研究报告

基于Apriori算法的小麦中多组分真菌毒素污染的关联规则挖掘

薛文博[1,2],王小丹[2],李明璐[2],唐昊[1,2],马宁[2],张磊[2],梁江[2],祝海江[1]
(1. 北京化工大学,北京 100029;2. 国家食品安全风险评估中心,北京 100021)

摘 要:目的 为分析主要影响小麦中多组分真菌毒素污染指标之间的关联性,研究不同毒素之间的共污染关系特征。**方法** 采用关联规则Apriori算法对小麦中多组分真菌毒素的污染监测指标之间的关联性进行数据挖掘分析。根据指标检测值对数据进行风险等级划分,构造出布尔类型的事务数据库,对事务数据库进行频繁项集挖掘,设置阈值最小支持度和最小置信度,迭代重复执行连接、剪枝操作确定出频繁项集,获取强关联规则。通过置信度、支持度、提升度等进行关联规则评价,最后将数据可视化应用在关联规则中,更直观地对规则进行展示与验证。**结果** 挖掘得到小麦中多组分真菌毒素之间共污染的潜在强关联规则,包括9条单项集共污染毒素强关联规则以及多条组合项集强关联规则。分析验证得到脱氧雪腐镰刀菌烯醇和雪腐镰刀菌烯醇、玉米赤霉烯酮和脱氧雪腐镰刀菌烯醇毒素之间存在共污染关系,置信度分别为92.0%和80.6%。**结论** 通过数据挖掘得到的强关联规则对小麦毒素风险预警和防控有一定的意义,同时为多毒素联合暴露评估提供依据。

关键词:小麦;真菌毒素;数据挖掘;关联规则;数据可视化
中图分类号:R155 **文献标识码:**A **文章编号:**1004-8456(2022)03-0451-08
DOI:10.13590/j.cjfh.2022.03.010

Association rule mining of multicomponent mycotoxins contamination in wheat based on Apriori algorithm

XUE Wenbo[1,2], WANG Xiaodan[2], LI Minglu[2], TANG Hao[1,2], MA Ning[2], ZHANG Lei[2], LIANG Jiang[2], ZHU Haijiang[1]
(1. Beijing University of Chemical Technology, Beijing 100029, China; 2. China National Center for Food Safety Risk Assessment, Beijing 100021, China)

Abstract: Objective To analyze the correlation of muti-mycotoxin contamination in wheat, the co-contamination characteristics of different mycotoxins were studied. **Methods** Data mining analysis of the association between monitoring data for multiple mycotoxins contamination in wheat was performed using the association rule Apriori algorithm. Boolean data type of transaction database was constructed according to the pollutant index values to risk hierarchy structure to mine frequent item sets of transaction database. To determine frequent item sets and obtain strong association rules, minimum threshold support and minimum confidence was set, and iterative connection and pruning operations were performed repeatedly. Association rules were evaluated by confidence, support and promotion degree, etc. Finally, data visualization was applied to association rules to display and verify rules more intuitively. **Results** The potential strong association rules of co-contamination of muti-mycotoxins in wheat were found, including 9 strong association rules of single common contamination toxin and several strong association rules of combined term sets. The co-pollution relationship between deoxynivalenol and nivalenol, zearalenone and deoxynivalenol was analyzed and verified. The confidence was 92.0% and 80.6%, respectively. **Conclusion** The strong association rules obtained by data mining have certain significance for the early warning, prevention and control of wheat toxin risk, which provides basis for the assessment of combined exposure to multiple toxins.

Key words: Wheat; mycotoxin; data mining; association rules; data visualization

收稿日期:2022-03-15
基金项目:国家重点研发计划(2019YFC1606500);国家食品安全风险评估中心"高层次人才队伍建设523项目"
作者简介:薛文博 男 硕士研究生 研究方向为数据挖掘 E-mail:540615266@qq.com
通信作者:梁江 女 研究员 研究方向为食品安全风险评估 E-mail:liangjiang@cfsa.net.cn
祝海江 男 教授 研究方向为图像处理及计算机视觉 E-mail:zhuhj@mail.buct.edu.cn

我国是世界最大的小麦主产国和消费国。小麦作为仅次于水稻的第二大粮食作物，其质量安全直接关系到广大消费者的健康。其中，小麦籽粒在生长及收获储存过程中可能受各种产毒真菌的污染，是影响小麦农作物质量安全的重要因素。小麦中的真菌毒素污染不仅会使小麦产量和品质降低并带来巨大的经济损失，更严重的是人类和动物摄入被污染的小麦还会威胁人畜的生命健康。因此研究小麦中真菌毒素的污染规律情况对于小麦安全风险防控有重要意义。

