数据挖掘技术在农产品质量安全监管中的应用

区晶莹,简 荣,俞守华　(华南农业大学,广东广州 510642)

摘要　数据挖掘技术可以在大量的农产品质量安全监管数据中提取有效的信息为政府监管决策服务。笔者以蔬菜批发市场为实例,在分析农产品质量安全监管数据特点的基础上,进行了可视化技术与关联规则的数据挖掘分析。结果表明:数据挖掘技术在农产品质量安全监管中的应用是可行的。

关键词　数据挖掘技术;农产品质量安全监管;数据可视化技术;关联规则
中图分类号　S11　　**文献标识码**　A　　**文章编号**　0517 - 6611(2009)32 - 16190 - 03

Application of Data Mining Technology in the Quality and Safety Monitoring of Agricultural Products

OU Jing-ying et al　(South China Agriculture University, Guangzhou, Guangdong 510642)

Abstract　Data mining technique can extract some useful information from a large number of agricultural products quality and safety data for the governmental supervision and making decision. Taking the vegetable wholesale market for an example, based on analyzing the characteristics of agricultural products' quality and safety monitoring data, the data mining analysis was carried on the and association rules. The results showed that the data mining technology used in the quality and safety monitoring of agricultural products was feasible.

Key words　Data mining technology; Agricultural products' quality and safety monitoring; Data visualization technology; Association rule

农产品质量安全问题是落实科学发展观,促进现代农业和新农村建设的重大问题。确保农产品质量安全关系到人民群众的身体健康和生命安全,关系到农民增收和社会稳定,关系到农产品市场竞争力和社会主义新农村建设,已经成为社会进步的重要指标,也是社会稳定的关键因素。我国各地依据《农产品质量安全法》的要求,逐步建立了从农田到餐桌全过程的农产品质量安全监管,在这些监管过程中产生了大量的农产品质量安全方面的数据,如何从这些数据中提取有效的信息为政府监管决策服务是迫切需解决的问题。为此,尝试利用数据挖掘技术,以批发市场蔬菜农药残留检测数据为实例进行挖掘分析,探讨数据挖掘技术在农产品质量安全监管中应用的可行性,为政府加强监管工作,提高监管效能,提供科学决策依据。

1　批发市场农产品质量安全监管数据的特点

目前农产品批发市场对农产品质量安全监管的主要手段是利用酶抑制率法进行农药残留检测,其检测对象为有机磷及氨基甲酸酯类等毒性农药[1]。测定出抑制率在0~0.2为优秀产品;在0.4~0.5为可疑农药残留超标产品;大于0.5则为农药残留超标产品。经分析批发市场农产品质量安全监管数据有如下特点:

1.1　数据存储方式逐渐规范化　我国《农产品质量安全法》第6章第34条规定:县级以上人民政府农业行政主管部门应当按照保障农产品质量安全的要求,制定并组织实施农产品质量安全监测计划,对生产中或者市场上销售的农产品进行监督抽查。监督抽查结果由国务院农业行政主管部门或者省、自治区、直辖市人民政府农业行政主管部门按照权限予以公布[2]。为此,国家农业部和各地农产品检测部门会定期发布农产品质量安全检测数据,各地主要农产品批发市场则会将每日农产品农药残留检测结果公布在企业网站上。因此,大量规范化的农产品质量安全监管电子数据,为数据挖掘技术的应用提供了基础。

1.2　数据间有较强的关联性　如某批发市场蔬菜农药残留检测部分数据如表1所示。

表1　某批发市场农残部分检测结果

Table 1　Some testing results for the pesticide residue of a wholesale market

名称 Name	生产地 Producing area	抑制率//% Inhibition rate	检测结果 Testing results	日期 Date
小白菜	大圩	11.10	合格	2007.12.17
白萝卜	舒城	10.20	合格	2007.12.17
西兰花	舒城	7.40	合格	2007.12.17
薄皮椒	山东	9.40	合格	2007.12.17
黄瓜	山东	4.80	合格	2007.12.17

