马高庭,蒋万春,申艳光. 基于关联规则的肉鸡产品质量安全预警模型[J]. 江苏农业科学,2015,43(3):271 – 274.
doi:10.15889/j.issn.1002 – 1302.2015.03.089

基于关联规则的肉鸡产品质量安全预警模型

马高庭[1], 蒋万春[2], 申艳光[1]
(1.河北工程大学信息与电气工程学院,河北邯郸 056038; 2.河北工程大学农学院,河北邯郸 056038)

摘要:针对肉鸡生产过程中的安全问题,基于改良关联规则挖掘算法(APTPPA)建立肉鸡产品质量安全预警模型。该模型以肉鸡养殖及屠宰过程中危害分析、关键控制点(HACCP)的异常数据为处理对象,采用模式指导树并行频繁项集挖掘算法(APTPPA),构造关联路径树,找寻最大频繁项集,提取预警关联规则,挖掘影响肉鸡产品安全的因素,通过试验验证预警模型的有效性。

关键词:肉鸡产品;质量安全;预警模型;关联规则;APTPPA;HACCP
中图分类号: TS207.7　　**文献标志码**: A　　**文章编号**:1002 – 1302(2015)03 – 0271 – 04

食品安全问题的频繁发生,引起了众多国家的高度重视[1]。发达国家早已开始研究构建一套广泛有效的食品安全预警模型。畜禽产品在日常养殖、加工过程中面临更多更复杂的安全风险,监管难度很大。因此国内外学者较为关注对畜禽产品质量安全预警模型的探讨和研究。我国肉鸡产业发展迅速,但产品品质参差不齐。如不及时改善产品质量,提高预警能力,国内肉鸡产业将难以抗衡外来企业[2]。

数据挖掘在食品安全领域的应用较少,而食品安全日常事务所产生的大量时序数据非常适合做数据分析,从中可挖掘出有效的预警条目[3]。选择合适、高效的挖掘算法对食品安全预警模型的精确度至关重要。本研究采用的关联规则挖掘算法最早由 Agrawal 等提出[4],其中以 Apriori 算法最为经典[5],后续学者提出的改进算法大多以 Apriori 算法为基础。由于 Apriori 算法存在固有缺陷,随后 Han 等提出基于 FP – tree 来生成频繁项目集的 FP – growth 算法[6]。近些年其他类型的关联规则挖掘算法也相继问世[7,8],明显进步于早期算法,但在食品安全领域的适用性并不理想。肉鸡养殖、屠宰的安全因素具有多值性、倾斜性、稠密性和负相关性等特点,使传统挖掘算法构建预警模型变得尤为困难。本研究针对食品安全因素的固有问题,结合 HACCP 管理体系,采用 Association Path Tree Pattern Parallel Algorithm(APTPPA)算法构建了肉鸡产品质量安全预警模型。

收稿日期:2014 – 05 – 01
基金项目:国家自然科学基金(编号:61075053);河北省自然科学基金(编号:G2014402027)。
作者简介:马高庭(1990—),男,浙江绍兴人,硕士研究生,主要从事数据挖掘研究。E – mail:magaoting@ msn.com。

1　肉鸡产品质量安全预警模型框架

本研究的预警模型是肉鸡产品质量控制与可追溯系统中的一个模块。该系统基于 B/S 架构,囊括肉鸡产品安全信息

其他拟除虫菊酯类农药和环境污染物的检测提供了借鉴。

参考文献:

[1] Mehler W T, Li H Z, Lydy M J, et al. Identifying the causes of sediment – associated toxicity in urban waterways of the Pearl River Delta, China[J]. Environmental Science & Technology, 2011, 45(5):1812 – 1819.
[2] Li H Z, Mehler W T, Lydy M J, et al. Occurrence and distribution of sediment – associated insecticides in urban waterways in the Pearl River Delta, China[J]. Chemosphere, 2011, 82(10):1373 – 1379.
[3] Shafer T J, Meyer D A, Crofton K M. Developmental neurotoxicity of pyrethroid insecticides: critical review and future research needs[J]. Environmental Health Perspectives, 2005, 113(2):123 – 136.
[4] Garey J, Wolff M S. Estrogenic and antiprogestagenic activities of pyrethroid insecticides[J]. Biochemical and Biophysical Research Communications, 1998, 251(3):855 – 859.
[5] Wijngaarden R P A V, Brock T C M, Brink P J V D. Threshold levels for effects of insecticides in freshwater ecosystems: a review[J]. Ecotoxicology, 2005, 14(3):355 – 380.
[6] 谢　海. 双水相萃取技术研究现状[J]. 化学工业与工程, 2006, 23(5):463 – 466.
[7] 邓凡政, 魏　迎, 陈　影, 等. 双水相体系中 Cu(Ⅱ), La(Ⅲ), U(Ⅵ), Ce(Ⅳ)光谱行为及萃取分离[J]. 光谱学与光谱分析, 2004, 24(12):1637 – 1639.
[8] 贾凤燕, 王文文, 刘振波, 等. 乙腈 – 无机盐 – 水双水相体系萃取 – 气相色谱法检测鱼肉中拟除虫菊酯[J]. 化学学报, 2012, 70(4):485 – 491.
[9] 刘　新, 尤民生, 廖金英, 等. 甲胺磷降解菌的分离与降解效能测定[J]. 武夷科学, 2001, 17(1):51 – 55.
[10] Tallur P N, Megadi V B, Ninnekar H Z, et al. Biodegradation of cypermethrin by *Micrococcus* sp. strain CPN 1[J]. Biodegradation, 2008, 19(1):77 – 82.
[11] Wang B Z, Ma Y, Zhou W, et al. Biodegradation of synthetic pyrethroids by *Ochrobactrum tritici* strain pyd – 1[J]. World Journal of Microbiology and Biotechnology, 2011, 27(10):2315 – 2324.
[12] Chen S H, Hu Q B, Hu M Y, et al. Isolation and characterization of a fungus able to degrade pyrethroids and 3 – phenoxybenzaldehyde[J]. Bioresource Technology, 2011, 102(17):8110 – 8116.

的监测、分析和追溯，能够挖掘溯源数据库中的异常数据，比对专家和历史数据库，生成有效的预警信息，并及时发出警报。肉鸡产品质量安全预警模型包括信息源、比对源、挖掘分析以及预警反馈 4 个模块。预警模型框架见图 1。