数据挖掘一般是指从大量的随机数据中利用算法挖掘产生数据内部某些特定隐含信息的过程[1]。目前对小麦中多毒素污染的表征分析多采用传统的统计学方法以及数学建模的方式，通过对监测数据统计计算分析获取不同毒素之间的相关性，如吴本刚等[2]对小麦中的呕吐毒素和玉米赤霉烯酮的相关性分析方法得到两种毒素协同污染的线性相关关系。而关联规则挖掘技术在食品预警与分析领域也有一些新的应用，LI等[3]提出了一种传统评估预警方法与数据挖掘技术相结合的思路，通过构建科学合理的风险评估指标体系来计算获取评估结果，并且对结果进行关联规则挖掘分析以获得安全预警信息，进而为监管机构提供相应的提示信息。CHEN等[4]通过收集大量的食品中污染物检测数据，以大豆中污染物数据为例，运用关联规则技术构建预警模型来帮助监管机构进行决策分析。BU等[5]着重研究食品安全案例中的数据信息，运用文本处理技术提取相关食品案件中的关键词汇，主要针对的是案例中饺子和包子等面食中的食品添加剂硫酸铝铵、瘦肉精以及时间和地区信息的挖掘，利用关联规则Apriori算法分析案例内部规律进行参考。ZHONG等[6]介绍了在面对食品安全大数据时应当采取不同的方式处理海量、多元化的数据，以及面对不同数据源如何采用高效处理算法提高数据的处理能力。WANG等[7]针对供应链上的乳品生产商数据同样采取数据挖掘技术和物联网技术建立预警模型并设计了食品安全预警系统，从海量数据中挖掘搜集食品安全风险信息和内在的关联规则，识别出风险进而做出事前预警提示，帮助食品企业管理者提前干预降低风险。宗万里等[8]针对山东食药局官方抽检数据对重金属、农兽药残留、食品添加剂、微生物等不同危害因素类别分析，挖掘得到部分与谷氨酸钠、氯化钠、菌落总数和重金属等信息相关的预警规则。顾小林等[9]针对猪肉生产加工过程中的评价指标作为输入项，对猪肉加工过程中的色度、氯化物、硫酸、氟化物等指标进行关联规则挖掘分析预警，构建出预警模型输出强关联规则做相关提示预警。晁凤英等[10]对食品安全检测数据库中的数据预处理，筛选出食品种类、产地、检测时间等预警指标，利用关联规则技术挖掘出与三个预警风险等级有关的强关联规则进行分析预警。白宝光等[11]建立了乳制品安全预警指标体系，筛选出影响乳品安全的预警指标，包括奶牛养殖环节以及乳品制作各过程中的指标作为神经网络的输入依据，如饲料质量合格率、兽药质量合格率等预警指标，把乳制品抽检合格率作为输出指标。通过神经网络训练，对最终的乳品抽检合格率做回归预测，构建出乳品质量预警模型，从各环节的合格率指标预测最终乳品的合格率。

本研究首次将数据挖掘中的关联规则Apriori算法应用在小麦多组分真菌毒素的相关性分析与预警，基于对监测数据的离散化处理和风险等级划分，通过挖掘的强关联规则分析研究小麦中多毒素的联合污染表征，为小麦中真菌毒素风险预警分析和进一步开展多毒素累积暴露评估和风险防控提供参考依据。

1 材料与方法

1.1 数据来源

数据来源于全国范围内11个省份小麦中的伏马菌素（Fumonisins，FB）、黄曲霉毒素（Aflatoxins，AFT）、脱氧雪腐镰刀菌烯醇（Deoxynivalenol，DON）、雪腐镰刀菌烯醇（Nivalenol，NIV）、赭曲霉毒素A（Ochratoxin A，OTA）、玉米赤霉烯酮（Zearalenone，ZEN）、T-2（Trichothecenes-2 toxin，T-2）毒素等7种真菌毒素的监测数据库，包含监测年份、监测省份、样品产地、样品编号、样品名称、检测方法、检出限、检测值等信息。由于挖掘算法是对数据库进行频繁项集挖掘，首先需要对所有样品毒素数据检测值划分等级，划分规则是根据各毒素指标数据检测值将其降序排序后划分为三个污染等级区间。一、二级为检出值高的数据，三级为低值数据和未检出数据。其中对不同检测方法的部分未检出数据以检出限一半值代替，并将其划分为三级低污染区间，重点关注一级高污染区间的指标频繁项，降低未检出三级污染区间的数据对整体研究的影响。Apriori算法是针对数据库中样品指标出现的频次计数，即算法执行过程需要对全部样品中各毒素指标出现的次数计数，因此该过程中处理所有数据保留生成新的布尔类型数据。

1.2 Apriori算法基本执行流程

Apriori算法首先扫描预处理后的污染物指标数据库，通过搜索扫描数据库确定出满足最小支持

度的频繁 1-项集，其中最小支持度的取值一般依照数据库中项集比例主观赋值 0-1 之间的数。频繁 1-项集再与自身执行连接操作产生候选 2-项集，将候选项集的支持度计数与最小支持度比较执行剪枝操作，确定出频繁 2-项集。不断搜索迭代直到不产生频繁项集为止，运用 Apriori 数据挖掘算法找出事务数据集之间隐藏的强关联规则。