由于每种农产品的生长都有其固有的生长周期,以及相应的来源地,因此数据属性之间有较强的关联性。在数据预处理阶段,可将具体日期转化为每个月上中下旬,将抑制率转化为各水平(如高抑制率、低抑制率等),利用关联规则的挖掘,可以得知不同地方、不同日期产出的农产品农药残留抑制率的变化规律,从而对高风险时期、高风险地区进行重点监控,从而达到有效监管目的。

1.3　数据维度较低　与企业中的数据仓库相比,农产品质量安全检测数据库表维度比较低,内容比较单一,这使得在构建农产品质量安全监管数据仓库时要注意ETL过程,即数据抽取(Extract)、转换(Transform)、装载(Load)的过程,尽可能从现有属性中抽取出更多的隐含信息。如表1只有5个维度,然而可以通过数据衍生的方法得到更多的信息,考虑到各种蔬菜的生长季节不一样,故不同时间上市的蔬菜被检测出的农药残留抑制率有明显的时间序列特征。为此,可以将日期进一步划分为"月上旬,月中旬,月下旬",以观察某一品种的蔬菜在每月不同时间上市的抑制率波动情况,为农产品质量监管部门重点监控提供科学依据。

2　数据挖掘技术在农产品质量安全监管中的应用

2.1　农产品质量安全监管数据来源　数据来自于互联网上安徽省国家重点龙头蔬菜批发市场按国家规定公布的农药残留检测数据。时间为2007年3月1日至2007年12月17日,每2天公布1次,每次抽取10个品种进行检测,共计146

作者简介　区晶莹(1964 -),女,广东佛山人,硕士,副研究员,从事管理科学与工程研究。
收稿日期　2009-07-20

天,1 460 项事务。数据利用 Access 建立数据库,导入至数据挖掘软件 SPSS Clementine 中进行数据挖掘分析。

2.2　利用数据可视化技术把握数据总体特征与趋向

2.2.1　数据可视化技术原理。数据可视化技术指的是运用计算机图形学和图像处理技术,将数据转换为图形或图像在屏幕上显示出来,并进行交互处理的理论、方法和技术[3],这涉及到计算机图形学、图像处理、计算机辅助设计、计算机视觉以及人机交互技术等多个领域。通过可视化技术,可以利用图像、曲线、二维图形、三维体和动画来显示数据,直观地表达出对象或事件数据的多个属性或变量,并按其每一维的值,将其分类、排序、组合和显示,并可对其模式和相互关系进行可视化分析。因此利用数据可视化技术有利于批发市场决策者快速地从总体宏观上掌握农产品质量安全监管数据的总体特征与趋向,从而进行有针对性的管理决策。

2.2.2　数据可视化技术具体应用。①对农产品的农药残留检测抑制率分布进行可视化,得出图 1 的结果。由图 1 得知,该批发市场蔬菜产品主要来自安徽、本地、长丰等 21 个省市县地区,其中本地(合肥)、昆明、山东、舒城 4 地供应量最多。从总体抑制率方面相比较,本地(合肥)主要集中在 0.2 ~ 0.4 和 0.4 ~ 0.6 两个区间,山东主要集中在 0 ~ 0.4 区间,这说明本地(合肥)供应的蔬菜农药残留检测抑制率指标比山东供应的蔬菜农药残留检测抑制率高。舒城也大量集中在 0.2 ~ 0.4 区间,总体比山东高,但比本地低。这说明 3 地比较,山东蔬菜能较好控制农药残留,舒城次之,而本地(合肥)对于农药残留的控制则不如前两者。作为本地农产品质量监管部门应加强监控本地蔬菜种植时使用农药的剂量或改进本地蔬菜种植方式。②对产品的来源地分布进行可视化。由于该蔬菜批发市场的产品来自全国各地,各地的种植方法、种植人员文化水平、土壤、自然环境等各不相同,这样有可能导致不同来源地的同一品种蔬菜的农药残留率不同。为此监管部门可以从产品-产地关联图中快速知道产品主要来源地,更好地做好产地追溯工作。③对产品的各个抑制率水平进行可视化,可得出如图 2 的结果。由计数图 2 可以从总体上把握某一段时期里该市场总体农药残留检测状况。如图 2 所示,0 ~ 0.4 这个区间的计数明显高于其他区间,但 0.4 ~ 0.6 的"可疑农残超标样品"仍然占据相当数量的计数。因此,可以认为在数据统计的时间范围内的蔬菜质量安全总体是合格的,但仍有部分品种的农药残留指标不理想,需要通过加强检测监管的手段对这些地区的农产品重点监控。

