层，即表现层、业务逻辑层和数据层。

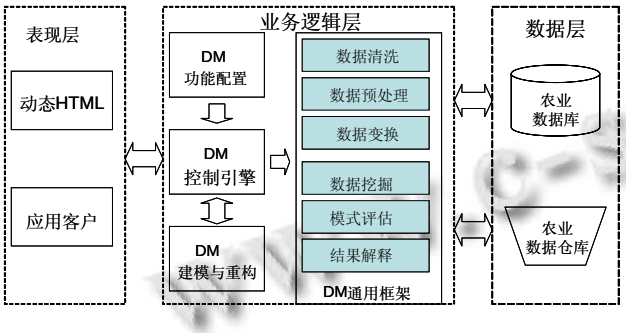

图 1 基于 Web 的农业数据挖掘平台的原理框架

(1) 表现层：为动态和可视化展示数据挖掘的效果，采用动态 HTML 技术和数据可视化技术(Data Visualization)，以远程图形化+Web 方式展示；为实现交互式管理和实时跟踪数据挖掘结果，对其数据挖掘的模式和数据挖掘结果的相互关系可视化。

(2) 业务逻辑层：包括数据挖掘(DM)功能配置、数据挖掘(DM)控制引擎、数据挖掘(DM)建模和重构、数据挖掘(DM)通用框架。数据挖掘通用框架包括数据清洗、数据预处理、数据变换、数据挖掘、模式评估和结果解释等模块，在数据挖掘(DM)控制引擎驱动下动态、有序调度运行，为灵活的数据挖掘提供支持。DM 控制引擎按 DM 功能配置文件描述这些数据挖掘执行流程，因此，只要简单修改数据挖掘(DM)功能配置文件，就可以实现数据挖掘的处理流程的重构，且避免修改源码、编译和部署等工作，从而保证业务逻辑层具有很高的灵活性。

(3) 数据层：农业数据由数据仓库和数据库存储，作为共享数据。数据库的数据存储是按照管理业务中事物处理的要求而存放的；数据仓库的数据存储是按决策分析需求而存放的，存放大量历史数据用于预测。

2 责任链结构实现基于Web的农业数据挖掘平台agri-Miner

为实现基于 Web 的农业数据挖掘平台 agri-Miner，借鉴面向服务开源软件 Apache Axis[6]的体系结构，将独立的数据挖掘模块封装成处理器，由数据挖掘控制引擎驱动一系列处理器，按照对农业数据的挖掘任务寻求动态调整处理器，这样既提升灵活性又保证较高性能。责任链式结构可降低耦合性，提高灵活性。

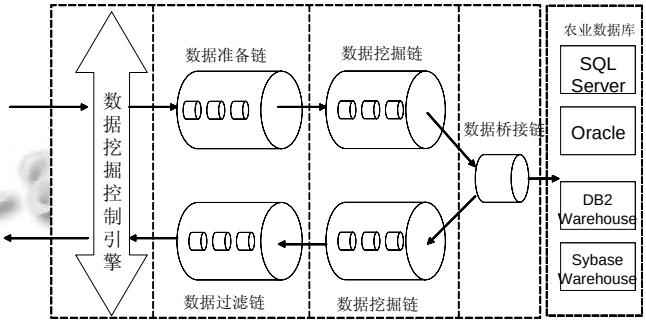

图 2 agri-Miner 的责任链式结构

根据基于 Web 的农业数据挖掘平台的原理框架，agri-Miner 包括请求链(Request Chain)和应答链(Response Chain)，在数据挖掘控制引擎按优先级处理 XML 消息的基础上，依次驱动请求链或应答链。