算法执行流程图如图 1 所示。

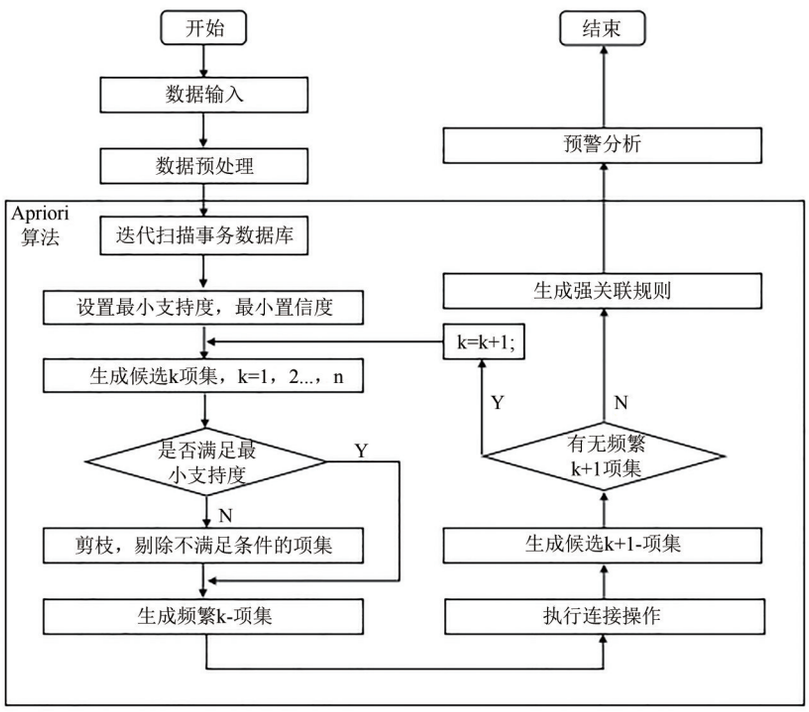


图 1　Apriori 算法流程图

Figure 1　Flow chart of Apriori algorithm

1.3　基于 Apriori 算法对小麦中多种真菌毒素污染的关联规则挖掘方法体系构建

1.3.1　构造事务数据库

事务数据库 $D=\{T_1,T_2,\cdots T_n\}$ 是数据库中所有事务构成的集合。事务数据库中的每个记录代表一个事务 T，也称为 *TID*。组成事务的各个属性称为项，项的集合为项集。例如集合{OTA，NIV，DON}称为 3-项集。设事务数据库中所有项的集合为项集 I，则事务数据库中每条记录事务 T 均为项集 I 的子集，即 $T\subseteq I$。若 X，Y 为事务中的项集，$X\subset I, Y\subset I, X\neq\varnothing, Y\neq\varnothing, X\cap Y=\varnothing$，则关联规则 $X\Rightarrow Y$ 成立，表示 X 中的项目出现时，Y 中的项目也跟着出现的规律。例如：

$X\Rightarrow Y$,(*support*,*confidence*,*lift*,*conviction*)

本研究首先基于小麦真菌毒素执行的国家标准《食品安全国家标准 食品中真菌毒素限量》(GB 2761—2017)中相关真菌毒素指标及相关领域的专家咨询意见，筛选出 7 种真菌毒素污染指标对其规律进行研究。该标准中对小麦中黄曲霉毒素、玉米赤霉烯酮、赭曲霉毒素 A、脱氧雪腐镰刀菌烯醇等真菌毒素指标限量。将 1 540 个小麦样品中 FB、AFT、DON、NIV、OTA、ZEN、T-2 毒素的污染水平抽检数据进行区间污染等级划分，即将 7 种毒素的污染水平按照检测值降序划分为三等分，分别对每种指标毒素检测值降序排序，样品总数为 1 540，三等分为 513 个样品。将检测值高的 513 个毒素检测值划分为一级污染区间，将检测值处于中间的 513 个毒素检测值划分为二级污染区间，将其余毒素检测值低的部分划分为三级污染区间。每种毒素指标取两个阈值分为三个区间构造出污染等级划分表，将样品数据根据污染等级划分表进行离散化处理生成布尔类型的数据，构造出不同毒素指标的事务数据库。为消除未检出数据对整体指标的影响，根据每个指标数据检测值降序标准划分三等分区间，在污染等级划分时将数据中高于检出限的数据划分为一级和二级污染区间，将低于检出限的部分以及检出值较低的部分数据化为三级污染区间。同时考虑算法挖掘时生成的一级高污染区间规则对毒素共污染预警的意义，因此本研究重点关注一级高污染区间内各毒素的共污染情况。

每个样品同时包含七项污染物指标。每一行表示一个样品，每一列表示指标污染物等级水平。事务数据库某一行样品 FB(三)、AFT(三)、DON(一)、NIV(三)、OTA(三)、ZEN(二)、T-2(三)代表样

