基于 Apriori 算法的食品抽检数据的关联规则挖掘

宗万里, 朱习军*

(青岛科技大学信息科学技术学院, 青岛 266061)

摘 要: 目的 为了发现检测数据的不合格项目之间有意义的关联规则, 并对挖掘出的关联规则进行分析解读, 进一步发掘了食品抽检数据的价值, 从而对食品安全监管具有一定的指导意义。**方法** 本文对利用 Apriori 算法对 2015~2019 年间山东食品药品监督管理局网站公布的食品安全抽检数据的不合格项目进行了关联规则挖掘。**结果** 通过挖掘得出最符合要求的 10 条规则。**结论** 利用关联规则挖掘算法对食品检验数据进行挖掘, 能够挖掘出有价值、有意义的规则, 对食品安全管理具有指导意义, 从中也可以看出数据挖掘技术在食品安全数据挖掘分析中具有广阔的应用前景。

关键词: 关联规则; Apriori 算法; 食品抽检数据

Mining association rules of food sampling data based on Apriori algorithms

ZONG Wan-Li, ZHU Xi-Jun*

(*School of Information Science and Technology*, *Qingdao University of Science and Technology*, *Qingdao* 266061, *China*)

ABSTRACT: Objective To find the meaningful association rules among the unqualified items of the test data, and to analyze and interpret the association rules, further discover the value of the food sampling data, so as to have a certain guiding significance for the food safety supervision. **Methods** In this paper, the association rules of the unqualified items of the food safety sampling inspection data published on the website of Shandong food and drug administration from 2015 to 2019 were mined by using Apriori algorithm. **Results** Through the excavation, we got 10 rules that most meet the requirements. **Conclusion** Using association rules mining algorithm to mine the food inspection data, we can mine the valuable and meaningful rules, which has the guiding significance to the food safety management. From this, we can see that data mining technology in food safety data mining analysis has a broad application prospect.

KEY WORDS: association rules; Apriori algorithm; food sampling data

1 引 言

随着全社会对食品安全问题的重视程度的不断提高, 政府食品安全监管部门对食品安全检测的投入的力度逐渐加大, 逐渐积累了大量的食品安全检测数据。抽样检测是食品安全监管的重要方式, 为监管提供了重要的技术支撑。每年各级政府投入了大量人财物对食品进行抽检, 逐渐积累了越来越多的检测数据。对这些检测数据进行进一步的分析挖掘利用, 得出一些有价值的知识, 从而进一步指导食品安全监管变得越来越有必要。数据挖掘是从海量数据中发现有趣模式的过程[1]。关联规则挖掘是数据挖掘的方法之一, 是一种描述性的方法, 用于发现隐藏在数据集背后的、项集之间的有意义的关联或相互关系。关联规则挖掘最早是针对购物篮分析而提出的, 目的是通过挖掘购物数据而得知有哪些商品经常被顾客同时购买, 以此分析商品间的关联规则和顾客的行为模式, 并针对这些规则对商品的摆放位置、营销策略等进行调整, 以得到更好的销售效果。例如, 如果顾客在一次超市购物时购买了牛奶, 他们有多大可能也同时

*通讯作者: 朱习军, 博士, 教授, 主要研究方向为数据挖掘、模式识别等。E-mail: 13156283299@163.com

*Corresponding author: ZHU Xi-Jun, Ph.D, Professor, Qingdao University of Science and Technology, No. 99, Songling Road, Laoshan District, Qingdao 266061, China. E-mail: 13156283299@163.com

