基于数据挖掘的植保预测系统

唐超礼[1], 魏圆圆　(1 . 安徽理工大学电气学院, 安徽淮南 232001 ;2 . 中国科学院合肥智能机械研究所, 安徽合肥 230031)

摘要　针对目前病虫害预测预报方法单一、适用面窄, 设计并实现了一个基于数据挖掘的综合植保预测系统。系统结合数据仓库技术, 采用分类、聚类、关联、时序搜索等多种挖掘模式, 对多方面的植保数据进行分析、建模, 达到预测的目的。

关键词　数据挖掘; 数据仓库; 植保; 预测

中图分类号　S126　　**文献标识码**　A　　**文章编号**　0517 - 6611(2008)12 - 05141 - 03

Plant Protection Forecast System Based on Data Collection

TANG Chao-li et al　(Department of Electrical Engineering , Anhui University of Science and Technology , Huainan , Anhui 232001)

Abstract　Aiming at the single method of forecasting plant disease and insect and its poor application ,in the paper we designed and realized an integrated plant protection forecast system based on data collection .The system was combined with the data warehouse technology and multi collection pattern ,such as classification ,clustering ,association ,time-sequence search ,etc .The plant protection data of multiform could be analyzed and then a forecasting report was made .

Key words　Data collection ; Data warehouse ; Plant protection ; Forecast

我国是一个农业大国, 植保工作极为重要[1-3]。对农作物病、虫、草害及时、有效防治是保证作物正常生长发育、获取高产的重要因素。利用计算机技术对病虫草害发生进行预测、识别、诊断是实施有效防治害虫的前提和关键。当前病虫害预测方法比较单一[4], 往往只能对特定形式的数据进行分析操作, 包括一些基于数据挖掘的预测工具, 也只是针对某种类型的数据采用某一种挖掘手段进行预测, 如基于决策树的预测、基于时序的预测等。在植保工作中, 积累了大量病害、虫害、气候、土壤等不同形式的数据, 对这些数据进行有效的描述以及构造一个综合预测平台显然很有意义。

数据挖掘任务一般可分为两类[5-6] : 描述和预测。笔者从数据挖掘的任务出发, 设计并实现了一个基于数据挖掘的植保预测系统, 系统中实现了关键的挖掘算法, 针对农业上不同类型的数据, 经过转换、集成后, 采用可视化技术和特定的挖掘模式, 对数据进行图形展示, 挖掘建模, 以达到对病虫草害等的预测与防治。实验评估表明: 挖掘方法各有所用, 且具有较高的预测质量。

1　数据挖掘

1.1　**数据挖掘的背景和意义**　近年来, 数据库技术得到迅速发展, 许多领域都建立了大型数据库[7], 其中隐藏着许多有价值的信息, 是不可多得的知识信息源, 而目前的数据库系统一般只限于一些基本的数据查询操作, 通过数据库管理系统只能对数据“粗加工”, 不能从这些数据中归纳出隐含的带有结论性的知识[8], 使得这些知识无法利用, 出现了所谓“数据丰富但知识贫乏”的局面, 实际上是对数据库信息资源的一种浪费[6]。因此, 对数据进一步加工和内容分析显得越来越重要。在这样的背景下, 数据仓库、数据挖掘等技术应运而生。

数据挖掘是一类深层次的数据分析[7], 它能从大量数据中抽取出具有一定规律的知识, 深层次的开发可进一步提高信息资源的使用价值, 充分利用信息资源, 提高使用效益。数据挖掘给决策分析带来了新的途径, 能更好地解决日益复杂多变的决策环境问题, 进一步提高了决策的准确性和可靠性, 为科学决策提供了基础。

1.2　**数据挖掘的定义和方法**　数据挖掘是一个从大量数据中抽取挖掘出未知的、有价值的模式或规律等知识的复杂过程。数据挖掘是一个交叉学科领域, 涉及的学科领域和方法包括数据库技术、人工智能、机器学习、神经网络、统计学、知识表示、可视化等。从数据到挖掘出知识的演化过程如下:

图1　数据到知识的演化

Fig .1 From data to knowledge

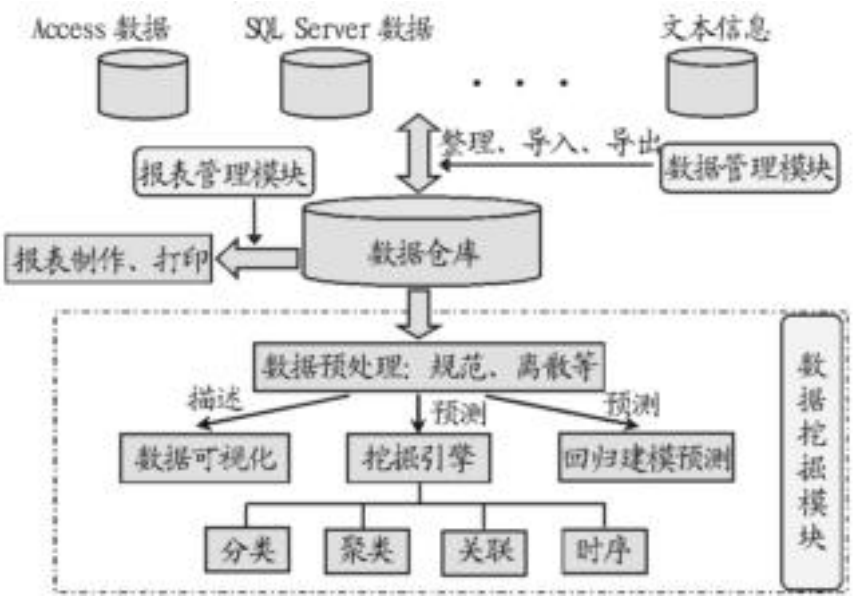

图2　植保预测系统结构框架

Fig .2 The structure frame of plant protection forecasting system

数据挖掘的主要方法模式有分类、聚类、关联分析、时序分析等。

2　基于数据挖掘的植保预测系统结构设计

根据植保工作的特点, 围绕数据仓库技术, 将植保预测系统设计为三大功能模块: 数据管理模块、报表管理模块、数据挖掘模块。数据仓库需要后端工具和实用程序来加载和刷新其数据(dm P56) , 该系统采用 SQL Server 2000 作为后台

基金项目　国家十一五科技支撑计划项目(2006BAD10A14) ; 国家自然科学基金项目(60774096) ; 国家 863 计划项目(2006AA10Z237) ; 安徽理工大学博士、硕士基金项目。

作者简介　唐超礼(1980 -) , 男, 安徽阜阳人, 硕士, 讲师, 从事自然计算、智能控制的研究。

收稿日期　2008-02-18

数据库环境,数据管理模块实现对数据仓库的维护。操作数据库和其他格式的信息源通过数据管理模块,构造数据仓库,数据仓库通过数据挖掘模块进行数据的描述和预测,报表管理实现报表的定制与打印。系统结构框图如图2 所示。

3 系统实现

3.1 数据管理模块 数据库管理分为3 个层次:数据表类别维护、数据表维护和数据维护。①类别维护。对类别进行添加、修改、删除。数据库中数据表进行分类别管理,根据众多数据表的信息特征,将数据表归为不同类别,以便于数据库的管理。②数据表维护。可进行数据表的新建、修改、删除,数据表结构的修改。③数据维护。可对表中记录进行添加、修改、删除,字段的统计,表的横向汇总、纵向汇总等操作。不同类型数据源可通过 ODBC 数据源形式导入到数据库中,数据库中数据表可以导出 RTF、文本、Access、HTML 等多种格式。

3.2 报表管理模块 报表管理进行自由报表表样设计以及报表的预览、打印。表样设计根据用户需求进行表头的自由定制,选择与之对应的数据表,并为报表每一列选择数据表中对应字段,然后进行简单的参数设置,即可对报表进行预览、打印。

