数据挖掘技术在食品检测数据中的探索

汪雪君，沈怡，杨慧元（上海市食品药品检验所，上海 201203）

摘要 目的： 将数据挖掘技术应用到食品安全检测数据中。**方法：** 首先，通过对食品安全检测的原始数据进行选择、清洗、转换和分类等预处理，转换成数据挖掘方法所需要的数据格式。然后，结合不同的数据挖掘方法，研究食品安全在时间、地区、种类等的关联性，以及在时间上的时序性。**结果与结论：** 通过数据挖掘技术，探索食品安全环节潜在的风险点。

关键词： 关联性；时序性；食品安全；数据挖掘

中图分类号： R155 **文献标识码：** A **文章编号：** 1002-7777(2019)03-0259-04
doi:10.16153/j.1002-7777.2019.03.004

Exploration of Data Mining Technology in Data of Food Control

Wang Xuejun, Shen Yi, Yang Huiyuan (Shanghai Institute of Food and Drug Control, Shanghai 201203, China)

Abstract Objective: This paper applied data mining technology to the data of food control. **Methods:** First of all, the original data of food safety control were pretreated through the methods of selection, cleaning, conversion and classification and were converted into the required data format of data mining. Then, association of food safety with time, area, types, as well as time sequence were analyzed by different data mining methods. **Results and Conclusion:** Potential risks of food safety were explored by data mining technology in data of food control.

Keywords: association; time sequence; food safety; data mining

食品的安全问题关系到全人类的生活、生存、繁衍，是人类发展的一个重要课题。如今食品安全已是我国消费者的“心头大患”，成为人们最普遍关心的一大主题，因此，防范食品安全风险、加强食品安全控制已成为食品药品监督管理部门的重要职责。食品检验检测也成了食品安全监管部门的重要工作。食品安全检验机构已累积了数年的检验检测数据，并且，每日都有大量的食品检验检测数据产生。这些检验数据中隐含了巨大的、潜在的、可利用的食品安全风险信息，如何从这些数据中提取可用的信息，为食品安全监管提供预警，是一项重要的工作。目前，许多数理统计方法用于对食品安全监测数据的分析，如控制图、移动平均线、线性回归等方法[1]。然而，食品的安全性是个比较复杂的指标，牵涉到的因素非常多，传统的数理统计分析方法已不能完全满足对食品安全检测数据的深度分析需求，在充分利用已有的食品安全检测数据来评估和预测未来的风险趋势上有所欠缺，更迫切需要利用数据分析工具从这些海量数据中提取出具有指导意义的“法则”，因此，需要引进更先进的数据分析技术[2]。

数据挖掘技术就是一种数据分析的新技术，它可从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中人们事先不知道的，但又是潜在有用的信息或知识[3]。数据挖掘常用的方法主要有关联规则[4]、神经网络[5]、决策树[6]、时序等。采用数据挖掘技术可以从大量的食品安全检测数据中提取隐含在其中的潜在的有用信息，实现

作者简介：汪雪君，硕士；Tel：（021）50798187；E-mail：wangxuejun@sifdc.org.cn

对食品安全检测数据的深度分析。比如数据挖掘中的关联规则挖掘能够发现大量数据中项集之间的相关联系，适用于食品安全检测数据的分析。

本文将涵盖食品安全检测数据特点以及数据挖掘技术，通过对食品安全检测的原始数据进行选择、清洗、转换和分类等预处理，构建食品安全数据仓库，并结合数据挖掘算法、食品安全评价指标体系等综合因素，摸索研究上海地区食品安全在时间、空间、种类等的关联性，建立食品安全监测模型，最终寻找上海地区食品安全环节潜在的风险点。

1 数据挖掘方法

1.1 关联规则定义

关联模型基于包含各事例的标识符及各事例所包含项的标识符的数据集生成。事例中的一组项称为“项集”。关联模型由一系列项集和说明组成。简单地说，关联规则可以用这样的方式来表示：A→B，其中A被称为前提或者左部（LHS），而B被称为结果或者右部（RHS）。

关联规则有几个重要的参数：1）支持度(Support)：支持度用来度量一个项集的出现频率。项集{A，B}的支持度是同时包含A和B的事务的总个数。2）概率(Probability)：也叫置信度（Confidence）；规则A=>B的概率是使用{A}的支持度除项集{A，B}的支持度来计算。3）重要性(Importance)：重要性可以用于度量项集和规则；数值越大，表示的规则的重要性越高。