2.3　利用关联规则掌握不合格产品的特征与模式

2.3.1　关联规则原理。利用关联规则可以从农产品检测数据库的大量事务中检测出蕴含在数据中的一些特定的模式,挖掘出形如"名称 = A and 产地 = B→检测结果 = 不合格"的规则,其中"名称 = A and 产地 = B→检测结果"称为规则前项,"检测结果 = 不合格"称为规则后项。通过了解数据中的规则而掌握农药残留指标不合格的产品模式,进而为监管决策进行服务。如建立主要农产品质量监管检测信息数据库和农产品质量安全监测、预报和预警系统[4],利用关联规则产生的产品模式可以加入到这些系统中的知识库里。

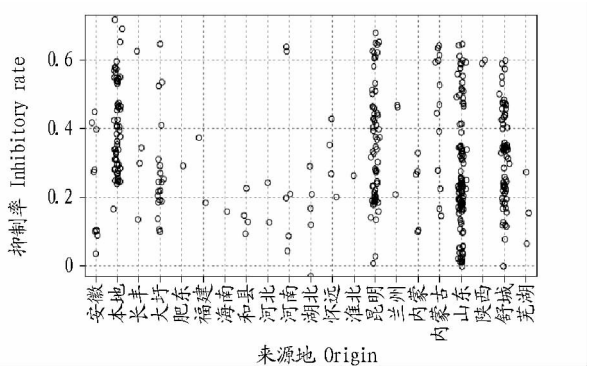

图 1　某蔬菜批发市场农药残留检测抑制率分布图

Fig. 1　The distribution of the inhibitory rate of agricultural products pesticide residues in a vegetable supermarket

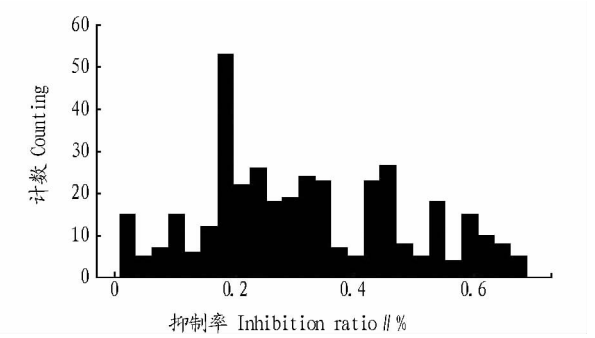

图 2　各抑制率水平计数图

Fig. 2　Counting of various inhibition ratio levels

设 $I=\{i_1,i_2,\cdots,i_m\}$ 是数据库中项的集合,D 是数据库事务的集合,A 是一个前项,B 是一个后项。设 S 是设定的最小支持度,也就是数据 D 中包含 A∪B 事务的百分比,即 $S_{A\cup B}\geqslant S$,计算公式为:

$$\sup port(A\cup B)=\frac{|A\cup B|}{|D|} \tag{1}$$

其中|A∪B|是出现 A 或 B 的事务数,|D|是 D 的事务数。

设 C 是设定的最小置信度,也就是数据 D 中任何一个包含 A∩B 的事务的百分比规则,即 $C_{A\cup B}\geqslant C$,计算公式为:

$$confident(A\cap B)=\frac{|A\cap B|}{|A|}\cdots\cdots \tag{2}$$

|A∩B|是同时出现 A 与 B 的事务数,|A|是出现 A 的事务数。

同时满足最小支持度和最小置信度的规则称作强规则。

关联规则数据挖掘一般可以分为以下两步进行,①找出所有频繁项集:根据定义,这些项集出现的频繁性至少和预定义的最小支持度一样。②由频繁项集产生强关联规则:根据定义,这些规则必须满足最小支持度和最小置信度[5-6]。