表 1 污染等级划分表

Table 1 Pollution levels partition

检测值/(μg/kg)	一级污染	二级污染	三级污染
FB	(30, +∞]	(7.5, 30]	(0, 7.5]
AFT	(1.5, +∞]	(0.5, 1.5]	(0, 0.5]
DON	(100, +∞]	(20, 100]	(0, 20]
NIV	(17.5, +∞]	(5, 17.5]	(0, 5]
OTA	(2, +∞]	(1, 1.5]	(0, 1]
ZEN	(2.5, +∞]	(1.5, 2.5]	(0, 1.5]
T-2	(15, +∞]	(5.5, 15]	(0, 5.5]

品中伏马菌素处于三级污染区间,黄曲霉毒素处于三级污染区间,脱氧雪腐镰刀菌烯醇处于一级污染区间,雪腐镰刀菌烯醇处于三级污染区间,赭曲霉毒素 A 处于三级污染区间,玉米赤霉烯酮处于二级污染区间,T-2 毒素处于三级污染区间。

构造出的事务数据库部分样品数据如表 2 所示。

表 2 事务数据库表

Table 2 Transaction database

编号	FB	AFT	DON	NIV	OTA	ZEN	T-2
1420	FB(三)	AFT(二)	DON(一)	NIV(三)	OTA(三)	ZEN(二)	T-2(三)
1421	FB(三)	AFT(一)	DON(一)	NIV(三)	OTA(三)	ZEN(一)	T-2(三)
1422	FB(三)	AFT(三)	DON(三)	NIV(三)	OTA(三)	ZEN(二)	T-2(三)
1423	FB(三)	AFT(三)	DON(一)	NIV(三)	OTA(一)	ZEN(二)	T-2(三)
1424	FB(三)	AFT(三)	DON(一)	NIV(三)	OTA(三)	ZEN(二)	T-2(三)
1425	FB(三)	AFT(三)	DON(一)	NIV(三)	OTA(三)	ZEN(二)	T-2(三)

1.3.2 Apriori算法执行过程

关联规则目的是找出事务数据库中事务数据之间的关联关系。通过对事务数据库中的数据进行算法挖掘,寻找出事务数据库中潜在的一些关联规则。采用 Apriori 算法对数据预处理后生成的污染物指标数据库进行关联性挖掘,获取置信度较高的关联规则,分析样品中污染指标之间关联性。

算法执行过程如下:

(1) 首次迭代扫描事务数据库中的 1-项集,分别对 7 种真菌毒素污染指标的三个等级区间出现频次计数,包括{FB(一)}、{FB(二)}、{FB(三)}、{AFT(一)}、{AFT(二)}、{AFT(三)}、{DON(一)}、{DON(二)}、{DON(三)}等共计 21 个 1-项集,统计其支持度计数,生成候选 1-项集。

(2) 执行剪枝操作。剔除不满足最小支持度计数的候选 1-项集,保留其余满足最小支持度计数的候选 1-项集生成频繁 1-项集集合。设置最小支持度为 0.1,样品总数为 1 540,最小支持度计数为 154。21 个 1-项集均满足最小支持度计数,保留生成频繁 1-项集。

(3) 进行连接操作。将频繁 1-项集与自身进行连接生成候选 2-项集。例如{FB(一)、AFT(一)}、{FB(一)、AFT(二)}、{FB(一)、AFT(三)}、{FB(一)、DON(一)}、{FB(一)、DON(二)}、{FB(一)、DON(三)}等候选 2-项集。

(4) 再次迭代遍历事务数据库,统计候选 2-项集支持度计数,剔除不满足最小支持度计数的项集,保留满足最小支持度的项集,得到频繁 2-项集。例如{ZEN(一)、DON(一)}、{AFT(一)、DON(一)}、{NIV(一)、DON(一)}、{ZEN(一)、AFT(一)}、{OTA(一)、AFT(一)}、{T-2(一)、DON(一)}、{T-2(一)、ZEN(一)}、{FB(三)、AFT(三)}、{NIV(三)、AFT(三)}、{AFT(三)、OTA(三)}等频繁 2-项集。

(5) 反复执行连接,剪枝操作获取频繁 N-项集,直到不再 产生频繁项集为止。执行连接过程中遵守频繁项集的所有子集一定为频繁项集,剔除非频繁项集。

算法伪代码如下:

输入:D:事务数据库

min_sup:最小支持度阈值

输出:L,D 中的频繁项集。

```
1  L1 = find_frequent_1_itemsets(D);
2  for(k=2; Lk-1≠null;k++) {
3      Ck = apriori(Lk-1);
4      for each T in D {
5          Ct = subset(Ck, T); // 获取事务 T 的候选子集
6      for each 候选 c in Ct
7          c. count++; // 支持度计数
8  }
9  L1 = {c(Ck | c. count >= min_sup )}
10 }
11 return L = ∪k Lk;//返回候选项集不小于最小支持度的频繁项
```

```
def apriori (Lk-1 e frequent(k-1) itemset)
1  for each 项集 ee in Lk-1
2      for each 项集 l2 in Lk-1
3      if(l1[1]= l2[1])∧ … ∧(l1[k-2] = l2[k-2])∧ … ∧(l1[k-1] = l2[k-1])) then{
4          c = l1⋈l2 //连接步:产生候选
5      if has_infrequent_subset(c, Lk-1) then
6          delete c;//剪枝步:删除非频繁的候选
7      else add c to Ck;
8  }
9  return Ck;
```

```
def has_ infrequent_subset (c : candidate k itemset ;
```