1.2 时序算法定义

时序模型根据用于创建该模型的原始数据集就可以预测趋势，如图1所示。

1）历史信息显示在竖线的左侧，表示算法用来创建模型的数据。

2）预测信息显示在竖线的右侧，表示模型所做出的预测。

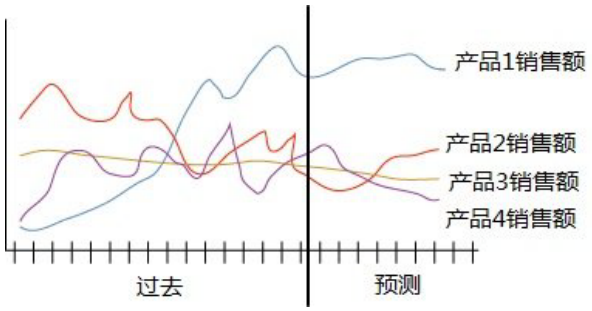


图1 时序模型

2 数据挖掘过程

2.1 数据选择

本文的检测数据取自实验室信息管理系统的数据库中。数据特点：1）系统中涵盖了不同业务种类的数据，比如食品、化妆品等，种类多，抽样区域分布广；2）系统中数据在数据库中的存储结构复杂，例如食品的基本信息、检验信息等都分布在数据库的不同的表中；3）由于不同抽样任务出具不同报告书，因此，结果值输入格式不统一，比如有的填写未检出、有的填写$P<0.5$等。

数据挖掘中不同的挖掘方法对数据集的要求不同。针对关联规则的数据集选择的是样品信息与样品结论，研究食品安全在时间、空间、种类等的关联性。针对时序算法分析的数据集我们选择的样品信息与检验信息中检验结果为数值型的数据。

2.2 数据的转换、清洗、装载

导出的数据集存在不规则、冗余，结果类型多样化等特点，需要对导出的数据进行数据整理、清洗和装载，主要包括以下几个方面：1）按照国标GB2762–2012食品类别（名称）说明，将导出的数据分为17个大类，以现有数据为基础；2）统一检验结果值单位，按统一的规则处理数值型数据以及日期型数据；3）将整理好的数据装载入数据挖掘的数据库。

2.3 数据挖掘工具

本文采用Microsoft SQL Server Analysis Services（SSAS）内嵌的关联规则和时序算法两种数据挖掘算法。因为SQL Server Analysis Service可用于创建、管理和浏览数据挖掘模型，并且SSAS含有数据挖掘扩展插件（DMX）语言，可用于管理挖掘

模型和创建复杂的预测查询，灵活性较高，更重要的是SSAS可以跟SQL数据库无缝连接，因此，选择用SSAS对整理好的数据进行挖掘，主要包括数据源的选择导入、数据源视图的创建、数据挖掘算法（时序预测）的部署，建立数据挖掘模型，可视化浏览数据挖掘模型，如图2所示。

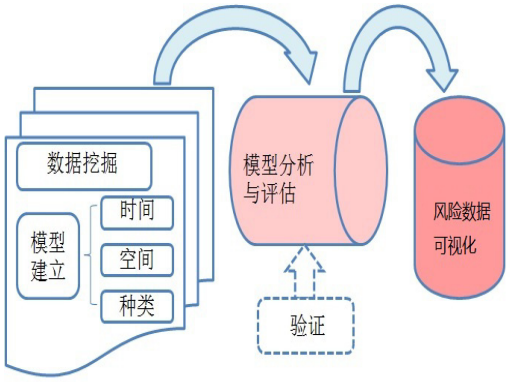


图2 可视化浏览数据挖掘模型

2.4 数据挖掘结果

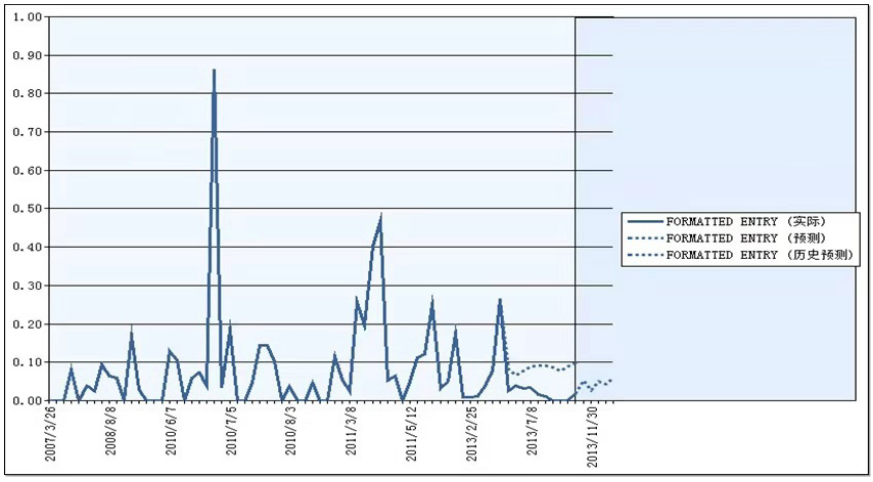
通过以上数据挖掘步骤，建立了两个数据挖掘模型。以下是数据分析结果：

1）关联规则

在清洗过的数据上，建立关联规则挖掘的数据源视图（数据量约为27619条），在此视图上训练数据挖掘模型，通过对食品检测数据的时间、产地、种类的关联性分析，我们得出一些重要的关联性规则。比如产地=**省，食品种类=谷物及其制品，→样品判断=不符合规定，置信度为0.65，支持度为111，重要性为 1.144。通过这些规则可以发现食品安全潜在的风险点，从而可以科学地确定危害物的施检频率，将有限的监管、检验资源放在风险度高的食品危害物检验。因此，关联规则适用于从大量的食品安全检测数据中提取隐含在其中的潜在有用信息，比如在时间、产地（或抽样点）、种类（基质）上的关联性，可实现对食品安全检测数据的深度分析；未来，还可以用于挖掘检测项目数据之间的关联性，发现食品安全风险点。当然，关联规则的产生是基于目前的数据量，在实际应用中需结合专业知识进行分析。