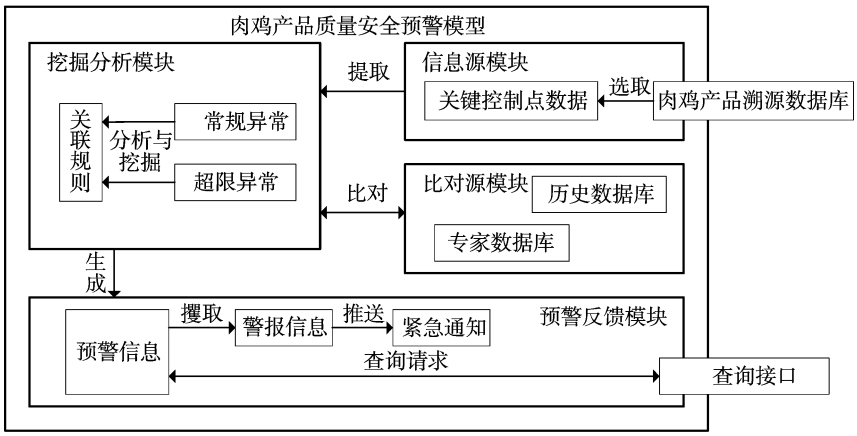


图1　肉鸡产品质量安全预警模型框架

信息源模块是预警模型数据的来源，以肉鸡溯源系统在肉鸡养殖、生产环节所收集的数据为基础，遵循 HACCP 体系，选取关键控制点中的记录进行预警挖掘。

比对源模块是专家数据和历史挖掘数据的数据源，在进行规则挖掘分析时，通常要与专家数据、历史数据对比，再得出挖掘规则。

挖掘分析模块是预警模型的核心，接收来自信息源的原始数据，经过对异常数据的分析，采用合适的关联规则挖掘算法，得出具有参考价值的规则，供下一个模块使用。

预警反馈模块是外部获得信息的窗口。当预警信息归类为紧急信息时，系统自动通知相关人员，即刻采取措施，避免造成食品安全事件和大规模损失。该模块还可供管理人员自主查询预警信息，从而提高预防能力，保证肉鸡产品质量，提高企业的行业竞争力。

2　肉鸡产品质量安全预警模型处理流程

肉鸡产品质量安全预警模型处理流程主要分为数据预处理、建立预警模型、挖掘结果检验 3 步[3]。预警模型详细处理流程如下：(1)进行数据预处理，并设置算法的支持度、置信度阈值。(2)利用关联规则挖掘算法搜索频繁项集。(3)对已找到的频繁项集进行剪枝操作。(4)判断是否完成频繁项集的搜索，若是则进入下一步，否则返回(2)。(5)根据找寻到的频繁项集生成关联规则，并在通过规则检验后更新预警数据库。

数据预处理主要是对异常数据进行逻辑转换和分类操作。逻辑转换针对监测数据为连续值的情况，连续值数据无法进行关联规则挖掘，因此要事先转换成逻辑值。分类是保证预警模型预警等级准确的前提，不同分类的异常数据后续处理方式也不同。根据提取食品安全预警事件特征的方法，可将异常数据分为常规异常和超限异常。

超限异常是指对于各项指标集合，具有影响食品安全状况的评价结果，它是最容易导致食品安全问题的因素[9]。

常规异常包括不规范异常、分布异常、趋势异常。(1)不规范异常。是指数据未按标准方式获得，具有不可信性，报警等级较低。(2)分布异常。通过区域的分布统计发现问题，将地区划分为 n 个区域，各区域内，超限异常总数为 $k_i(i\leqslant n)$，监测总数为 $l_i(i\leqslant n)$，异常情况数量均值 W_i 计算公式为：$W_i=Ck_i/l_i$(C 为常数)，当 W_i 超过预置阈值时，进行报警。(3)趋势异常。从历史数据库中分析得知，将历史数据划分为 n 个时间段，第 i 个时间段内出现的异常情况数量为 u_i，n 个时间段内出现的异常情况均值为 u^*，n 个时间段的中值为 i^*，趋势异常系数 r 计算公式为：$r=\dfrac{\sum_{i=1}^{n}[(u_i-u^*)(i-i^*)]}{\sqrt{\sum_{i=1}^{n}[(u_i-i^*)^2]\sum_{i=1}^{n}[(i-i^*)^2]}}$。趋势异常系数 r 与显著性指标 P 的关系如表 2 所示。P 由查询 t 分布函数得到，当 P 超过预置阈值时，进行报警。

表 2　趋势异常系数与显著性指标关系表

r	P	差异显著程度
$\dfrac{r}{1-r^2}\geqslant\dfrac{t_p(q-2)}{\sqrt{q-2}}$	≤0.01	极显著
$\dfrac{r}{1-r^2}\geqslant\dfrac{t_p(q-2)}{\sqrt{q-2}}$	≤0.05	显著
$\dfrac{r}{1-r^2}<\dfrac{t_p(q-2)}{\sqrt{q-2}}$	>0.05	不显著

建立预警模型就是把预处理后的异常数据采用 APTPPA 算法进行数据挖掘，找到频繁项集，抽取关联规则的过程。

挖掘结果检验即把新生成的预警规则与原有规则库进行对比，并分析实际预警效果。如果原有库中不存在该条规则，并且印证规则具有实际预警效果时，则将该规则更新到现有规则库中。

