文章编号：2095-2163(2021)05-0188-05 中图分类号：TP391.9 文献标志码：A

基于信息熵的食品安全事件聚类方法研究

辜萍萍

（厦门大学嘉庚学院 信息科学与技术学院，福建 漳州 363105）

摘 要：食品安全是社会各界日益关注的民生问题，政府部门正在逐步完善监管体制、加大监管力度，构建社会共治的格局。本文针对已经曝光的食品安全事件，经过清洗筛选建立统一规范的数据存储，利用改进的基于信息熵模糊聚类分析算法对其进行数据挖掘，以便发现这些事件中具有象征性的现象以及典型性的安全事件，从而为政府制定管理决策和为民众提高防范意识提供参考性依据。实验中将改进的算法运行在 UCI 数据集上验证算法的有效性，结果表明该算法进一步提高了聚类的正确率、类精度及召回率。

关键词：食品安全；数据挖掘；聚类分析；信息熵

Clustering analysis model simulation for food safety events based on information entropy

GU Pingping

（Computer Science Department, Tan Kah Kee College, Zhangzhou Fujian 363105, China）

【Abstract】 Food safety is an increasingly concerned issue of people´s livelihood. At present, the national government is gradually improving the regulatory system and increasing regulatory efforts, and building the pattern of joint governance. In view of the food safety events that have been exposed, a unified and standardized data storage is established after screening and cleaning, and the improved clustering algorithm based on information entropy is used to analyze and study them, so as to find the rules in these events and the typical safety problems which can provide a reference for the government to make management decisions and improve the public awareness of prevention. In the experiment, the improved fuzzy K-Modes clustering algorithm is run on the UCI data set to test the effectiveness of the algorithm. The results show that the improved algorithm further improves the clustering quality.

【Key words】 food safety; data mining; cluster analysis; information entropy

0 引 言

近几年，中国的食品安全危机成为发展小康社会的绊脚石，主要集中表现在化学添加剂与农药等滥用、生产流通环节卫生状况差、安全标准不够完善、监管水平有待提高几个方面。2017 年 2 月，国务院印发《“十三五”国家食品安全规划》文件，明确指出“保障食品安全是建设健康中国、增进人民福祉的重要内容，是以人民为中心发展思想的具体体现”[1]。规划中制定了针对食品安全管理的基本原则：预防为主、风险管理、全程控制与社会共治。现阶段，国家不断加大整治力度，从完善政府监管体制到增强全民自主防范意识多渠道多维度进行共治，全面打造老百姓可信赖的食品安全环境 。

在智能信息化时代背景下，提高食品安全智慧的监管能力、实施“互联网+”食品安全监管项目、推进食品安全监管大数据资源共享和应用是切实的目标与手段[2]。目前，数据挖掘方法正越来越多地被应用于食品安全相关数据进行知识发现。例如，江苏省食品安全研究基地的李清光等人运用 Hierarchical Cluster 方法对 2005～2014 年间发生的有明确时空定位的 2 617 起食品安全事件进行聚类分析，总结了事件集的区域分布特点以及随时间变化的空间转移趋势[3]。本研究提出一种基于信息熵的食品安全事件聚类分析模型，对食品安全事件中的相关属性特征进行提取，从聚类模式中挖掘代表性的问题，分析不同种类食品的各种不安全因素是否受到日期与地区的某种影响，发现各地区之间是否在食品安全危机状况上存在相似点，从而为监管部门在措施制订过程中提供决策支持，也为普通消费者在采购食用方面提供信息参考。

1 食品安全数据挖掘的意义

数据挖掘中的聚类分析、关联规则、决策树及偏差检测等技术在食品行业管理领域不断被应用及推广，利用这些挖掘技术找出潜藏在数据中的知识，为

作者简介：辜萍萍（1982-），女，硕士，副教授，主要研究方向：数据挖掘、软件工程。

收稿日期：2021-01-12

监管及预防提供依据。其意义主要体现在以下 2 方面：

(1)数据挖掘技术用于食品安全事件数据的分析,可以发现事件发生的特点、规律以及未来发展趋势,对监管部门分析事件发生的原因、制定监管措施,更有效地预防事件的发生都存在积极的作用,从而利于社会的安定与发展。

(2)通过公众媒体渠道发布数据挖掘的相关结果,让民众了解食品安全事件中更深层次的信息,提醒民众结合平时的采购及饮食习惯,警惕可能存在问题的食物,远离风险源头,保障自身的饮食健康。

