doi:10.3969/j.issn.2095-6002.2017.02.014　　文章编号:2095-6002(2017)02-0089-06
引用格式:肖克晶,左敏,王星云,等. 改进的关联规则在食品安全预警上的应用[J]. 食品科学技术学报,2017,35(2):89-94.
XIAO Kejing, ZUO Min, WANG Xingyun, et al. Application of improved association rules on food safety early warning [J]. Journal of Food Science and Technology, 2017,35(2):89-94.

改进的关联规则在食品安全预警上的应用

肖克晶[1], 左　敏[1,*], 王星云[1], 刘　婷[2]
(1. 北京工商大学 计算机与信息工程学院, 北京　100048;
2. 中国食品药品检定研究院, 北京　100050)

摘　要: 为将海量的食品检测数据有效地应用于食品安全预警,首先分析了食品检测数据的特点,以及传统的 Apriori 算法在挖掘食品检测数据上的不足,进而提出过滤算法,并将其作为 Apriori 算法的前置组件对算法进行改进,然后建立了食品安全预警模型,最后将实际的食用油检测数据用改进后的算法进行挖掘,发现其存在的潜在安全隐患进而做出风险预警。通过实验对比 Apriori 算法,发现改进后的算法摒弃了大量的伪关联规则,能有效提高食品安全预警的效率和准确度,具有十分重要的实际意义。

关键词: 关联规则; 频繁项集; 稀疏数据; 过滤算法
中图分类号: TS201.6; TP391.9　　**文献标志码:** A

食品安全预警是食品安全监管的重要环节,合理有效的预警方法能大大提高食品安全整体水平。因此,如何识别食品安全风险并做出预警,降低食品安全事故的发生概率,成为监管部门面临的重要挑战。国内外学者目前在食品安全领域做了很多研究,包括对国内外食品安全预警机制的对比[1]、基于互联网舆情监测的食品安全预警[2]、如何在流通领域加强食品安全[3]以及对食品安全风险因素的分析等[4]。在食品安全预警方面,国外学者 McMeekin 等[5]研究了欧盟"食品与饲料快速预警系统",Kadir 等[6]用自适应模糊推理系统 ANFIS 预测粮食安全问题,国内学者 Zhu 等[7]研究了基于支持向量机分类的食品风险预警模型,章德宾等[8]研究了基于 BP 神经网络的食品安全预警方法,王海明等[9]通过对食品安全综合评价指数的研究提出风险监测预警系统,顾小林等[10]研究了基于关联规则挖掘的食品安全信息预警模型。

已有研究为本文的研究奠定了一定的基础,但是这些研究大多是分析预警模型如何建立[11],以及针对生产和流通过程中的影响因素做定性分析[12],并没有针对食品检测数据的预警研究。食品检测数据来源于监管部门在超市等市场终端的抽检结果,主要用于监管部门的统计分析和决策。这些数据隐含了很多有价值的信息,必须对其进行分析挖掘才能找到其中的安全隐患因素[13]。因此对食品检测数据采用更加科学的挖掘算法进行分析、提高挖掘的准确度和效率成为目前研究的重点。本文首先介绍了关联规则挖掘算法,分析了食品检测数据的特点并提出了过滤算法,然后将过滤算法作为 Apriori

收稿日期: 2015-12-03
基金项目:"十二五"国家科技支撑计划项目(2015BAK36B04)。
作者简介: 肖克晶,女,硕士研究生,研究方向为食品安全数据挖掘;
*左　敏,男,教授,博士,主要从事人工智能方面的研究,通信作者。

算法的前置组件对其进行改进,建立了相应的预警模型,最后对实际的食用油检测数据进行挖掘,发现其中存在的安全风险因素,并根据分析结果做出风险预警。

1 关联规则挖掘方法

关联规则是 Agrawal 等在 1993 年提出的,其目标是找到同时满足最小支持度和最小置信度的强关联规则,过程分为 2 个步骤:1)找到所有满足支持度的频繁项集;2)使用频繁项集生成强关联规则[14]。其中 Apriori 算法是目前应用最广泛的算法,其原理可以表示如下:

设 $I=\{i_1,i_2\cdots i_m\}$ 是由 m 个不同的数据项组成的集合,其中元素称为项,项的集合称为项集。给定一个事务数据库 $D=\{T_1,T_2\cdots T_n\}$,其中每一个事务 T 是项集 I 的一个子集,$|D|$ 表示 D 中的事务总数。其中 X 和 Y 都是 T 中的项集且不相交,假设 $num(\ )$ 表示事务数据库里特定项集出现的次数,那么就可以得到形如 $X\rightarrow Y$ 的关联规则表达式。

其中,支持度(*Support*)表示项集 $\{X,Y\}$ 在总项集里出现的概率,计算如公式(1)。

$$Support(X\rightarrow Y)=P(XY)=num(X\cup Y)/|D|。\tag{1}$$

置信度(*Confidence*)表示在含有 X 的项集中,同时含有 Y 的概率,计算如公式(2)。

$$Confidence(X\rightarrow Y)=P(X|Y)=num(X\cup Y)/num(X)。\tag{2}$$

Apriori 算法首先扫描数据库找到所有频繁 1-项集,然后由 Apriori_gen 算法找到所有候选 2-项集并扫描数据库,对每个项进行计数找到所有频繁 2-项集,以此类推直至找到所需要的频繁 k-项集,然后由频繁项集生成强关联规则[15]。

但是 Apriori 算法有着自身的缺点和不足,包括:1)对数据库的扫描次数过多,I/O 代价很高;2)产生大量的中间项集;3)对不同的数据集动态更新适应性较差。因此在使用 Apriori 算法挖掘时,时间和空间的消耗成为制约挖掘效率的关键。根据已有研究,频繁项集挖掘的时间复杂度是 $O(2^m)$,即属性项的数目 m 对频繁项集挖掘的复杂度影响很大[16]。本文将根据具体的食品检测数据特点,针对此问题给出相应的解决办法。

2 数据来源与预处理

2.1 数据来源

所用数据全部来自国家食品安全抽检监测信息系统,该平台保存了各省市的食用油检测数据,选取 2014 年的数据,对这些数据进行清理噪声、一致性检查等预处理操作,然后将数据转换成挖掘算法需要的格式。

2.2 数据预处理

经过数据的准备工作以后,针对属性间关联规则的挖掘目标,将食用油的检测数据按以下方式组织存储,每一行记录代表一种特定的食用油产品,一项产品对应着 20 种属性,包括它的生产商所在省份、食用油细类、生产时间以及 17 种质量检测指标:酸值/酸价(KOH)、总砷、反式脂肪酸、丁基羟基茴香醚(BHA)、黄曲霉毒素 B1、二丁基羟基甲苯(BHT)、月桂酸、苯并[a]芘、饱和酸、二十四烷酸、过氧化值、没食子酸丙酯(PG)、溶剂残留量、铅、二十二碳二烯酸、游离棉酚、癸酸。然后对每一项食用油检测结果进行标号,省份、食品细类保持不变,生产时间转换成季度,17 种检测项分别用 A-Q 表示,例如 A 表示酸值/酸价(KOH),B 表示溶剂残留量等;将合格项标记为 0,不合格项标记为 1,未检验项标记为 null。例如,若检测项 A 为不合格项,则将其标记为 A_1,若某检测项 B 为合格项,则将其标记为 B_0,预处理后的数据如表 1。

表 1 预处理后的食用油检测数据

Tab. 1 Oil detection data after pretreatment

省份	食用油细类	生产时间	酸值/酸价(KOH)	溶剂残留量	反式脂肪酸	黄曲霉毒素 B1	饱和酸	……
广西	花生油	第 2 季度	A_0	B_0	null	D_1	null	……
吉林	其他食用植物油	第 2 季度	A_0	B_1	null	D_0	null	……
江西	其他食用植物油	第 1 季度	A_0	B_0	null	D_0	null	……

2.3　数据特点

分析可知表 1 具有如下特征：

1)多维属性。每一个属性都有几种不同的取值,即数据具有多维属性。

2)类别型。每一个属性的取值都是离散的,即数据为类别型。

3)稀疏性。如果一个数据集有大量属性是 null,则认为这个数据集具有稀疏性[17]。由表 1 可知各检测项目有很多属性的取值为 null,各属性 null 值统计结果如表 2(共 14 例)。