3.3 数据挖掘模块 数据挖掘包括数据预处理、数据可视化、挖掘引擎和回归建模预测4 部分。预处理方法有数据的规范化、离散化以及缺损处理,为挖掘的前期操作。数据可视化进行图表展示,能将表中相关字段数据以坐标形式形象表示出来,另外该部分中集成了简单的记录抽样和字段抽样功能,以便记录或字段很多时,只显示出用户关心的数据。挖掘引擎引导用户选择特定的挖掘模式针对符合要求的数据挖掘知识,可对规则知识进行导出。由于植保过程中广泛采用回归方法进行病虫预测,该模块中实现了回归建模、预测以及模型库的维护操作。下面对模块中实现的关键挖掘方法进行介绍。

3.3.1 数据规范化(Data Normalization)。 为适应基于对象距离的挖掘算法,常需要对数据进行规范化处理。实现中采用两种规范化策略:最大- 最小规范化和零- 均值规范化。对于最大- 最小规范方法,要求用户输入合适的区间范围。

3.3.2 数据离散化(Data Discretization)。 用于减少给定连续属性值的个数,对属性进行概念分层,用高层次的概念替代低层次的概念。用户选择需要离散的数值属性,根据实际情况划分离散区间,并为每个区间映射一个区间标号,替代实际的数据值,如“温度”可离散为“高”、“中”、“低”。

“小麦病害”分类结果

分类规则:

if 病斑形状='卵圆形' and 受害部位颜色='褐色' and 发病部位='叶部' and 发生时期='穗期' and 株高表现='正常' then 病害名称='纹枯病'
if 病斑形状='卵圆形' and 受害部位颜色='褐色' and 发病部位='叶部' and 发生时期='返青期' and 株高表现='正常' then 病害名称='纹枯病'
if 病斑形状='卵圆形' and 受害部位颜色='褐色' and 发病部位='叶部' and 发生时期='起身拔节期' and 株高表现='正常' then 病害名称='纹枯病'
if 病斑形状='卵圆形' and 受害部位颜色='褐色' and 发病部位='叶部' and 发生时期='分蘖期' and 株高表现='正常' then 病害名称='纹枯病'
if 病斑形状='卵圆形' and 受害部位颜色='褐色' and 发病部位='叶部' and 发生时期='苗期' and 株高表现='正常' then 病害名称='纹枯病'
if 病斑形状='卵圆形' and 受害部位颜色='褐色' and 发病部位='茎部' and 发生时期='苗期' and 株高表现='正常' then 病害名称='纹枯病'
if 病斑形状='卵圆形' and 受害部位颜色='褐色' and 发病部位='茎部' and 发生时期='穗期' and 株高表现='正常' then 病害名称='纹枯病'
if 病斑形状='卵圆形' and 受害部位颜色='褐色' and 发病部位='茎部' and 发生时期='返青期' and 株高表现='正常' then 病害名称='纹枯病'
if 病斑形状='卵圆形' and 受害部位颜色='褐色' and 发病部位='茎部' and 发生时期='起身拔节期' and 株高表现='正常' then 病害名称='纹枯病'
if 病斑形状='卵圆形' and 受害部位颜色='褐色' and 发病部位='茎部' and 发生时期='分蘖期' then 病害名称='锈病'
if 病斑形状='卵圆形' and 受害部位颜色='褐色' and 发病部位='穗部' then 病害名称='纹枯病'
if 病斑形状='无病斑' and 受害部位颜色='黑色' and 发生时期='返青期' then 病害名称='全蚀病'
if 病斑形状='无病斑' and 受害部位颜色='黑色' and 发生时期='分蘖期' then 病害名称='全蚀病'
if 病斑形状='无病斑' and 受害部位颜色='黑色' and 发生时期='发穗期' then 病害名称='全蚀病'
if 病斑形状='无病斑' and 受害部位颜色='黑色' and 发生时期='起身拔节期' then 病害名称='全蚀病'
if 病斑形状='无病斑' and 受害部位颜色='黑色' and 发生时期='苗期' then 病害名称='全蚀病'
if 病斑形状='无病斑' and 受害部位颜色='黑色' and 发生时期='穗期' then 病害名称='腥黑穗病'
if 病斑形状='卵圆形' and 受害部位颜色='白色' then 病害名称='纹枯病'
if 病斑形状='无病斑' and 受害部位颜色='红色' then 病害名称='赤霉病'
if 病斑形状='圆形' then 病害名称='白粉病'
if 病斑形状='条纹状' then 病害名称='霜霉病'