2）时序模型

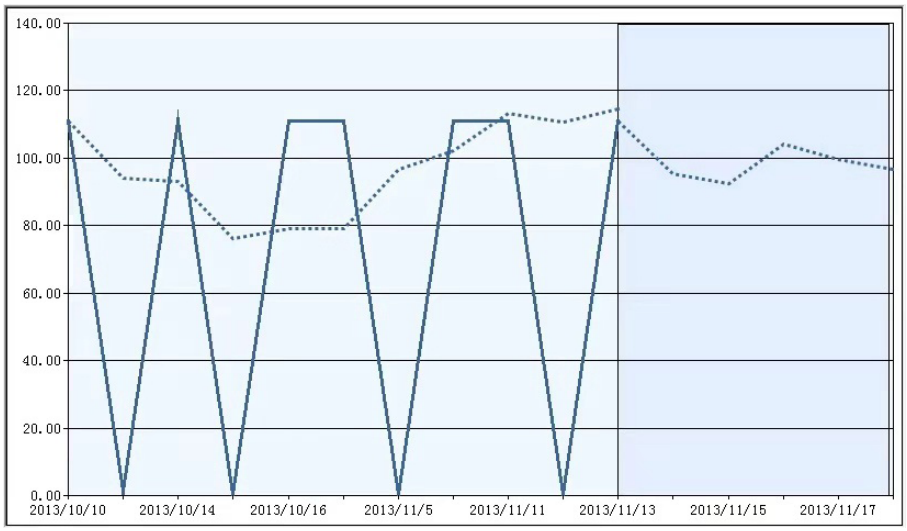
选择畜禽肉类含有铅元素项目的检测数据建立数据源视图，在该视图上建立数据挖掘模型，得到了图3的时序图。



实线部分表示历史数据；虚线部分表示预测数据。

图3 畜禽肉类含有铅元素的时序图

为了寻找微生物类的时序关系，在数据源的选择上，我们也选择了微生物类的检验数据，通过建立数据源视图，在视图上建立数据挖掘模型，同样，得到了图4的大肠菌群的时序图。



实线部分表示历史数据；虚线部分表示预测数据。

图4 大肠菌群的时序图

可以看出一段时间内的历史预测图和实际数据趋势图非常吻合，可用于提供食品安全监管的预警作用。时序分析可用于所有结果值为数值型的检测项目。需要指出，数据量越大，覆盖面越全，预测的趋势将会与实际更吻合。未来也可用于多个检测项目的交叉预测，以挖掘各个检测项目的趋势相关性。

3 结论

通过对数据挖掘技术在食品检测数据中的探索，我们建立了关联模型和时序模型。发现数据挖掘中的关联规则挖掘能够发现大量数据中项集之间的相关联系，适用于食品安全检测数据的分析，可以探索食品安全环节潜在的风险点，可为食品监管部门抽验任务布置、监管决策提供依据。时序模型可以动态监测未来的趋势，对食品安全监管提供数据支持和预警提示。同时，数据挖掘模型的准确性也受限于数据量、数据分布和数据规范性等因素。

通过数据挖掘技术的应用，我们不仅建立了以上两种数据挖掘模型，还总结出了数据挖掘技术对数据规范性和完整性的要求，这为实验室信息管理系统的数据规范性和完整性建立提供了参考意见。未来，随着数据的不断累积，检测数据覆盖范围的增加以及数据格式的规范，将更有利于数据挖掘的应用，更好为食品监管服务。

参考文献：

[1] 秦燕，李辉，李聪. 控制图分析在食品安全预警中的应用[J]. 中国公共卫生，2004，20（9）：1089-1090.

[2] 李聪. 食品安全监测与预警系统（食品安全关键技术系列图书）[J]. 分析仪器，2006，（3）：42-42.

[3] 吉根林，孙志挥. 数据挖掘技术[J]. 中国图象图形学报，2001，6（8）：715-721.

[4] Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules[C]. VLDB，1994：487-499.

[5] 焦李成. 神经网络系统理论[M]. 西安：西安电子科技大学出版社，1990.

[6] Quinlan J.R. C4.5: Programs for Machine Learning[M]. Morgan Kaufmann，1993.

（收稿日期 2018年3月21日 编辑 范玉明）