表 2　食用油检测数据 null 值所占比例

Tab. 2　Null value proportions of oil detection data

属性	取值个数	null 值个数	null 值比例/%
省份	22	0	0
生产日期(月份)	22	0	0
所属细类	22	0	0
酸值/酸价(KOH)	22	0	0
总砷	22	0	0
苯并[a]芘	22	0	0
黄曲霉毒素 B1	22	0	0
月桂酸	22	15	68
饱和酸	22	14	63
二十四烷酸	22	15	68
二十二碳二烯酸	22	11	50
游离棉酚	22	15	68
反式脂肪酸	22	22	100
癸酸	22	22	100

由表 2 可以看出,部分属性如二十四烷酸、游离棉酚等检测项目 null 值比例超过 50%,反式脂肪酸、癸酸的 null 值甚至达到了 100%,说明食用油检测数据具有一定的稀疏性。当 null 作为其属性取值进行关联规则挖掘时,会产生大量含 null 的频繁项集,最后会生成很多含有 null 的伪关联规则。因此,为了使挖掘出的频繁项集更有实际意义,必须采用相应的算法对含有大量 null 值的数据集进行过滤,以保证关联规则挖掘的有效性。

3　食品安全预警系统设计

针对关联规则挖掘的目标和食品检测数据的特点,建立了基于关联规则挖掘的食品安全预警系统,主要是对食品检测数据库中的实际检测数据进行挖掘,并根据挖掘结果判断其风险情况从而做出预警。预警系统设计主要包括:数据源模块、预警分析模块、反应模块等,如图 1。

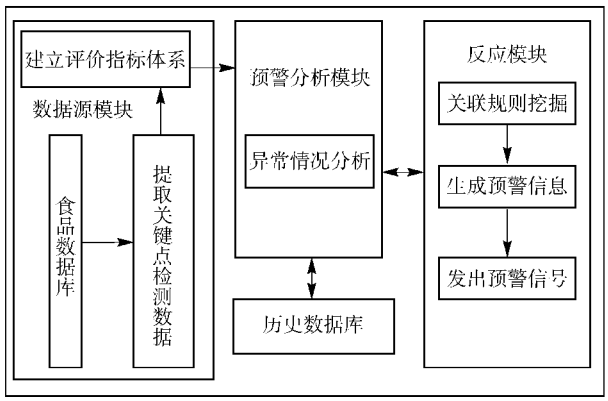

图 1　基于关联规则挖掘的食品安全预警模型

Fig. 1　Food safety early warning model diagram based on association rules

其中,数据源模块是预警系统的数据来源,是来自国家食品安全抽检监测信息系统,数据源模块主要是对数据进行相应的预处理、提取关键点检测数据、建立相应的评价指标体系。预警分析模块是依据历史数据库对异常情况进行分析,反应模块采用关联规则算法挖掘数据项中有意义的规律,发现潜在的安全隐患,生成预警信息并发出预警信号通知相关监管部门,以利于相关部门下一步的工作决策和安排。

4　过滤算法的设计与实现

4.1　算法意义

频繁项集挖掘的时间和空间消耗主要在于:1)计算了过多的候选项集;2)多次扫描数据库计算每个候选项集的支持度[18]。由于 Apriori 算法本身没有过滤稀疏数据的机制,如果不对 null 值进行处理,就会计算过多无意义的候选项集,这样就导致 I/O 代价很大,挖掘效率降低,时间复杂度呈指数增长。采用过滤算法对食用油检测数据集进行过滤,只过滤无效的 null 值数据,得到一个较小但价值密度更高的数据集,能提高挖掘效率和准确度,而且不会破坏原有数据集的有效性和完整性[19-20],从而解决了前文所提到的时间复杂度过大的问题。

4.2　算法原理

根据 Apriori 算法的基本原理:所有频繁项集的子集也是频繁项集,所以频繁 1-项集越少则频繁 k-项集越少[21]。过滤算法第一步是把原始数据集的属性加以分类,标记不含 null 的属性,然后遍历标记

后的新的数据集生成候选 1-项集,将所有 null 值产生的候选 1-项集过滤掉,生成只包含非 null 取值的候选 1-项集,以此候选项集作为关联规则挖掘算法的输入。