测试属性:发生时期 发病部位 株高表现 病斑形状 受害部位颜色
类别属性:病害名称
训练比例:90%
训练样本个数:74
测试样本个数:9
分类规则对测试样本预测的正确率:0.888888888888889

图3 “小麦病害”的分类规则及评估结果

Fig.3 The classification rule and evaluation results of “wheat disease”

3.3.3 数据缺损处理(Blank Filling)。 对含有缺损数据的数据表,选择缺损字段进行填充,系统中提供4 种填充方式:以平均值填充、以最大值填充、以最小值填充和指定值填充。

3.3.4 分类(Classification)。 通常数据挖掘中,将预测离散无序类别值的数据归纳方法称为分类方法。系统中实现了基于决策树归纳方法的分类模式。通过候选测试属性(判定属性)和类别属性构造一棵决策树,得到的分类模型以分类规则(IF-THEN)形式加以描述输出,用于预测未知数据实例的归属类别。采用常用的保持(holdout)方法对模型分类准确率进行估计,根据训练比例,将样本集分为训练样本集和测试样本集,训练样本用于得到决策树,测试样本用于计算分类规则的准确率。要求所有候选属性和类别属性都为离散属性,连续值必须离散化。以某地区小麦病害为例,分类和评估结果如图3 所示。

3.3.5 聚类(Clustering)。聚类是一个将数据集划分为若干组或类的过程,并使得同组内的数据对象具有较高的相似度,不同组中的数据相似度尽可能小。与分类不同,聚类分析在归类预测时所分析处理的数据均是无类别归属,是一种无教师监督学习方法。这里采用k-means 算法实现聚类,距离采用欧氏距离,聚类的结果由参与聚类的属性(数值属性)决定,新数据实例通过计算其与各聚合中心的距离归为距离最小的组。

3.3.6 关联(Association)。关联规则挖掘就是从给定的数据集中搜索数据项之间存在的有价值联系。以病虫害诊断或预测为例,根据病虫害某些症状特征的有无,推断其他症状或病害存在的可能性,或通过某一种病害的有无,推断症状出现的可能性。采用 Apriori 算法为布尔数据表产生关联规则,用户确定最小支持度和最小置信度,挖掘过程分为得到频繁项目集和得到关联规则两个过程进行。每条规则都匹配一个置信度,表明根据该规则进行预测的可信程度。以布尔数据"事务数据表"为例,选择最小支持度20%、最小置信度60%,关联分析结果如图4 所示。

3.3.7 时间序列相似搜索(Similarity Search on Time Sequences)。农业上数据常常和时间紧密相连,针对时间序列数据库,输入查询序列(当前几天的虫害资料),在历史数据中搜索与其最匹配的序列,根据历史上相似情况下接下来几天的虫害资料,预测当前可能会出现的病虫害指标,并制订相应的预防措施,达到对病虫害的预测与防治目的。采用文献[9]中搜索策略,描述如下:①建立检索结构。采用滑动窗口(长度 w)将多维时间序列(长度为 n)分成 n - w + 1 个子序列,且从每个子序列提取形状特征向量 SF 和最小边界方形 MBR,并根据 SF 构造一棵 k-d 树。②查询处理。计算查询序列的 SF 和 MBR,通过搜索 k-d 树找到与查询序列外形相似的子序列,接着比较这些子序列和查询序列的 MBR,若某一子序列的 MBR 与查询序列的 MBR 相差满足误差要求,则进一步比较它们的实际距离(欧氏距离)。