(1) 数据准备链包括数据抽样、数据清理、数据约简、数据融合和数据变换等处理器，负责将数据变成干净的、规范的、可供数据挖掘模型使用的数据集；

(2) 数据挖掘链主要选择具体的分类、聚类、关联

规则和序列模式发现等处理器对具体的数据对象进行挖掘处理；

(3) 数据桥接链通过数据库桥接器访问后台数据库或数据仓库；

(4) 数据过滤链负责对数据挖掘的结果过滤、整合，并将结果转化为可视化对象最后以远程图形化+Web 方式展示，包括数据挖掘结果过滤器和数据可视化器。

整个链式结构由数据挖掘控制引擎驱动按序来执行，根据数据挖掘的任务以及根据灵活修改数据挖掘流程的配置文件来灵活定制各种处理器，从而保证 agri-Miner 平台应用的灵活性和有效性。图 3 给出关联规则发现处理器(Association Rule Handler)实现示例[7]：

```
public class AprioriHandler extends AbstractAssociator implements CARuleMiner, BasicHandler {
    public invoke() throws Exception {
    double[] confidences, supports;     //置信度，支持度向量
    FastVector[] sortedRuleSet;
    instances = new Instances(instances);
    //如果 m_car 和 m_metricTyle 不等于置信度
    if (m_car && m_metricType != CONFIDENCE)
    getCapabilities().testWithFail(instances);
    //确定下界至少包含一个实例
    double lowerBoundMinSupportToUse=
    m_lowerBoundMinSupport;
    if(m_car){
        m_instances = LabeledItemSet.divide(instances,false);
        m_onlyClass = LabeledItemSet.divide(instances,true); }
    else{ // 增加最小支持度直到找到合适的规则
        m_minSupport = m_upperBoundMinSupport - m_delta;
m_minSupport = (m_minSupport < loweMinSupportToUse)
         ? lowerBoundMinSupportToUse: m_minSupport;}
     // 找到最大项集及其规则
    findLargeItemSets();
if (m_significanceLevel!= -1 || m_metricType != CONFIDENCE) findRulesBruteForce();
    else findRulesQuickly(); }
    // 按照支持度对规则排序
    for (int i = 0; i < m_allTheRules[2].size(); i++)
    supports[i] =
(double)((AprioriItemSet)m_allTheRules[1].elementAt(i)).supportindices = Utils.stableSort(supports);}
```

图 3　关联规则发现处理器实现

3　应用案例研究

农作物病虫害直接危及农产品的产量和质量。二化螟是水稻生产中最重要的虫害之一，严重影响水稻产量和质量。病虫害的发生量主要受温度、湿度和降雨的影响，随着气候条件的变化，其发生量变化呈现出很大的不同。通过收集我国某地区 15 年越冬代二化螟相关数据，采用统计学中层次分析方法或主成份分析方法，选取出四个最重要的影响因子：1 月份平均气温(℃)、1 月份降水量(mm)，4 月份平均气温(℃)、4 月份降水量(mm)[8,9]，具体数据见表 1。

表 1　越冬代二化螟 15 年重要影响因子数据

序号 No.	1月份平均气温/℃ T.Avg Jan.	1月份降水量/mm R. Jan.	4月份平均气温/℃ T.Avg Apr.	4月份降水量/mm R. Apr.	越冬代二化螟量(总量) Chilo suppressalis
1	1.4	31.4	16.1	86.3	99
2	-1.0	32.3	17.6	150.2	132
3	-0.8	73.0	14.3	170.2	232
4	1.4	27.0	15.0	230.0	256
5	-0.9	97.4	14.2	307.1	284
6	1.4	31.1	15.9	234.8	137
7	4.0	29.3	15.9	72.2	38
8	2.7	27.3	17.7	12.1	198
9	3.8	3.5	16.2	94.0	102
10	-0.1	16.5	16.2	150.0	410
11	-0.5	20.9	17.2	87.0	130
12	3.4	2.09	17.0	83.0	128
13	2.2	1.5	17.6	127.7	223
14	1.4	32.5	14.5	232.0	312
15	0.2	80.5	14.6	250.0	267

对 agri-Miner 的数据挖掘流程进行修改，在关联规则发现器(Association Rule Handler)指定 Apriori 算法。经 Apriori 算法进行相关规则发现挖掘得到 1 月份平均温度、1 月份降水量、4 月份平均温度、4 月份降水量和越冬代二化螟的潜在规律，具体的挖掘结果如下：

Best rules found:

(1) Chilo suppressalis='(-inf-131]' 5==> R.Apr.='(71.1-100.6]' 5 conf: (1)

(2) R. Apr.='(71.1-100.6]' 5 ==> Chilo suppressalis='(-inf-131]' 5 conf: (1)