L_{k-1}e frequent(k-1) itemset)

1 for each(k-1)subset s of c

2 if s not in L_{k-1}

3 return TRUE;

4 return FALSE;

实验环境为 Windows 7，编程语言为 Python，集成软件为 PyCharm。实验调用 Python 中“efficient_apriori”包，实验参数设置最小支持度=0.1，最小置信度=0.5。支持度过低挖掘出规则较多且指标项集占比过低，挖掘出无用的规则也会占很大一部分；支持度过高会导致挖掘不到强关联规则。参数设置采用穷举搜索法（exhaustive search）遍历搜索最优参数，选取针对本数据集适合的最小支持度与最小置信度。最小支持度设置 0.1，即在剪枝过程中保留的项集大于最小支持度计数，强关联规则中指标项集同时出现的样品数占全部样品数比例大于 0.1。最小置信度为 0.5，即规则前项发生时规则后项发生的概率。

其中，支持度：X 和 Y 中所包含的项在事务集 D 中同时出现的概率。

$$support(X \Rightarrow Y) = p(XY)$$

置信度：包含 X 的事务中又包含 Y 的比例。

$$confidence(X \Rightarrow Y) = p(Y|X)$$

提升度：X 出现对 Y 出现概率发生的变化，值为 1 说明两者独立，值越大表明二者关联性越强。

$$lift(X \Rightarrow Y) = p(Y|X)/p(Y)$$

确信度：值为 1 时表明项集 X 和项集 Y 具有独立性，确信度越大说明该规则越可信。

$$conviction(X \Rightarrow Y) = (1 - p(Y))/(1 - p(Y|X)$$

强关联规则：关联规则既满足最小置信度，又满足最小支持度则为强关联规则。

频繁模式是在数据中频繁出现的模式。频繁项集作为频繁模式的一种，指的是频繁出现在事务数据库中项的集合，并且它在事务数据库中满足最小支持度。频繁项集的所有子集都是频繁项集。

2 结果

2.1 Apriori算法挖掘的关联结果

通过设置最小支持度与最小置信度，对 1 540 条样品数据进行 Apriori 算法挖掘可得到多条强关联规则，其中包含单项集毒素与多项集毒素。由于数据中存在部分未检出低值数据，所以重点分析一级污染区间多种毒素之间的潜在关系。通过整理挖掘出的单项集毒素强关联规则对两种毒素之间的共污染关系开展分析，主要得到单项集毒素的 9 条强关联规则及若干多项集强关联规则。

在满足最小支持度与最小置信度的条件下，将置信度与支持度高的关联规则排序，挖掘获取小麦中真菌毒素指标之间的强关联规则，其中 DON 与 NIV、OTA 与 AFT、ZEN 与 DON 为关联性最强的三条共污染规则。通过应用算法挖掘出频繁项集，得到 1-项集的单因素共污染毒素符合预警条件的前 9 条强关联规则如表 3 所示，多项集的规则如表 4 所示。根据九条关联规则构造出的置信度矩阵如表 5 所示。

表 3 1-项集共污染毒素强关联规则表

Table 3 1-itemsets pollution toxin strong association rules

规则	conf:置信度, supp:支持度, lift:提升度, conv:确信度
{NIV(一)} -> {DON(一)}	(conf: 0.920, supp: 0.157, lift: 2.319, conv: 7.555)
{OTA(一)} -> {AFT(一)}	(conf: 0.899, supp: 0.162, lift: 3.133, conv: 7.079)
{ZEN(一)} -> {DON(一)}	(conf: 0.806, supp: 0.130, lift: 2.033, conv: 3.117)
{ T-2 (一)} -> {DON(一)}	(conf: 0.742, supp: 0.106, lift: 1.870, conv: 2.339)
{ T-2 (一)} -> {ZEN(一)}	(conf: 0.715, supp: 0.103, lift: 4.439, conv: 2.943)
{ZEN(一)} -> {AFT(一)}	(conf: 0.673, supp: 0.108, lift: 2.346, conv: 2.183)
{ZEN(一)} -> { T-2 (一)}	(conf: 0.637, supp: 0.103, lift: 4.439, conv: 2.360)
{AFT(一)} -> {OTA(一)}	(conf: 0.566, supp: 0.162, lift: 3.133, conv: 1.887)
{AFT(一)} -> {DON(一)}	(conf: 0.505, supp: 0.145, lift: 1.272, conv: 1.218)