3.3.8 回归。对连续数据的预测通常利用统计回归方法所建的模型来实现。系统中实现了基于最小二乘法的线性回归,选择要建模的数据表,在此基础上选择预测变量和响应变量,生成样本空间,计算回归系数,得到一元或多元线性回归模型,最后假设检验,保存模型,根据该模型输入预测变量值,对响应变量进行预测。

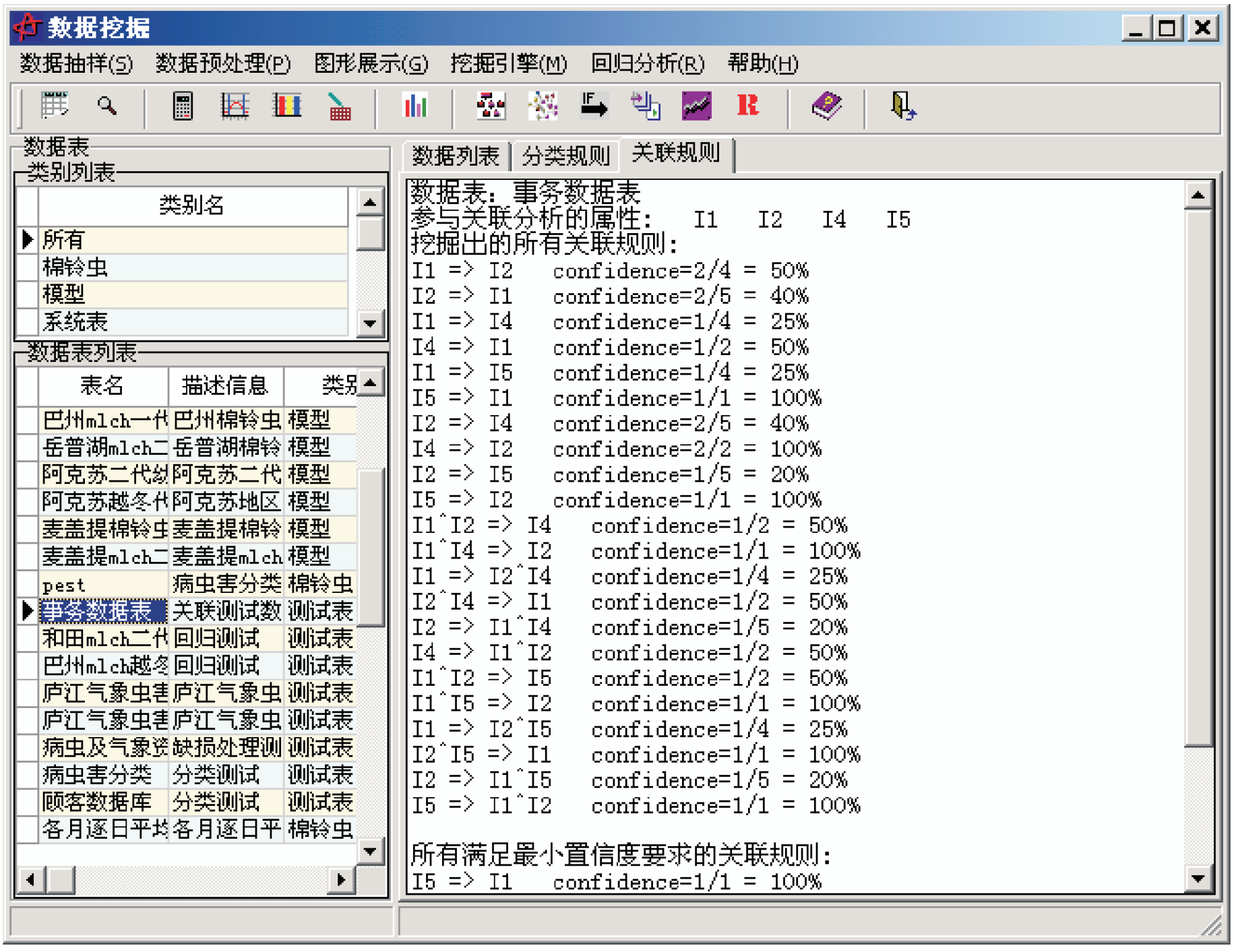

图4　关联分析得到的结果

Fig.4　The results of correlation analysis

4　结束语

该文结合植保工作的任务和特点,介绍了一个基于数据挖掘的植保预测系统的设计及其关键技术的实现。本地化数据通过数据管理模块装入数据仓库,基于数据仓库进行数据描述、知识挖掘,最后达到预测预报的效果。系统集数据管理、报表制作及预测预报于一体,为植保工作提供了一个相对完备的应用工具。该系统运用于新疆植保系统取得了

(下转第5202 页)

业全面发展,才能保证国民财富的增长。并且从中可看出,对各行业已有明确的安排和管理,已经有意识地运用农业结构调整的方式方法进行农业的生产了。

2.2 土地资源管理思想 农业生产离不开土地,对土地资源的管理和合理运用是保证农业持续发展、提高生产效益的关键因素。因此,几千年来我国学者对土地利用的研究一直未间断。

2.2.1 合理开发土地资源,反对滥用地力。我国古代儒家主张开发利用土地资源应合理有度,并且土地资源应受国家法律保护,不得随意开垦,如"善战者服上刑,连诸侯者次之,辟草莱、任土地者次之"[5]。"辟草莱,任土地",这种随意开发土地的行为,可以与"善战"、"连诸侯"这种重罪相提并论。说明当时合理开发土地资源已经是一项重要的农业法律法规,人们对土地的开垦已经有章可循,有法可依了。这样充分保障了土地资源的高利用率,达到了可持续利用的目的。《周礼》在土地制度中把土地分为上、中、下三等,按质量将土地授予劳动者耕种。"不易之地家百亩,一易之地家二百亩,再易之地家三百亩"[6]。从中可看出,我国早在先秦时期,对于土地的分配利用已有了初步的认识。