购买面包以及何种面包？这种信息可以帮助零售商做选择性销售和安排货架空间, 导致增加销售量。

目前关联规则挖掘已经被广泛应用在许多领域之中, 主要有食品领域[2−5]、教育领域[6−8]、交通领域[9,10]、化工领域[11]、医疗领域[12]、通讯领域[13]、军事领域[14]等。本研究提取了 2015~2019 年在山东省食药局网站上公布的不合格食品抽检数据信息, 经过对这些数据进行预处理后, 利用关联规则挖掘方法—Apriori 算法对这些数据的不合格项目进行挖掘, 目的是为了发现不合格项目之间的关联关系, 从而对食品安全管理进行更好的指导。本研究尝试并成功的将数据挖掘技术应用于食品安全数据的挖掘分析, 同时对挖掘出的关联规则进行了进一步的解读, 得出了一些有价值的规则, 证明数据挖掘技术可以并且很好的应用于食品安全数据分析中, 并且从不同的角度可以挖掘出从常规的统计分析中得不到的有价值的信息, 以期为数据挖掘技术在食品安全数据分析领域的应用研究提供科学依据。

2　材料与方法

2.1　数据挖掘相关理论知识

2.1.1　关联规则

数据集 D 是数据库中所有事物的集合, 数据集中每一条记录的各个属性称为项, 属性的集合称为项集。每一条非空记录称为一个事务 T。设 X 和 Y 是事务 T 中包含的 2 个项集, 即 X、Y 均为 T 的真子集。若存在 X 为非空子集, Y 也为非空子集, 且 X 与 Y 的交集为空集, 则 X=>Y 构成事物集 T 中的一条关联规则。也就是说关联规则是一个形如 X=>Y 的表达式, X称为前项, Y 称为后项。

项集中同时含有 X 和 Y 的概率称作 X=>Y 的支持度, 记做 support(X=>Y)=P(X,Y)。

在关联规则的先决条件 X 发生的条件下, 关联结果 Y 发生的概率, 即含有 X 的项集中, 同时含有 Y 的概率, 称为关联规则 X=>Y 的置信度, 记作 confidence(X=>Y)=P(Y|X)=P(X,Y)/P(X)。

含有 X 的条件下同时含有 Y 的可能性, 与没有这个条件下项集中含有 Y 的可能性之比称为关联规则的提升度, 记作 Lift(X=>Y)=P(Y|X)/P(Y)=confidence(X=>Y)/P(Y)。提升度在某种程度上提升度弥补了置信度的缺陷, 其值越大表明 X 对 Y 的提升度越大, 表明关联性越强。

关联规则的挖掘过程就是根据用户给定的最小支持度和最小置信度, 在数据集中通过挖掘频繁项集进而挖掘出强关联规则的过程。

2.1.2　Apriori 算法

Apriori 算法是 Agrawal 和 R.Srikant 于 1994 年提出的, 为布尔型关联规则挖掘频繁项集的原创性算法。Apriori 算法使用一种称为逐层搜索的迭代方法, 利用 k 项集搜索(k+1)项集。首先, 找出所有频繁 1-项集的集合 L_1, 然后用 L_1 生成候选 2-项集的集合 C_2, 通过探查候选 2-项集来形成频繁 2-项集 L_2。以此类推, 使用 L_2 寻找 L_3, 如次迭代, 直至不能找到频繁 k 项集为止。

Apriori 算法主要基于以下 2 点原则:

(1) 如果一个项集为频繁项集, 那么其所有的子集一定也为频繁项集。

(2) 如果一个项集不是频繁项集, 那么其所有的超集一定也不是频繁项集。

2.2　数据挖掘平台 Weka 软件

Weka 软件是一个由新西兰怀可托大学开发的数据挖掘工作平台, 全名是怀卡托智能分析环境。Weka 得到了广泛的认可, 被称为数据挖掘软件工具历史上的里程碑, 是现今最完备的数据挖掘工具之一。Weka 是一款免费的、非商业化的基于 JAVA 环境下开源的机器学习以及数据挖掘软件, Weka 及其源代码可以在官方网站下载, 本研究使用的是 3.8.3 版本。