2.2 小麦中毒素污染的关联性分析结果验证

表 5 中点（NIV，DON）的矩阵高度值为 0.92 表示强关联规则 $NIV(一) \Rightarrow DON(一)$，*confidence* = 0.92 包含 NIV（一）的事务中又包含 DON（一）的概率，即在小麦真菌毒素污染的事务数据库中雪腐镰刀菌烯醇在一级污染区间时，脱氧雪腐镰刀菌烯醇有 92% 的概率也处于一级污染区间。

对挖掘得到的单因素共污染规则验证并对其部分规则进行讨论，图 2 依次对应表 3 中 9 条联合污染毒素的强关联规则毒素指标散点图，依次为 a、b、c、d、e、f、g、h、i。其中每个图的 X 轴代表规则前项毒素指标检测值，Y 轴代表规则后项毒素指标检测值。纵轴红线右侧代表横坐标一级污染区间，水平轴红线上侧代表纵坐标的一级污染区间。

{NIV（一）}->{DON（一）}（conf：0.920，supp：0.157，lift：2.319，conv：7.555）规则关联性强。由散点图 2a 很明显看出，当横坐标处于一级污染状态时，绝大多数点都处于纵轴预警上方。与挖掘规则

表4 部分多项集毒素强关联规则表

Table 4 Partial multi-item toxin strong association rule

规则	conf:置信度, supp:支持度, lift:提升度, conv:确信度
{FB(一), NIV(二)} -> { T-2 (二)}	(conf: 0.972, supp: 0.227, lift: 1.656, conv: 14.826)
{ T-2 (二), NIV(二)} -> {FB(一)}	(conf: 0.810, supp: 0.227, lift: 2.648, conv: 3.649)
{ T-2 (二), FB(一)} -> {NIV(二)}	(conf: 0.975, supp: 0.227, lift: 2.745, conv: 25.649)
{FB(二), OTA(一)} -> {AFT(一)}	(conf: 0.979, supp: 0.119, lift: 3.410, conv: 33.510)
{AFT(一), FB(二)} -> {OTA(一)}	(conf: 0.864, supp: 0.119, lift: 4.785, conv: 6.019)
{OTA(一)} -> {AFT(一), FB(二)}	(conf: 0.662, supp: 0.119, lift: 4.785, conv: 2.548)
{FB(二), OTA(一)} -> {AFT(一)}	(conf: 0.979, supp: 0.119, lift: 3.410, conv: 33.510)
{AFT(一), OTA(一)} -> { T-2 (二)}	(conf: 0.656, supp: 0.106, lift: 1.118, conv: 1.201)
{ T-2 (二), NIV(二)} -> {AFT(二)}	(conf: 0.840, supp: 0.235, lift: 2.483, conv: 4.133)
{ T-2 (二), AFT(二)} -> {NIV(二)}	(conf: 0.892, supp: 0.235, lift: 2.510, conv: 5.950)
{NIV(二)} -> { T-2 (二), AFT(二)}	(conf: 0.662, supp: 0.235, lift: 2.510, conv: 2.177)
{AFT(二)} -> { T-2 (二), NIV(二)}	(conf: 0.695, supp: 0.235, lift: 2.483, conv: 2.360)