2.2.2 休养地力,因地制宜。从《禹贡》中可以看出,当时的人们已认识到不同地质的土壤,决定了其适宜生长的农作物种类,并且耕作与管理的方法也大不相同。"巡邦野之稼,而辨种之种,周知其名,与其所宜地,以为法而县于邑间"[6]。"观地宜…使五谷桑麻皆安其处。"[7] 这样做"辨于地利",从而达到"民可富"的目的。

农业生产有很强的季节性,我国古代的思想家、政治家都很强调农时,冬季是不适宜农作物生长的时期,应使地力得到休养。《周礼》中,把土地分为上地、中地、下地三级,并规定上地可年年耕种,中地耕种一年需休耕一年,下地耕种一年后需要休耕二年。这是维护地力的具体措施,并为后世沿用。另外,早在战国时代就已提出使用精耕细作的方式,提高耕地的单位面积产量。"今是土之生五谷也,人善治之,则亩益数盆。"[7] 说明对土地进行集约利用有利于人均产量的提高;从中还可看出,当时人们已认识到土地经由"人善治之"可取得意想不到的效果,肯定了人力维护、利用土地资源的作用。

2.3 水资源管理思想 水是保障农业生产顺利进行的重要条件,但由于季节不同,水的流量和对水的需求量也不同,因此农业生产中水资源的保护利用显得非常重要,不仅要重视保护已有的水资源,还应加强对农田水利设施的兴修及管理。如荀子主张大兴水利兴修及管理,"修堤梁,通沟浍,行水潦安水藏,以时决塞,岁虽凶败水旱,使民有所耕耘"[7]。并且有专门负责管理水利设施修建的官员——司空。司空每年春季之日要"循行国邑,周视原野,修利堤防,道达沟渎,开拓道路,毋有障塞。"[8] 可见,早在我国先秦时期,对水利的管理就已经有一套系统的措施,目的是为了防御水旱。《周礼》中更是对农田水利的修造进行了详细的阐述,"凡治野,夫间有遂,遂上有径;十夫有沟,沟上有畛;百夫有洫,洫上有涂;千夫有浍,浍上有道;万夫有川,川上有路,以达于畿。"[6] 从这段话中可看出,在当时我国的水利设施已有网络式的沟通渠道,与地界、田埂、道路都有相应的配合,非常规律,对水利的管理、利用已达到较高的水平,并在当时发挥了极重要的作用。而自此以后,我国各朝代都将水利设施的兴修提到重要日程上,以此来防御水旱灾害,保证农业生产。