2.3　食品安全检测数据挖掘

2.3.1　挖掘对象

本研究的数据来源为 2015~2019 年山东食药局官方网站[15]公布的不合格食品检测样品信息,将所有的不合格的样品汇集到一起, 共计 3848 条数据。不合格样品信息主要包括食品名称、标称生产企业名称、标称生产企业地址、被抽样单位名称、被抽样单位地址、规格型号、商标、生产日期、不合格项目、食品分类、任务来源、检验机构等, 本研究主要挖掘不合格项目之间的关联规则, 因此更多关注不合格属性。

2.3.2　选定数挖掘算法

本研究使用关联规则挖掘算法 Apriori 进行关联规则挖掘。

2.3.3　数据预处理

本研究的挖掘目的是寻找不合格样品之间的关联规则, 在数据预处理过程中去掉了与数据挖掘不相关的属性, 只保留不合格项目的信息, 将同一样品中含有多个不合格项目分为 item1、item2、item3 等字段。将同时有 2 个不合格项的样品信息汇总为一张表, 将同时有 3 个不合格项目的样品信息汇总为一张表, 将同时有 4 个不合格项目的样品信息汇总为一张表, 分别进行关联规则挖掘。

由于 weka 软件对于汉字的支持性能不是太好, 本文将不合格项目名称以汉语拼音的形式表示。转换后的表格如下图 1~2 所示。最后将 xls 格式的文件另存为 csv 格式。

3　结果与分析

3.1　使用 Weka 软件对经过数据预处理后的食品安全检测数据进行挖掘

数据经过预处理过程后, 开始对食品检测数据进行的数据挖掘实验。启动 Weka 软件, 进入 Explorer 界面, 选

择 preprocess 选项卡，打开经过预处理后的 csv 格式的数据表格文件，点击 save，将文件保存为 arff 格式，重新打开该 arff 格式的文件，选择 Associate 选项卡中的 Apriori 算法，设定最小支持度阈值 0.01，最小置信度阈值 0.7，分别对 2 个不合格项目、3 个不合格项目、4 个不合格项目的数据集进行挖掘，最终分别得到 3 个数据集的挖掘结果，每个数据集最符合要求的 10 条关联规则如图 3~5 所示。

item1	item2
yaliusuanyan	benjiasuan
yaliusuanyan	benjiasuan
yanzhihong	riluohuang
lvhuana	guansuanna
zongdan	tangjingna
lvhuana	guansuanna
lvhuana	guansuanna

图 1 2 个属性数据集

Fig.1 Two kinds of attribute data sets

item1	item2	item3
yanzhihong	riluohuang	benjiasuan
yanzhihong	riluohuang	benjiasuan
tianmisu	guotangputaotang	benjiasuan
ningmenghuang	eryanghualiu	benjiasuan
tianmisu	tangjingna	benjiasuan
yanzhihong	riluohuang	benjiasuan
yanzhihong	riluohuang	benjiasuan

图 2 3 个属性数据集

Fig.2 Three kinds of attribute data sets

3.2 处理结果分析

不合格的食品检测指标主要分为金属元素、着色剂、防腐剂、甜味剂、农药残留、兽药残留、理化指标等类别。金属元素不合格产生的原因可能有环境污染、加工过程污染、生物体蓄积。着色剂、防腐剂、甜味剂不合格的原因主要为超量、超范围使用添加剂，农兽药残留不合格的原因可能为超量使用农兽药、违禁使用国家明令禁止使用的农兽药，理化指标不合格的原因可能为生产质量不合格的食品，生产过程中工艺参数控制不当等。通过关联规则挖掘尝试得出同一样品的多个不合格项目之间的关联关系。

对部分挖掘出的关联规则的解读:

1、谷氨酸钠不合格的样品同时氯化钠也可能不合格，菌落总数不合格的样品，大肠菌群也可能会不合格，亚硫酸盐不合格的样品也可能会苯甲酸钠不合格，防腐剂占各自最大用量比例之和的样品同时山梨酸不合格，甲醛次硫酸氢钠不合格的样品同时二氧化硫不合格，那可丁不合格的样品罂粟碱同时会不合格，亚硫酸盐不合格的样品同时会有苯甲酸钠不合格，酒精度不合格的样品同时会有甜蜜素不合格，日落黄不合格的样品同时会有柠檬黄不合格，铝不合格的样品同时会有日落黄、柠檬黄不合格等。

```
Best rules found:

 1. item2=guansuanna 7 ==> item1=lvhuana 7    <conf:(1)> lift:(28.43) lev:(0.03) [6] conv:(6.75)
 2. item1=lvhuana 7 ==> item2=guansuanna 7    <conf:(1)> lift:(28.43) lev:(0.03) [6] conv:(6.75)
 3. item1=dachangjunqun 7 ==> item2=junluozongshu 7    <conf:(1)> lift:(19.9) lev:(0.03) [6] conv:(6.65)
 4. item2=yaliusuanyan 5 ==> item1=benjiasuan 5    <conf:(1)> lift:(9.48) lev:(0.02) [4] conv:(4.47)
 5. item2=fangfujigeziyongliangzhanzuidabilizhihe 4 ==> item1=shanlisuan 4    <conf:(1)> lift:(24.88) lev:(0.02) [3] conv:(3.84)
 6. item1=jiaquancilisuanqingna 3 ==> item2=eryanghualiu 3    <conf:(1)> lift:(39.8) lev:(0.01) [2] conv:(2.92)
 7. item2=nakeding 3 ==> item1=yingsujian 3    <conf:(1)> lift:(66.33) lev:(0.01) [2] conv:(2.95)
 8. item1=yingsujian 3 ==> item2=nakeding 3    <conf:(1)> lift:(66.33) lev:(0.01) [2] conv:(2.95)
 9. item1=yaliusuanyan 2 ==> item2=benjiasuan 2    <conf:(1)> lift:(4.74) lev:(0.01) [1] conv:(1.58)
10. item2=jiujingdu 2 ==> item1=tianmisu 2    <conf:(1)> lift:(6.86) lev:(0.01) [1] conv:(1.71)
```

图 3 2 个属性的关联规则挖掘结果

Fig.3 Results of mining association rules with 2 kinds of attributes

```
Best rules found:

 1. item1=riluohuang 13 ==> item2=ningmenghuang 13    <conf:(1)> lift:(3.4) lev:(0.18) [9] conv:(9.18)
 2. item3=lv 13 ==> item1=riluohuang 13    <conf:(1)> lift:(3.92) lev:(0.19) [9] conv:(9.69)
 3. item1=riluohuang 13 ==> item3=lv 13    <conf:(1)> lift:(3.92) lev:(0.19) [9] conv:(9.69)
 4. item3=lv 13 ==> item2=ningmenghuang 13    <conf:(1)> lift:(3.4) lev:(0.18) [9] conv:(9.18)
 5. item2=ningmenghuang item3=lv 13 ==> item1=riluohuang 13    <conf:(1)> lift:(3.92) lev:(0.19) [9] conv:(9.69)
 6. item1=riluohuang item3=lv 13 ==> item2=ningmenghuang 13    <conf:(1)> lift:(3.4) lev:(0.18) [9] conv:(9.18)
 7. item1=riluohuang item2=ningmenghuang 13 ==> item3=lv 13    <conf:(1)> lift:(3.92) lev:(0.19) [9] conv:(9.69)
 8. item3=lv 13 ==> item1=riluohuang item2=ningmenghuang 13    <conf:(1)> lift:(3.92) lev:(0.19) [9] conv:(9.69)
 9. item1=riluohuang 13 ==> item2=ningmenghuang item3=lv 13    <conf:(1)> lift:(3.92) lev:(0.19) [9] conv:(9.69)
10. item2=ningmenghuang 15 ==> item1=riluohuang 13    <conf:(0.87)> lift:(3.4) lev:(0.18) [9] conv:(3.73)
```