2.3.2　利用关联规则掌握不合格产品的特征与模式。利用关联规则对某蔬菜批发市场的农药残留检测数据进行分析,考虑到农产品质量安全事故带来的危害性,将支持度分别设置为 0.3%(即一个事务前项在该数据库中发生 4 次,就为规则候选对象)0.4%、0.5%,置信度设置为 50%(即一个事务后项在候选对象中发生概率是 50%,就为强规则),对 1 460 项事务进行数据挖掘分析,可以得出如下有意义的结果,见表 2、3、4 所示。

通过以上分析可知,设置不同的支持度可以得到不同风

险级别的农产品质量安全情况。在上述 1 460 项事务数据中,菠菜属高风险产品,要重点监管;其次是内蒙古下旬生产的产品和本地中旬生产的豆角;再次是山东产的豆角和昆明产的菠菜。

表 2 支持度设置为 0.3%时数据挖掘结果

Table 2 The data mining results with the minimum support degree of 0.3%

后项 Consequent item	前项 Antecedent item	支持度//% Support degree	置信度//% Confidence degree
检测结果 = 不合格	名称 = 菠菜 and 生产地 = 昆明	0.342	80
检测结果 = 不合格	名称 = 豆角 and 生产地 = 山东	0.342	80
检测结果 = 不合格	名称 = 菠菜	0.548	50
检测结果 = 不合格	生产地 = 内蒙古 and 时间 = 下旬	0.411	50
检测结果 = 不合格	名称 = 豆角 and 生产地 = 本地 and 时间 = 中旬	0.411	50

表 3 支持度设置为 0.4%时数据挖掘结果

Table 3 The data mining result with the minimum support degree of 0.4%

后项 Consequent item	前项 Antecedent item	支持度//% Support degree	置信度//% Confidence degree
检测结果 = 不合格	名称 = 菠菜	0.548	50
检测结果 = 不合格	生产地 = 内蒙古 and 时间 = 下旬	0.411	50
检测结果 = 不合格	名称 = 豆角 and 生产地 = 本地 and 时间 = 中旬	0.411	50

表 4 支持度设置为 0.5%时数据挖掘结果

Table 4 The mining result with the minimum support degree of 0.5%

后项 Consequent item	前项 Antecedent item	支持度//% Support degree	置信度//% Confidence degree
检测结果 = 不合格	名称 = 菠菜	0.548	50

3 结论与讨论

(1)利用可视化技术,可以快速形象地掌握批发市场中农产品农药残留抑制率检测的总体特征,从而了解农产品质量安全总体状况。并且可以及时掌握各品种农产品的来源地,在发生质量安全事故的时候,可以有效地支持追溯调查工作。

(2)利用关联规则,可以通过抑制率检测的历史数据掌握不合格产品的风险程度,为监管工作提供有力的量化决策支持,为农产品质量安全监测、风险管理和预警系统进行数据分析和建立监管数据库提供了技术支持。

(3)以某蔬菜批发市场的 1 460 项事务数据为实例进行了可视化技术与关联规则的数据挖掘分析,从分析结果来看数据挖掘技术应用在农产品质量安全监管工作中是可行的和有效的,但在实际应用中还需要足够多的数据和进一步提高数据的维度,如增加生产地的天气状况、种植规模、运输路途等,并且规范监管工作中的数据录入、存储、上报等工作,这样才能提供全面性、及时性和有效性的数据,才能更有效地为农产品质量安全监管决策服务。

参考文献

[1] 郭维胜,赵作朋,王风洲,等. 酶抑制率法检测农药残留技术[J]. 北京农业,2006(7):43 -44.