2.4 山林资源管理思想 我国古代对自然资源的利用有一定的原则,本着节约的目的有节制地利用,国家对林业资源、渔业资源、野生动物资源都实行了保护政策。对山林川泽资源的开发和利用,一要反对滥砍滥伐,履行"时禁"保护树木的正常生长;二要防火灾,保证森林资源不受毁坏。因为汉代以后,人们已认识到森林对于水旱灾害起到了重要的防护作用。另外,人们按照野生动物、鱼类的生长繁衍规律,规定人们捕捞的时间。"毋杀畜生,毋拊卵,毋伐木,毋夭英,毋折竿,所以息百长也。"[4] 在春季草木萌发时,动物孵育季节禁止砍伐、捕捞,以保证自然界生物生生不息的生长繁殖。

3 小结

我国古代的自然资源丰富,人们在对其利用的过程中体现出了一定的农业生态管理思想,这种理念贯穿了我国传统农业生产的整个时期,对我国农业生产的具体管理措施起到了指导作用。通过以上论证可知,我国古代农业生产合理安排了各生产要素间的关系,使得天、地、人、物间相互协调,统一发展,保证了我国自然资源的持续利用和农业的稳定发展,已将生态思想融入到农业思想中,影响了我国传统农业的生产实践,引导了生态农业管理思想几千年的发展进程。

参考文献

[1] 邹德秀. 绿色的哲理[M]. 北京: 农业出版社,1990:7.
[2] 王征兵. 中国农业经营方式研究[M]. 北京: 科学文化出版社,2001:7.
[3] 张磊. 中国传统农业文化转型研究[M]. 西安: 陕西人民出版社,2007:67.
[4] 刘向. 管子[M]. 桂林: 广西师范大学出版社,2005.
[5] 孟子. 孟子[M]. 合肥: 安徽人民出版社,2002.
[6] 郑玄. 周礼注疏[M]. 北京: 中华书局,1980.
[7] 荀况. 荀子[M]. 西安: 山西古籍出版社,2004.
[8] 戴圣. 礼纪[M]. 郑州: 中州古籍出版社,1993.

(上接第5143页)

很好的效果。

参考文献

[1] 沈光斌, 刘家成, 郑兆阳, 等. 开展病虫电视预报创新植保服务模式[J]. 中国植保导刊,2007,27(9):35-37.
[2] 耿爱军, 李法德, 李陆星. 国内外植保机械及植保技术研究现状[J]. 农机化研究,2007(4):189-191.
[3] 鞠瑞亭, 严巍, 池杏珍, 等. 数据库技术在园林植保中的应用现状及趋势[J]. 中国森林病虫,2005,24(5):31-34.
[4] 吕昭智, 沈佐锐, 田长彦. 棉蚜预测预报网络数据库系统的设计与开发[J]. 植物保护学报,2003,30(3):237-242.
[5] HAN J W, MICHELINE K. Data Mining: Concepts and Techniques[M]. New York: Morgan Kaufmann Publishers, Inc,2001.
[6] 朱明. 数据挖掘[M]. 合肥: 中国科学技术大学出版社,2002.
[7] 陈京民. 数据仓库与数据挖掘技术[M]. 北京: 电子工业出版社,2002.
[8] 史忠植. 知识发现[M]. 北京: 清华大学出版社,2002.
[9] 黄河. 时间序列相似搜索及其在农业上的应用[D]. 合肥: 中科院合肥智能所,2003.