3　基于 APTPPA 算法的肉鸡产品质量安全预警模型

经典的 Apriori 算法在执行过程中会产生大量中间项集，必须多次扫描数据库，需要很多辅助空间结构，且要求数据为二值逻辑。本研究采用的 APTPPA 算法在压缩数据的同时保证了原始数据集的基本形态，使其在多值数据、倾斜数据和负关联规则的挖掘中比其他同类算法更加有效。APTPPA 算法主要包括关联路径树生成、频繁项集挖掘和寻找最大频繁项集 3 个步骤[3,10]。

3.1　关联路径树生成

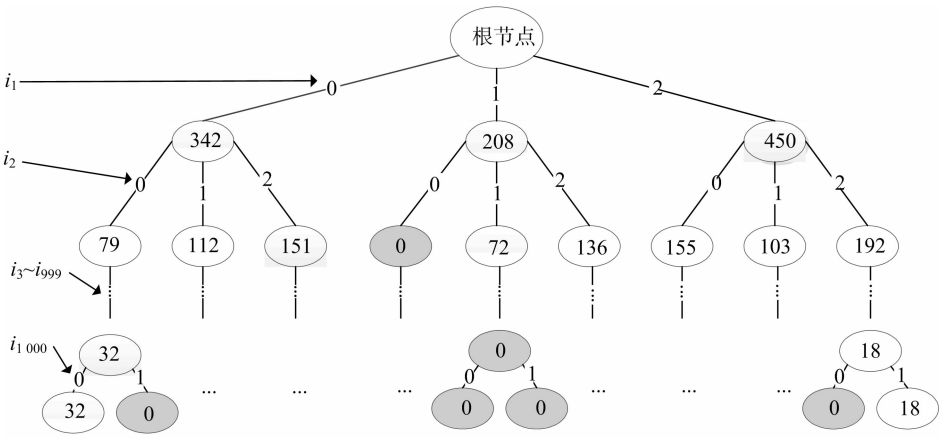
3.1.1　关联路径树的基本思想　将事务数据库 D 中每个数据项 i_m 均进行逻辑化处理会导致项数大量增加，造成维灾难。为了减少项数，将项值进行标号化处理，每类项值都用标号 v_n 表示。将标号化结果构造成树形结构就是关联路径树。以 1 000 组 15 项肉鸡超限异常数据为例，标号化的数据集 D 如表 3 所示。

数据集 D 进行逻辑化、标号化处理后，各项的值域显著减小，内部存在较多相同的事务数据。此时为数据集 D 增加 count 属性，对相同的事务数据进行统一计数，删除冗余，得到无重复数据的数据集 D'。由于没有冗余事务，每条事务 T_i 包含项集的一种取值构成最大项集，其支持度计数就是事务计数 count_i 的值。

3.1.2　构建基于树的路径表　数据集 D'中的每个事务都是项值的组合，D'中所有事务可构成 1 棵关联路径树，每个事务都是 1 条分支（图 2）。

表 3　数据集 D 标号化处理结果

序号	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}	i_{11}	i_{12}	i_{13}	i_{14}	i_{15}
1	2	1	2	1	1	3	1	0	2	0	0	0	2	3	0
2	1	0	1	1	1	3	0	0	0	0	0	3	1	3	0
3	1	2	0	1	1	3	1	0	0	1	1	2	2	3	0
4	1	1	1	2	0	2	0	3	1	0	1	2	1	1	0
5	0	2	1	2	1	2	0	0	2	1	2	3	3	2	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1 000	0	1	1	1	1	1	1	1	3	0	0	2	2	0	0



阴影部分不存在

图2　对数据集D'生成的关联路径树

根节点到第 1 层节点的路径为项 i_1 的取值，第 1 层节点到第 2 层节点的路径为项 i_2 的取值，以此类推。节点内的数字是经过该结点的事务计数。

基于上述关联路径树，可生成数据集 D'的关联路径表，如表 4 所示。

表 4　对数据集 D′生成的关联路径表

序号	路径	计数（count）
1	001002011200130	32
2	001003011000330	25
3	001110233300121	22
4	001200233300101	14
5	010000123310010	44
⋮	⋮	⋮
52	222110033230031	18