[2] 第十届全国人民代表大会常务委员会. 中华人民共和国农产品质量安全法[EB/OL]. (2006-04-30) http://www.agri.gov.en/zefg/nyfl/t20060430_604147.htm.

[3] 王衍. 基于信息可视化技术的税务决策支持系统分析[J]. 数量经济技术经济研究,2004(4):148 -153.

[4] 吴广红. 如何有效加强农产品质量安全监管[J]. 中国质量技术监督,2008(2):54.

[5] 秦国锋,李启炎. 基于数据挖掘的知识获取与发现[J]. 计算机工程,2003(3):206 -208.

[6] 叶瑾,周瑞凌,谢康林. 关联规则数据挖掘方法的改进和实现[J]. 小型微型计算机系统,2002,23(3):347 -349.

(上接第 16186 页)

[22] 仰榴青,茆广华,吴向阳,等. 银杏外种皮多糖的免疫活性研究[J]. 时珍国医国药,2009,20(4):872 -873.

[23] 许爱华,陈华圣,王玲,等. 银杏外种皮多糖对不同状态小鼠血清 SOD 和 MDA 形成的影响[J]. 中国中药杂志,1998,23(12):7446 -747.

[24] 许爱华,王玲,陈华圣,等. 银杏外种皮多糖延缓小鼠衰老的实验研究[J]. 中药材,1996,19(9):466 -468.

[25] 许爱华,王玲,李永华,等. 银杏外种皮多糖延缓荷瘤小鼠衰老的实验研究[J]. 辽宁中医杂志,1997,24(9):429 -430.

[26] 费文勇,彭爱军,王爱萍,等. 银杏外种皮多糖拮抗 D-半乳糖致小鼠衰老作用的实验研究[J]. 辽宁中医学院学报,2004,6(1):56 -57.

[27] 王爱萍,史明仪,费文勇,等. 补充银杏外种皮多糖对 D-半乳糖致衰老小鼠运动能力的影响[J]. 中国运动医学杂志,2004(6):695 -697.

[28] 彭爱军,王爱萍,费文勇,等. 银杏外种皮多糖对衰老模型小鼠学习记忆能力及脑内酶系活力的影响[J]. 中国行为医学科学,2004,13(2):136 -137.

[29] 许爱华,陈华圣,褚澄,等. 银杏外种皮多糖对人癌细胞株的抑制作用及与阿霉素的协同效应[J]. 中国新药杂志,2000,9(11):753 -755.

[30] 许爱华,陈华圣,孙步蟾. 银杏外种皮多糖对 HL-60 细胞的体外实验研究[J]. 中药材,2004,27(5):361 -363.

[31] 许爱华,贾筱琴,陈华圣,等. 银杏外种皮多糖抑制小鼠肝癌及诱导肝癌细胞凋亡的研究[J]. 中药新药与临床药理,2001,12(5):340 -341,375.

[32] 许爱华,陈华圣,孙步蟾,等. 银杏外种皮多糖对人胃癌细胞凋亡及其凋亡诱导基因表达的影响[J]. 中国药理与临床,200319(3):11 -13.

[33] 许爱华,陈华圣,陈钢,等. 银杏外种皮多糖对 SGC-7901 细胞 p53 基因的表达及端粒酶活性的影响[J]. 中国药理学通报,2003,19(10):1174 -1176.

[34] 许爱华,褚云飞,陈华圣,等. 银杏外种皮多糖对胃癌的临床及超微结构研究[J]. 中国新药杂志,2002,11(9):724 -726.

[35] 翟范,陈华圣. 银杏外种皮多糖制剂治疗中晚期癌症 84 例[J]. 辽宁中医杂志,2002,29(9):564.

[36] 陈华圣,翟范,褚云飞,等. 银杏外种皮多糖胶囊制剂治疗中晚期上消化道恶性肿瘤的临床研究[J]. 中西医结合学报,2003(3):189 -191.