(3) R. Apr.='(218.6-248.1]' 3 ==> T.Avg Jan.='(1-1.5]' 3 conf: (1)

(4) T.Avg Apr.='(-inf-14.55]' 3 ==> Chilo suppressalis='(224-317]' 3 conf: (1)

(5) T.Avg Apr.='(17.35-inf)' 3 ==> Chilo suppres-

salis='(131-224]' 3 conf: (1)

(6) T.Avg Jan.='(3.5-inf)' 2 ==> R. Apr.= '(71.1-100.6]' 2 conf: (1)

(7) T.Avg Jan.='(3.5-inf)' 2 ==> Chilo suppressalis='(-inf-131]' 2 conf: (1)

(8) T.Avg Jan.='(-inf--0.5]' Chilo suppressalis='(224-317]' 2 ==> T.Avg Apr.='(-inf-14.55]' 2 conf: (1)

(9) T.Avg Jan.='(-inf--0.5]' T.Avg Apr.='(-inf-14.55]' 2==> Chilo suppressalis='(224-317]' 2 conf: (1)

(10) R.Jan.='(30.27-39.86]' R. Apr.='(218.6-248.1]' 2 ==>T.Avg Jan.='(1-1.5]' 2 conf: (1)

由以上 10 条规则可以得到如下结论：

(1) 越冬代二化螟发蛾量与气候因子有密切关系。四个气候因子对发蛾量的影响顺序是：4 月份降水量>4 月份平均气温>1 月份平均气温>1 月份降水量。

(2) 越冬代二化螟发蛾量与降水量关系。当 4 月份降水量为 71.1-100.6 mm 时，越冬代二化螟发蛾量<131，说明降水量较少时则虫害量较低；当 1 月份平均温度>3.5℃时，则 4 月份降水量为 71.1-100.6mm、越冬代二化螟发蛾量<131，说明 4 月份降水量对发娥量的影响占主导地位；当 4 月份降水量大于 250mm 时，越冬代二化螟发蛾量受气候变量较小，一般维持在 224-317。

(3) 越冬代二化螟发蛾量与气温关系。当 4 月份平均温度<14.55 或>17.35℃时，越冬代二化螟发蛾量分别为 224-317 和 131-224；当 1 月份平均温度<-0.5℃、4 月份平均温度<14.55℃时，越冬代二化螟发蛾量为 224-317。因此可以看出平均气温低时发娥量较高，平均气温高时发娥量较低[10]。

4 结论

针对现有的数据挖掘平台普遍存在数据挖掘算法与数据紧耦合，体系结构扩展性不理想等问题，本文首先提出一种基于 Web 的农业数据挖掘平台的原理框架用以指导平台开发，并采用责任链模式实现一个基于 Web 的农业数据挖掘平台 agri-Miner，能保证体系结构的可扩展性，数据挖掘算法的灵活替换。最后，成功应用到农业病虫害以验证平台的有效性。

参考文献

1 郑向群,高怀友,周军,王菲,王跃华,赵玉杰.农业环境信息数据分析中数据挖掘技术的应用.农业环境与发展,2003,(1):35–37.

2 郑向群,赵政,刘东生.基于数据仓库的土壤环境监测综合挖掘模型构架.农业工程学报,2008,24(8):162–168.

3 查骏雄.数据挖掘在土壤侵蚀分析系统中的应用.南昌工程学院学报,2005,24(2):46–48.

4 赵怿甦.数据挖掘在农业环境中的应用[硕士学位论文].天津:天津大学,2004.

5 孟军.数据挖掘方法在农业生产决策支持系统上的应用.农业网络信息,2007,(1):63–64.

6 The Axis Development Team. 2009. Axis Architecture Guide. http://xml.apache.org/axis

7 Witten IH, Frank E. 2010. WEKA: Machine Learning Algorithm in Java. http://www.cs.waikato.ac.nz/ml/weka

8 黄光明.Apriori 算法在农业病虫害分析中的应用.安徽农业科学,2009,37(13):6028–6029.

9 程新意,杨崇瑞.用模糊分析方法预报越冬代二化螟的发生量.安徽农学院学报,1992,19(3):308–312.

10 罗凤娥.基于 Web 的农业数据挖掘平台技术研究[硕士学位论文].长沙:湖南农业大学,2010.