4.3 算法描述

设原始数据集的属性为 A_i,其中($i=1,2,\cdots m$),设每个 A_i 有 k_i 个取值,A_i 的值域为 $\{a_i[j]\}$,($j=1,2,\cdots k_i$),候选 1-项集 $C_1=\{c_1,c_2\cdots c_n\}$,输入的是原始数据集,过滤参数 $p\in\{a_i[j]\}$,以 $p=null$ 为例,输出候选 1-项集 C_1。

```
1  for (i=0; i<m; i++)
2      for(j=0; j<k_i; j++)
3          if ∀j<k_i, a_i[j]≠null then mark A_i, 0<i<u
4  for (i=0; i<u ; i++)
5      for (j=0; j<k_i; j++)
6          c_i = { A_i = a_i[j] }
7  for (i=u;i<m;i++)
8      for (j=0;j<k_i;j++)
9          if a_i[j]≠null   c_i = {A_i = a_i[j]}
10 output C_1 = {c_1,c_2⋯c_n}
```

4.4 预警流程

首先用过滤算法遍历食用油检测数据库,同时计算出不包含 null 值的所有候选 1-项集作为挖掘算法的输入,然后对数据进行关联规则挖掘,具体流程如图 2。

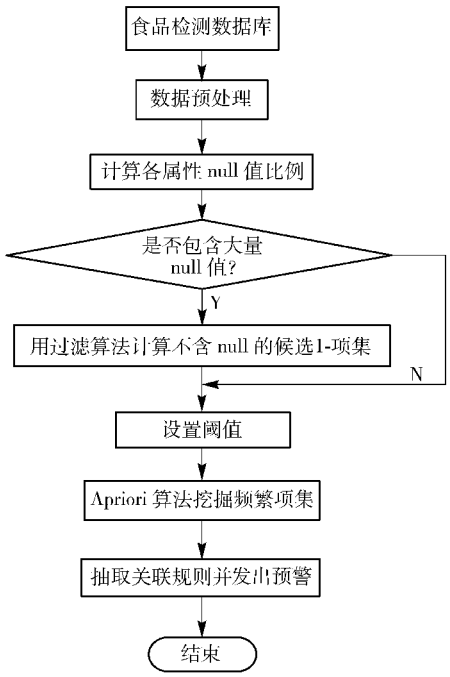

图 2 基于关联规则挖掘的食品安全预警流程

Fig. 2 Flowchart of food safety early warning model diagram based on association rules

4.5 实验与结果分析

实验使用过滤算法产生候选 1-项集和不使用过滤算法两种情况下,用 Apriori 算法对关联规则挖掘的性能进行对比。采用 2014 年全国各省市部分食用油实际检测数据为实验数据,其中记录数 $D=8\,343$,实验环境为 window 7,所用的挖掘软件是 R,变换支持度 S 并比较二者的挖掘效率和结果,如表 3 和表 4。

表 3 未使用过滤算法时的挖掘结果

Tab. 3 Results without using filtering algorithm

支持度 S	挖掘所用时长/s	关联规则数目
0.001	30.05	2 230
0.002	19.23	1 022
0.003	8.14	233
0.004	2.56	87

表 4 使用了过滤算法以后的挖掘结果

Tab. 4 Results with using filtering algorithm

支持度 S	挖掘所用时长/s	关联规则数目
0.001	2.59	268
0.002	1.32	65
0.003	1.21	15
0.004	0.12	6

由表 3 和表 4 的对比可以明显看出:在不使用过滤算法的情况下,当支持度相同时,挖掘到的关联规则数目总是多于使用了过滤算法以后的 Apriori 算法挖掘到的关联规则数目,如当支持度 $S=0.001$ 时,不使用过滤算法时关联规则数目达到了 2 230,而使用了过滤算法以后关联规则数目是 268,可见使用过滤算法可以过滤掉由大量含有 null 值的频繁项集生成的伪关联规则,提高了挖掘的精度。比较挖掘所用时长,当 $S=0.002$ 时,不使用过滤算法的情况下用时 19.23 s,而使用了过滤算法以后只需要 1.32 s,说明使用过滤算法可以大大提高挖掘效率,节省一定的时间。由此可见,使用过滤算法可以同时提高挖掘的精度和挖掘效率,具有重要的实际意义。当支持度 $s=0.005$ 时,在使用过滤算法的情况下,挖掘得到强关联规则部分示例如下:

{省份 = 海南省,检测项目 = 黄曲霉毒素 B1,时间 = 第 3 季度}→{结果判定 = D_1}

{食品细类 = 其他食用植物油,检测项目 = 溶剂残留量,时间 = 第 4 季度}→{结果判定 = B_1};

{食品细类 = 其他食用植物油,检测项目 = 过氧化值,时间 = 第 4 季度}→{判定结果 = K_1}。

示例第一条强关联规则的意思是:当产地所在省份是海南,生产时间是第 3 季度时,黄曲霉毒素 B1 不合格的风险很大。其他关联规则的含义可以此类推。

由以上关联规则可以分析出 2014 年食用油生产存在的安全问题及其应对措施主要有:

1)海南省第 3 季度生产的食用油中,黄曲霉毒素 B1 不合格的风险很大,应当发出风险预警,在下一年第 3 季度的抽检工作中加强对海南省食用油黄曲霉毒素 B1 含量的抽检力度。

2)第 4 季度生产的食用油中,溶剂残留量和过氧化值不合格的风险很大,应当发出风险预警,在下一年第 4 季度全国范围内的抽检工作中将其作为重点抽检对象。

综上所述,食品检测数据具有很明显的稀疏性,即当支持度很低时,才能挖掘到属性间的关联规则。这时产生频繁项集在使用了经过滤算法改进的 Apriori 算法进行挖掘后,产生有意义的规则数目要少很多,易于观察和理解。这说明对食品检测数据采用过滤算法之后,再用 Apriori 算法来挖掘频繁项集是正确且有效的,它比传统 Apriori 算法模型的预警效果更好。

5 结　论

基于食品检测数据的特点和频繁项集挖掘的难点,提出过滤算法并对 Apriori 算法进行改进。通过实验表明,改进的算法对于带有 null 值的稀疏数据集非常有效,避免生成含有 null 值的伪关联规则,提高了算法的性能。最后通过对实际的食用油检测数据进行挖掘,得出风险情况并进行预警。

但是在实验中也发现了一些问题,还需进一步改进,可以从以下 3 方面进行:

1)以后的研究中需要增大实验数据的数量,以提高挖掘的准确性。

2)食品安全预警模型还不够完善,需要进一步对其结构进行研究。

3)食品检测数据量十分庞大,应探索更加科学高效的数据预处理方法。

参考文献:

[1] ZHOU Qiang, GONG Chen, ZHOU Yi. Public food safety pre-warning system of crisis management[C]// Information Systems for Crisis Response and Management (ISCRAM), 2011 International Conference on. Piscataway, United States, November 25-27, 2011. IEEE, 2011:158 – 162.

[2] LI Hui, XIAO Hang, QIU Tianchen, et al. Food safety early warning research based on internet public opinion monitoring and tracing[C]// Agro-Geoinformatics (Agro-Geoinformatics), 2013 Second International Conference on. Washingtong DC, United States, August 13-16, 2013. IEEE,2013:481 – 484.

[3] YI Ming. How to strengthen food safety in circulation field by HACCP[C]// Logistics Systems and Intelligent Management, 2010 International Conference on. Piscataway, United States, January 9-10, 2010. IEEE,2010:1746 – 1750.

[4] XU J, DENG Y. Food safety risk analysis based on generalized fuzzy numbers[C]// Advanced Management Science (ICAMS), 2010 IEEE International Conference on. IEEE, 2010: 699 – 702.

[5] MCMEEKIN T A, ROSS T. Predictive microbiology: providing a knowledge-based framework for change management[J]. International Journal of Food Microbiology, 2012, 78(1): 133 – 153.

[6] KADIR M K A, HINES E L, AROF S, et al. Grain security risk level prediction using ANFIS[C]// Computational Intelligence, Modelling and Simulation (CIMSiM), 2011 Third International Conference on. Piscataway, United states, September 20-22, 2011. IEEE, 2011: 103 – 107.