图 4 3 个属性的关联规则挖掘结果

Fig.4 Mining results of association rules with 3 kinds of attributes

```
Best rules found:

 1. item2=tianmisu 1 ==> item1=xiancaihong 1    <conf:(1)> lift:(5) lev:(0.16) [0] conv:(0.8)
 2. item1=xiancaihong 1 ==> item2=tianmisu 1    <conf:(1)> lift:(5) lev:(0.16) [0] conv:(0.8)
 3. item3=jiujingdu 1 ==> item1=xiancaihong 1    <conf:(1)> lift:(5) lev:(0.16) [0] conv:(0.8)
 4. item1=xiancaihong 1 ==> item3=jiujingdu 1    <conf:(1)> lift:(5) lev:(0.16) [0] conv:(0.8)
 5. item1=xiancaihong 1 ==> item4=benjiasuan 1    <conf:(1)> lift:(2.5) lev:(0.12) [0] conv:(0.6)
 6. item2=nakeding 1 ==> item1=yingsujian 1    <conf:(1)> lift:(5) lev:(0.16) [0] conv:(0.8)
 7. item1=yingsujian 1 ==> item2=nakeding 1    <conf:(1)> lift:(5) lev:(0.16) [0] conv:(0.8)
 8. item3=mafei 1 ==> item1=yingsujian 1    <conf:(1)> lift:(5) lev:(0.16) [0] conv:(0.8)
 9. item1=yingsujian 1 ==> item3=mafei 1    <conf:(1)> lift:(5) lev:(0.16) [0] conv:(0.8)
10. item4=kedaiyin 1 ==> item1=yingsujian 1    <conf:(1)> lift:(5) lev:(0.16) [0] conv:(0.8)
```

图 5 4 个属性的关联规则挖掘结果

Fig.5 Mining results of association rules with 4 kinds of attributes

2、同一样品可能会产生多项指标不合格，不合格指标之间存在一定的关联关系。生产厂家可能会使用多种甜味剂来实现甜味的功能，使用多种防腐剂来实现防腐的功能。

3、柠檬黄、日落黄不合格的样品同时会有铝不合格。可能的原因是在使用柠檬黄、日落黄这些人工合成着色剂时是以柠檬黄铝色淀、日落黄铝色淀的形式加入的，这些添加剂本身含有铝元素，从而造成了最终产品中铝含量超标。

4、罂粟碱、可待因、那可丁、吗啡会同时出现在同一样品中是由于罂粟壳内的成分包括吗啡、可待因、那可丁、罂粟碱等 30 多种生物碱，在食物中添加的非法添加物为罂粟壳而不是单纯的具体某一种物质。

5、苋菜红、甜蜜素、酒精度的指标之间有较强的关联性说明了对于果酒、葡萄酒等酒类制品的生产中。一些不法厂商为了谋取较高的利润，通过使用色素、甜味剂来生产一些劣质的酒类供应市场的现象仍然存在。

4　结　论

通过利用数据挖掘算法 Apriori 算法对经过预处理后的食品抽检数据进行关联规则挖掘，挖掘了出不合格项目之间的关联规则，通过对规则进行解读，对于食品安全监管具有一定的帮助。同时也说明数据挖掘技术能够应用在食品安全数据分析之中，并且具有较大的应用前景和现实意义，值得进行进一步的、深入的应用研究。

参考文献

[1] Han JW, Kamber M, Pei J. 数据挖掘概念与技术[M]. 北京: 机械工业出版社, 2012.
Han JW, Kamber M, Pei J. Data mining concepts and technique [M]. Beijing: China Machine Press, 2012.

[2] 汪海萍, 胡小崧, 何苏, 等. 基于食品检验数据的关联规则研究[J]. 现代食品, 2017, 24: 35–38.
Wang HP, Hu XS, He S, *et al*. Research on association rules based on food inspection data [J]. Mod Food, 2017, 24: 35–38.

[3] 汪雪君, 沈怡, 杨慧元. 数据挖掘技术在食品检测数据中的探索[J]. 中国药事, 2019, 33(3): 259–262.
Wang XJ, Shen Y, Yang HY. The exploration of data mining data of food control [J]. Chin Pharm Affa, 2019, 33(3): 259–262.