至此数据集 D 得到极大压缩，使数据在组合计数时尽可能一次性读入内存中。

3.2　频繁项目集挖掘

3.2.1　按模式指导求频繁项集　根据 Apriori 性质，可利用模式指导在关联路径树之上找寻出频繁项集。所谓模式即形如“xxooxxxxo”的某种排列组合，将事务中“x”位处项值忽略，而把“o”位处项值相同的事务计数并求和，就是该模式下的频繁项集及其计数。对于非倾斜数据，在“o”增加的同时，此模式下的事务计数会锐减，从而有效收敛。对于倾斜数据，事务计数原本大于支持度阈值，模式计数退化。此时为了保证算法的快速收敛，将包含全部项的频繁项集计数置零，再进行模式计数。还可通过设定最小支持度阈值对项集组合进行直接的剪枝操作。以 1 000 组 15 项肉鸡超限异常数据为例，找到的频繁项集如表 5 所示。

3.2.2　并行递归求频繁项集　m 个项有 m 个 1－“o”模式（k－“o”模式指包含频繁 k 项集的模式）的初始项集。并行递归就是在关联路径树上以 1－“o”模式为起始条件递归生成其他模式的方法。单 CPU 时，所有模式按 1－“o”模式的生成过程逐个递归完成。多 CPU 时，每个 CPU 分配 1 个 1－“o”模式，显著提高递归速度。当事务计数小于递归阈值时，递归终止，算法收敛有效。

3.3　寻找最大频繁项集

为了使挖掘结果更有意义，有必要在挖掘过程中剔除相似关联规则，防止重复规则出现。寻找最大频繁项集是剔除

表 5　挖掘出的频繁项集

序号	路径	计数(count)
1	00xxxxxxxxxxxx	402
2	001xxxxxxxxxxx	333
3	0011xxxxxxxxxx	218
4	00111xxx1xxxxx	167
5	0011x1xx0xxxxx	78
⋮	⋮	⋮
301	xxxxxxxxxxxx1x	590

相似关联规则的一条途径。对于 APTPPA 算法而言,在模式指导树上取路径 a 与其他任意路径 b 进行比较,当 a 的"o"位包含于 b 中时,把 b 赋值给 a,重复上述过程,直到不能发现路径 b 为止。以 1 000 组 15 项肉鸡超限异常数据为例,挖掘出的最大频繁项集如表 6 所示。

表 6　挖掘出的最大频繁项集

序号	路径	计数(count)
1	001xxx0xxxxxxx	234
2	001x1xxxxxxxxx	201
3	00xx1xxxxxxxxx	194
4	00x0xxxxxxxxxx	166
5	00xx1x1x1xxxxx	253
⋮	⋮	⋮
182	xxxxxxxxxx1xx1	265

4　试验与分析

抽取河北某食品公司肉鸡产品溯源数据库中的 1 000 组 15 项历史超限异常数据,在 Windows 7 操作系统下,采用 Java 编程语言,通过 Eclipse 集成平台,验证预警模型的有效性。将采用 APTPPA 与 Apriori 算法的肉鸡预警模型进行对比试验,验证 APTPPA 算法在食品安全预警领域的应用具有高效性。

4.1　基于 APTPPA 算法的挖掘结果及分析

试验参数设置如下:最小支持度 = 0.3,最小置信度 = 0.8,最大标号数 = 4,最大规则数 = 500。试验后从中选取 3 条报警记录如表 7 所示。

表 7　APTPPA 算法挖掘的最大关联规则

序号	最大关联规则	置信度
1	{$A1$ = (100,200],$A2$ = 4.0],$A3$ = (5.0,9.0], $A13$ = (250,450],$A14$ = 5],$A15$ = n} {$A4$ = 6]}	1.0
2	{$A1$ = (200,300],$A2$ = (3.5,4.0],$A4$ = 6],$A7$ = (4.2,4.3],$A8$ = (8.5,$A14$ = 5],$A15$ = y} {$A5$ = (12,15]}	1.0
3	{$A1$ = (100,200],$A4$ = 6],$A5$ = (12,15],$A13$ = 450],$A14$ = 5],$A15$ = n} {$A2$ = (3.0,3.5]}	1.0