表5 单项集置信度矩阵表

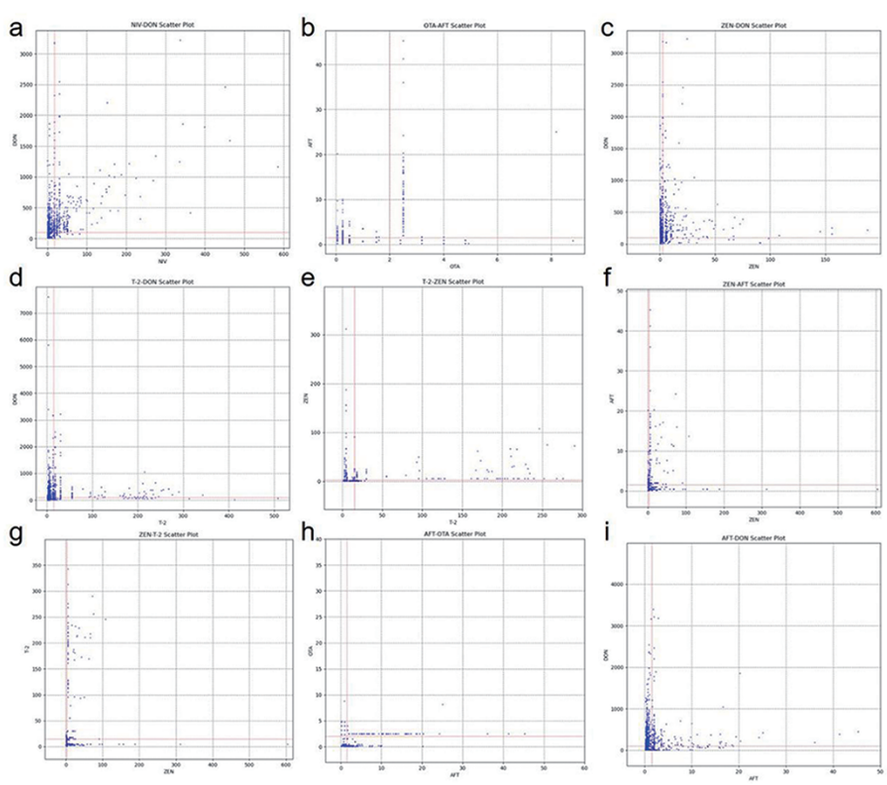
Table 5 Single set of confidence matrix

毒素类别	OTA	T-2	FB	DON	NIV	ZEN	AFT
OTA	0	0	0	0	0	0	0.899
T-2	0	0	0	0.742	0	0.715	0
FB	0	0	0	0	0	0	0
DON	0	0	0	0	0	0	0
NIV	0	0	0	0.920	0	0	0
ZEN	0	0.637	0	0.806	0	0	0.673
AFT	0.566	0	0	0.505	0	0	0

注:列指标表示单因素的强关联规则中的前项,记为X;行指标表示其规则后项,记为Y。其中,(X,Y)表示X对Y的置信度;(Y,X)表示Y对X的置信度

与已有文献毒素联合暴露评估结果相符。同时DON与NIV的协同污染正相关关系也已被多个文献证实[12-15]。如程天笑等[16]利用Spearman技术对四省小麦中的真菌毒素相关性分析得出NIV与DON有很强的正相关性。

{ZEN(一)}->{DON(一)}(conf:0.806, supp:0.130, lift:2.033, conv:3.117)对应散点图2c,样品中ZEN在一级污染区间时DON有80%的概率处于一级污染区间。根据现有的研究成果吴本刚等[2]的小麦中呕吐毒素和玉米赤霉烯酮与赤霉病粒关



注:a图X轴:NIV检测值,Y轴:DON检测值;b图X轴:OTA检测值,Y轴:AFT检测值;c图X轴:ZEN检测值,Y轴:DON检测值;d图X轴:T-2检测值,Y轴:DON检测值;e图X轴:T-2检测值,Y轴:ZEN检测值;f图X轴:ZEN检测值,Y轴:AFT检测值;g图X轴:ZEN检测值,Y轴:T-2检测值;h图X轴:AFT检测值,Y轴:OTA检测值;i图X轴:AFT检测值,Y轴:DON检测值;依次对应表3中9条单项集强关联规则每个样品中前后项毒素指标的检测值

图2 小麦中单因素毒素污染水平关联规则验证散点图

Figure 2 Scatter diagram for verifying association rules of single-factor toxin pollution level in wheat

系研究中的DON与ZEN相关性分析，2种真菌毒素是相伴产生存在较高的正相关关系；程天笑[17]对于小麦中多种真菌毒素协同污染调查得出ZEN和DON也有较强的正相关关系。与挖掘得到的ZEN与DON为共污染强关联规则相符合。

{AFT(一)}->{DON(一)}(conf：0.505，supp：0.162，lift：3.133，conv：1.887)对应散点图2i，当AFT处于一级污染区间时，对应DON的散点个数基本均分在水平轴上下两侧，与挖掘规则相符合。规则置信度最低为50.5%，在挖掘得到的单项集强关联规则中最低，与以上置信度最高的两条规则对比，同时毒素相关性显著低于以上两条规则。李星等[18]在对上海地区小麦中霉菌毒素的污染调查发现，AFT和DON并不存在明显的相关关系。

3 结论

本文通过数据处理等级划分模式，利用数据挖掘关联规则中的Apriori算法建立了对小麦中多种真菌毒素监测数据的关联分析方法体系，获取小麦中共污染的真菌毒素之间内在强关联规则，为进一步开展小麦多毒素共污染联合暴露风险评估及风险预警提供科学依据。有助于监测人员在检测出较高污染毒素检测值后评估与其有共污染关系的未检毒素含量污染水平范围，同时可以对一些样品中未测毒素指标数据产生预警效果。还可以利用一些相应管控措施对共污染毒素进行提前干预与治理。

本文将数据挖掘相关的技术应用在小麦多种真菌毒素的相关性分析中，并且基于客观监测数据利用无监督学习挖掘获取不同毒素间的关联规则，将机器学习的结果与现有相关研究成果分析比对，为多毒素相关性污染分析提供一种新的思路。利用数据分析技术对样品数据库分析处理获取指标关联性，相较于传统的统计分析方式可以基于数据基础利用数据挖掘算法模型更迅速、直观挖掘出毒素指标之间的相关性来分析其特点。后续可以针对不同类别食品建立相应的软件系统设置指标划分污染等级直接导入相关数据获取有用的强关联规则结果。

参考文献

[1] HAN J W，KAMBER M. 范明，孟小峰译. 数据挖掘概念与技术[M]. 机械工业出版社：计算机科学丛书，2012：1-468.
HAN J W，KAMBER M. Edited by FAN M，MENG X F. Data mining concepts and techniques[M]. China Machine Press：Computer Sciences，2012：1-468.