[4] 张洋, 陈伟炯, 付姗姗. 基于数据挖掘的食品供应链风险预警系统研究[J]. 广西大学学报(自然科学版), 2018, 43(3): 1118–1125.
Zhang Y, Chen WJ, Fu SS. Research on food supply chain risk early warning system based on data mining [J]. J Guangxi Univ (Nat Sci Ed), 2018, 43(3): 1118–1125.

[5] 李香串, 乔丽芳, 任飒. 基于关联规则法研究黄芪在保健食品配方中的应用规律[J]. 山西医科大学学报, 2019, 50(1): 40–49.
Li XC, Qiao LF, Ren S. Application law of huangqi in the formula of health-care food by data association rule [J]. J Shanxi Med Univ, 2019, 50(1): 40–49.

[6] 亓文娟, 晏杰, 黄书城, 等. 关联规则挖掘在大学生心理健康测评系统中的应用研究[J]. 湖南工业大学学报, 2013, 27(6): 94–99.
Qi WJ, Yan J, Huang SC, *et al*. The Application of association rule mining in college students'mental health assessment system [J]. J Hunan Univ Technol, 2013, 27(6): 94–99.

[7] 陆鑫赟, 王兴芬. 基于领域关联冗余的教务数据关联规则挖掘[J]. 计算机科学, 2019, 46(6A): 427–430, 435.
Lu XY, Wang XF. Educational administration data mining of association rules based on domain association redundancy [J]. Comp Sci, 2019, 46(6A): 427–430, 435.

[8] 张鸿雁. 基于 Apriori 算法的校园教学质量评价系统设计[J]. 电子技术与软件工程, 2019, 18: 188–189.
Zhang HY. Design of campus teaching quality evaluation system based on Apriori Algorithm [J]. Electron Technol Software Eng, 2019, 18: 188–189.

[9] Shi CH, Ding Y, Yue GF, *et al*. Analysis of causes of traffic accidents based on improved Apriori association rules [C]. Sydney: 2019 9th International Conference on Management, Education and Information, 2019.

[10] 黄常海, 高德毅, 胡甚平, 等. 基于 Apriori 算法的船舶交通事故关联规则分析[J]. 上海海事大学学报, 2014, 35(3): 18–22.
Huang CH, Gao DY, Hu SP, *et al*. Association rule analysis of vessel traffic accidents based on Apriori algorithm [J]. J Shanghai Maritime Univ, 2014, 35(3): 18–22.

[11] Wang J, Li HG, Huang JW, *et al*. Association rules mining based analysis of consequential alarm sequences in chemical processes [J]. J Loss Prev Process Ind, 2016, 41: 178–185.

[12] Dong GL, Kwang SR, Mohamed Bashir, *et al*. Discovering medical knowledge using association rule mining in young adults with acute myocardial infarction [J]. J Med Syst, 2013, 37(2): 1–10.

[13] Guo Y, Wang MX, Li X. Application of an improved Apriori algorithm in a mobile e-commerce recommendation system [J]. Ind Manage Data Syst, 2017, 117(2): 287–303.

[14] 曹冠平, 王跃利, 张立韬. 关联规则挖掘在作战实验数据分析中的应用[J]. 计算机控制与仿真, 2019, 41(2): 70–74.
Cao GP, Wang YL, Zhang LT. Application of association rule mining in combat experiment data analysis [J]. Comput Control Simul, 2019, 41(2): 70–74.

[15] 山东省食品药品监督管理局官网[Z]. http://www.sdfda.gov.cn/
Shandong food and drug administration website [Z]. http://www.sdfda.gov.cn/

(责任编辑：于梦娇)

作者简介

宗万里，硕士研究生，主要研究方向为数据挖掘与智能系统。
E-mail: zwanLzjq@163.com

朱习军，博士，教授，主要研究方向为数据挖掘、模式识别等。
E-mail: 13156283299@163.com