[7] ZHU Changxing, WANG Feng. Study on risk pre-warning model of China food based on SVM classification[C]// E-Pro-duct E-Service and E-Entertainment (ICEEE), 2010 International Conference on. Piscataway, United States, November 7-9, 2010. IEEE, 2010: 1 – 3.

[8] 章德宾,徐家鹏,许建军,等. 基于监测数据和 BP 神经网络的食品安全预警模型[J]. 农业工程学报, 2010, 26(1):221 – 226.
ZHANG Debin, XU Jiapeng, XU Jianjun, et al, Model for food safety warning based on inspection data and BP neural network[J]. Transactions of the Chinese Society of

Agricultural Engineering, 2010, 26(1):221 – 226.
[9] 王海明,郑培,潘海虹. 食品安全风险监测预警系统研究[J]. 中国卫生监督杂志, 2010, 17(6):529 – 533.
[10] 顾小林,张大为,张可,等. 基于关联规则挖掘的食品安全信息预警模型[J]. 软科学, 2011, 25(11): 136 – 141.
GU Xiaolin, ZHANG Dawei, ZHANG Ke, et al. The information pre-warning model of food safety based on association rules mining[J]. Soft Science, 2011, 25(11): 136 – 141.
[11] 肖宛凝. 吉林省食品安全风险监测预警系统构建研究[D]. 长春: 吉林大学,2014.
[12] 胡春林. 基于供应链管理的食品安全风险预警系统研究[J]. 经济师,2012(7):35 – 37.
[13] WANG Yuhong, TANG Jianrong, CAO Wenbin. Grey prediction model-based food security early warning prediction[C]// Proceedings of 2011 IEEE International Conference on Grey Systems and Intelligent Services (GSIS), Piscataway, United States, September 15-18, 2011. IEEE, 2011: 281 – 285.
[14] STEINBACH M, KUMAR V. Introduction to data mining[M]. 2th ed. Beijing: Post & Telecom Press, 2011: 202 – 208.
[15] 郭秀娟. 基于关联规则数据挖掘算法的研究[D]. 长春:吉林大学,2004.
[16] 田春元. 基于数据挖掘的食品安全风险评价与预警系统[D]. 青岛:青岛理工大学, 2012.
[17] 徐燕伟. 增量关联规则算法及其在食品安全监管中的应用[D]. 杭州:浙江大学,2008.
[18] 晁凤英, 杜树新. 基于关联规则的食品安全数据挖掘方法[J]. 食品与发酵工业, 2007, 33(4): 107 – 109.
CHAO Fengying, DU Shuxin. Data mining technics for food safety based on association rules[J]. Food and Fermentation Industries, 2007, 33(4): 107 – 109.
[19] 罗艳,文锡梅,谭红. 基于改进型 AHP 的食品质量安全时间序列预警模型的研究[J]. 贵州科学, 2012, 30(6):35 – 39.
LUO Yan, WEN Ximei, TAN Hong. A study on time series early warning model of food quality safety based on improved AHP[J]. Guizhou Science, 2012, 30(6): 35 – 39.
[20] 刘文. 食品安全指数的构建及应用[D]. 武汉:华中农业大学, 2013.
[21] 黄驱冥. 多维量化关联规则在食品安全检测中的应用[D]. 杭州:浙江工业大学,2007.

Application of Improved Association Rules on Food Safety Early Warning

XIAO Kejing[1], ZUO Min[1,*], WANG Xingyun[1], LIU Ting[2]
(1. *School of Computer and Information Engineering, Beijing Technology and Business University, Beijing* 100048, *China*;
2. *National Institutes for Food and Drug Control, Beijing* 100050, *China*)

Abstract: In order to the effective application of the massive detection data in food safety early warning, this paper analyzed the characteristics of the food detection data, and the insufficient of traditional Apriori algorithm on food detection data, then proposed the filtering algorithm, which is a pre-components of Apriori algorithm. An early warning model was established, which was applied to excavate the real oil detection data, and the potential safety problems were founded to make an early warning. Compared with the Apriori algorithm, the improved algorithm abandoned a lot of pseudo-association rules, and also could effectively enhance the efficiency and accuracy of food safety early warning, which has a very important practical significance.

Keywords: association rules; frequent item sets; sparse data; filtering algorithm

(责任编辑:檀彩莲)