[2] 吴本刚，孙宝胜，徐存宽，等. 小麦中呕吐毒素和玉米赤霉烯酮与赤霉病粒关系研究[J]. 粮食与油脂，2017，30(7)：105-108.
WU B G，SUN B S，XU C K，et al. Study on relationship of DON content，ZEN content and *Gibberella* damaged kernels in wheat[J]. Cereals & Oils，2017，30(7)：105-108.

[3] LI L J，SHEN Y Y，YUAN Z Q，et al. The analysis and research on food safety risk assessment based on data mining[C]. 2015 International Conference on Intelligent Systems Research and Mechatronics Engineering，2015.

[4] CHEN K，TAN H，GAO J，et al. Application of association rule mining on food safety test data[J]. Applied Mechanics and Materials，2014，556-562：4681-4684.

[5] BU K，LI X L，WANG K，et al. Data analysis of public food safety cases based on Apriori[C]. 2020 Chinese Control and Decision Conference (CCDC). New York：IEEE Press，2020：343-348.

[6] ZHONG Y L，HE C H，TANG J Q. Research on the application of big data in smart food safety[C]. 2020 4th Annual International Conference on Data Science and Business Analytics (ICDSBA). New York：IEEE Press，2020：36-39.

[7] WANG J，YUE H L. Food safety pre-warning system based on data mining for a sustainable food supply chain[J]. Food Control，2017，73：223-229.

[8] 宗万里，朱习军. 基于Apriori算法的食品抽检数据的关联规则挖掘[J]. 食品安全质量检测学报，2020，11(4)：1334-1337.
ZONG W L，ZHU X J. Mining association rules of food sampling data based on Apriori algorithms[J]. Journal of Food Safety & Quality，2020，11(4)：1334-1337.

[9] 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 X L，ZHANG D W，ZHANG K，et al. The information pre-warning model of food safety based on association rules mining[J]. Soft Science，2011，25(11)：136-141.

[10] 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 F Y，DU S X. Data mining technics for food safety based on association rules[J]. Food and Fermentation Industries，2007，33(4)：107-109.

[11] 白宝光，朱洪磊，范清秀. BP神经网络在乳制品质量安全风险预警中的应用[J]. 中国乳品工业，2020，48(7)：42-45，57.
BAI B G，ZHU H L，FAN Q X. Application research of BP neural network in dairy product quality and safety risk[J]. China Dairy Industry，2020，48(7)：42-45，57.

[12] 李兵，梁晋刚，朱育攀，等. 我国小麦赤霉病成灾原因分析及防控策略探讨[J]. 生物技术进展，2021，11(5)：647-652.
LI B，LIANG J G，ZHU Y P，et al. Epidemiological analysis and management strategies of fusarium head blight of wheat[J]. Current Biotechnology，2021，11(5)：647-652.

[13] 李听听. 玉米和小麦储藏中真菌多样性及真菌毒素的研究[D]. 泰安：山东农业大学，2015.
LI T T. Study on fungal diversity and mycotoxins from corn and wheat during storage[D]. Taian：Shandong Agricultural University，2015.

[14] 康会欣. 小麦中脱氧雪腐镰刀烯醇的污染调查及检测方法研

究[D]. 青岛：青岛农业大学，2017.
KANG H X. Contamination Surveys and Detection Methods of the Deoxynivalenol in Wheat. Qingdao：Qingdao Agricultural University，2017.

[15] 李雅静，秦曙，杨艳梅，等. 中国谷物真菌毒素污染研究现状[J]. 中国粮油学报，2020，35(3)：186-194.
LI Y J，QIN S，YANG Y M，et al. Research status of mycotoxin contamination in grains in China[J]. Journal of the Chinese Cereals and Oils Association，2020，35(3)：186-194.

[16] 程天笑，韩小敏，王硕，等. 2018年中国4省脱粒小麦中9种真菌毒素污染情况调查[J]. 食品安全质量检测学报，2020，11(12)：3992-3999.
CHENG T X，HAN X M，WANG S，et al. Investigation on contamination situation of 9 mycotoxins in wheat kernel from 4 provinces of China in 2018[J]. Journal of Food Safety & Quality，2020，11(12)：3992-3999.

[17] 程天笑. 中国四省脱粒小麦中多种真菌毒素协同污染情况的调查[D]. 北京：北京工业大学，2020.
CHENG T X. Investigation on synergistic contamination of various mycotoxins in threshed wheat in four provinces of China[D]. Beijing：Beijing University of Technology，2020.

[18] 李星，郭红卫，许洁，等. 上海地区小麦中霉菌毒素的污染调查[J]. 上海预防医学杂志，1997，9(9)：413-415.
LI X，GUO H W，XU J，et al. Studies on mycotoxins contamination or wheat in Shanghai[J]. Shanghai Journal of Preventive Medicine，1997，9(9)：413-415.