将上述最大关联规则与历史超限异常数据进行检验,匹配度达 80% 以上,超标报警也较为准确,体现了本研究预警模型的有效性。由以上最大关联规则可分析出肉鸡养殖、屠宰加工过程中的安全隐患因素,主要有:肉鸡养殖环境中氨气水平、可吸入颗粒物同时超标,需要对栋舍进行清理;养殖用水中氯化物、硝酸盐同时超标,需要对水质进行改良;屠宰车间中氧气浓度、氨气水平同时超标,需要对屠宰车床进行消毒。

4.2　APTPPA 与 Apriori 算法挖掘效率的分析

为了验证 2 种挖掘算法的预警效率,采用上述 1 000 组 15 项超限异常数据分别测试 APTPPA 和 Apriori 算法预警挖掘的速度和精度并进行比较,在相同参数设置下,比较结果如表 8 所示。

表 8　APTPPA 与 Apriori 算法预警效率对比表

算法	速度(ms)	精度(%)
APTPPA	16 454	88.71
Apriori	62 342	85.33

由表 8 可知,在相同的规则覆盖率下,APTPPA 算法产生的规则更少,速度更快,效率更高。Apriori 算法没有结合食品安全预警信息的特点,产生较多冗余和不符合实际情况的规则。综上所述,在肉鸡产品安全预警时,基于 APTPPA 算法的肉鸡产品质量安全预警模型比传统 Apriori 算法预警模型更加有效。

5　总结与展望

基于关联规则的肉鸡产品质量安全预警模型采用了 APTPPA 算法,该算法能够在海量复杂多变的影响因素中,挖掘出导致肉鸡产品质量安全问题的要素,及时发现肉鸡养殖、屠宰、加工过程中的安全隐患并预警,在实时监控的同时有效减少和消除食品安全事故。但本研究的预警模型尚有不足,仍需进一步改进,主要体现在以下几方面:关键控制点囊括的异常因素不够全面;异常因素之间没有主次之分;逻辑值分类转换过程中没有用到较为准确的分类算法等。

参考文献:

[1]李　倩,张圣忠,王　芳. 基于博弈分析的食品安全风险监管策略研究[J]. 江苏农业科学,2013,41(9):268 - 270.

[2]赵金石. 我国肉鸡质量追溯系统应用现状分析[J]. 中国畜牧杂志,2011,47(8):45 - 48.

[3]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.

[4]Agarwa R,Imielinski T,Swmai A. Mining association rules between sets to items in large databases[C]. Porceedings of ACM SIGMOD Int'l Conf Management of Data,Washington DC,1993:207 - 216.

[5]Agrawal R,Srikant R. Fast algorithms for mining association rules in large databases[C]. 20th International Conference on Very Large Data Bases,1994:478 - 499.

[6]Han J,Pei J. Mining frequent patterns without candidate generation [C]. Porc 2000 ACM IGMOD Int Conf on Management of Data, SIGMOD'2000,Dallas TX,2000:1 - 12.

[7]宋　威,杨炳儒,徐章艳,等. 基于索引数组与集合枚举树的最大频繁项集挖掘算法[J]. 计算机科学,2007,34(7):146 - 149.

[8]王黎明,张　卓. 基于 iceberg 概念格并置集成的闭频繁项集挖掘算法[J]. 计算机研究与发展,2007,44(7):1184 - 1190.

[9]翁道磊. 食品安全追溯系统的分析和研究[D]. 重庆:重庆大学,2008.

[10]张大为,黄　丹,嵇　敏,等. 利用模式指导树的并行频繁项集挖据方法[J]. 计算机工程与应用,2010,46(22):147 - 150.