2　数据采集与建模

2.1　食品安全数据的采集

无论是食品生产环境受到污染还是制造过程的非法添加,近些年食品安全事件层出不穷。由复旦大学研究生吴恒创办的“掷出窗外”网站从各大主流新闻媒体搜集了最近八年以来全国发生的食品安全事件将近 3 600 条数据记录,构建成一个相对集中的问题食品资料库[4]。本次研究利用爬虫技术,建立数据爬取模型,将网站地址导入模型,随即获取网页信息,读取每个页面上的 50 条事件记录,直到所有页面访问完成。每条记录包含新闻标题、新闻日期、事件发生地区、事件主题、相关食品、品牌、不安全因素等多个标签,由于网站中的原始事件列表存在若干无关项或者标签类别不统一,则需要后续的数据集预处理工作。数据筛选的准则是保留所有有助于挖掘算法实施的记录,所以需要完成特征选择与特征工程的相关工作,针对数据记录中确实存在不统一的属性标签或属性值有所缺失的情况,将数据集的属性标签做统一化预处理,将多余属性排除,将残缺数据补全。

2.2　食品安全数据的存储与描述

每一个食品安全事件都是一个数据样本,这些数据样本的所有属性值都是文本型数据,各属性的取值范围见表 1。

数据存储采用电子表格,以 CSV 纯文本形式存取二维数据表。表格中的每一行为食品安全事件记录,每一列为一种属性。预处理后得到的食品安全事件数据集格式见表 2(因篇幅所限,此处仅提供数据表片段)。

3　算法改进与实现

在现实中的分类往往伴随着模糊性,所以用模糊理论来进行聚类分析也更趋于最优。同时,食品安全数据的最大特点是样本的每一个属性经过标准化均取离散型数值,因此该数据模型最适合采用模糊 K-Modes 算法进行挖掘[5-6]。为了进一步提高算法的有效性,本模型利用加权平均密度的方法自动选取初始聚类中心,并依据信息熵理论重新计算数据集的属性重要性,以此改进传统的算法。

表 1　数据对象各属性取值范围

Tab. 1　Value range of each attribute of data object

属性	属性取值范围及说明
日期	该属性对范围不做限制,以实际读取日期为准
地区	属性范围涉及全国 34 个省级行政区域,如果涉及多个地区,则此项标记为“多地区”,如果记录中未提及,则标记为“未知”
食品品牌	如果某些食品事件中未提及品牌信息,则此项标记为“未知品牌”
食品种类	水果、水产、肉类、蔬菜、粮油、酒类、饮品、零食、调味品、速冻食品、药品、婴儿食品
不安全因素	添加剂、致癌物、转基因、农药、造假、过期、细菌超标,如果记录中未给出的,则此项标记为“不合格”

表 2　食品安全数据集

Tab. 2　Food safety data set

日期年份	地区	食品品牌	食品种类	不安全因素
2015	广东	未知品牌	粮油	致癌物
2017	北京	三只松鼠	零食	添加剂
2017	陕西	未知品牌	饮品	大肠菌
……	……	……	……	……

3.1　模糊 K-Modes 算法思想

传统的模糊 K-Modes 算法总体思想是对文本型的数据集进行划分归类,假设对于一个由 n 个对象构成的非空集合 $U=\{x_1,x_2,\cdots,x_n\}$,首先随机选取 k 个对象作为 k 个初始聚类中心,通过 0 或 1 匹配的方法计算相异度矩阵,再根据相异度矩阵计算隶属度矩阵,而后通过隶属度矩阵将 n 个对象划分到最近的初始聚类中心中,形成 k 个聚类簇,完成一次聚类,计算收敛函数,再通过更新聚类中心的方法在每个聚类簇中重新定义一个新的中心,重复之前的内容计算相异度矩阵、隶属度矩阵、分配对象,形成 $k+1$ 个聚类簇,计算收敛函数,比较 2 次的收敛函数。多次迭代这样的过程,交替更新聚类中心和隶属度矩阵,直到式(1)所示收敛函数的值趋于稳定,即聚类中心不再发生偏移,算法结束。

$$E_c(W,Z)=\sum_{l=1}^{k}\sum_{i=1}^{n}w_{i,l}^{\alpha}d(X_i,Z_l). \qquad (1)$$

Fuzzy K-Modes 算法利用模糊的概念对数据进

行软聚类,大大提高了聚类过程的鲁棒性。

3.2 改进的 Fuzzy K-Modes 算法

信息熵是同一信源发出的各种信息量的平均,可以表征信息环境的无序程度,在一定程度上解决了系统有序度的度量问题[7-8]。根据信息熵理论来计算每个聚类对象的属性对于分类的贡献度,可以排除无用属性。基于传统的算法思想以及信息熵理论,改进的 K-Modes 算法具体流程如下:

(1)输入目标的初始聚类中心数 k,$k \geqslant 1$;

(2)根据公式(2)计算属性总集合 A 的信息熵 $E(A)$:

$$E(A) = -\sum_{i=1}^{p} \frac{|A_i|}{|U|} \log_2 \frac{|A_i|}{|U|}. \quad (2)$$

$E(A)$ 表示整体的信息熵,即所有的属性将数据集 U 划分的情况。其中,A 将数据集 U 划分成了一个新的集合 C,$C = \{A_1, A_2, A_3, \cdots\cdots, A_p\}$,对于 C 中的任意一个元素 A_i 表示数据集 U 中与 B_i 的属性值完全相等的数据集子集,所以 $A_i \subseteq U$,且 $|A_1| + |A_2| + |A_3| + \cdots\cdots + |A_p| = |U|$,$|A_i| / |U|$ 即表示属性值与 A_i 完全相等的元素在数据集 U 中出现的概率;

(3)计算属性总集合中缺少每个属性后的信息熵 $E(A - \{a\})$,其中 $E(A - \{a\})$ 表示去掉 a 属性后,剩余的属性对 U 的划分情况,计算公式与 $E(A)$ 相同;

(4)根据步骤(2)和步骤(3)获取的结果,计算每个属性的权值 Sig(a),式(3):

$$Sig(a) = \frac{E(A) - E(A - \{a\})}{E(A)}. \quad (3)$$

若属性 a 对数据集 U 毫无影响则 $E(A) = E(A - \{a\})$,说明 a 对数据集 U 的划分没有起到作用,即 $Sig(a) = 0$,说明 a 的属性重要性为 0;反之若属性 a 对数据集 U 影响越大,则排除 a 属性的 $E(A - \{a\})$ 与 $E(A)$ 就相差越大;

(5)遍历数据集 U,计算每个属性的平均密度,平均密度的计算公式(4)为:

$$Dens(x) = \frac{\sum_{a \in A} Dens_a(x)}{|A|}. \quad (4)$$

其中,$Dens_a(x)$ 表示对于 A 中的任意元素 a,对象 x 在属性 a 上的平均密度计算方法为式(5):

$$Dens_a(x) = \frac{|\{y \in U | f(x,a) = f(y,a)\}|}{|U|}. \quad (5)$$

该公式为了计算得出数据集 U 中与对象 x 的属性 a 的属性值相同的对象的总数;

(6)对于数据集 U 中的每一个对象 x,计算其加权密度 WDens(x),式(6):

$$WDens(x) = \frac{\sum_{a \in A} Sig(a) \times Dens_a(x)}{|A|}. \quad (6)$$

(7)选取所有对象中加权密度 WDens(x) 最大的一个,将其设为第一个初始聚类中心,加入聚类中心集合 Z;

(8)遍历数据集 U 中已经选取为聚类中心以外的每个对象 x,保存对象的加权密度 WDens(x),计算公式与步骤(6)所述相同;

(9)采用 0-1 相异度度量方法计算对象 x 与每个已分配好的初始聚类中心的距离之和 $d(x)$,式(7)和式(8):

$$d(x_i, x_j) = \sum_{l=1}^{m} \sum_{a \in A} \delta_l(x_{i,al}, x_{j,al}), \quad (7)$$

$$\delta_l(x_{i,al}, x_{j,al}) = \begin{cases} 0 & x_{i,al} = x_{j,al} \\ 1 & x_{i,al} \neq x_{j,al} \end{cases}. \quad (8)$$

其中,$x_{i,al}, x_{j,al}$ 分别表示数据集中 x_i 和 x_j2 个对象在对应属性上的属性值,如果相等则当前属性间的距离赋值为 0,如果不相等则赋值为 1,累加所有属性的属性间距离,最后得出 2 个对象之间的距离,即差异度;

(10)对每一个对象 x,计算 $m(x)$,式(9):

$$m(x) = WDens(x) + d(x). \quad (9)$$

(11)比较所有的 $m(x)$,选取 $m(x)$ 最大的那个对象作为新的初始聚类中心,加入聚类中心集合 Z。

(12)判断聚类中心数是否达到 k 个,即 $|Z| > k$ 是否成立,如果成立跳转到步骤(13),如果不成立则跳转到步骤(8),继续选择其它的初始聚类中心;

(13)根据步骤(4)所述的 $Sig(a)$,计算每个属性的权值 $weight(a)$,式(10):

$$weight(a) = \begin{cases} \frac{2}{3} & \text{if } Sig(a) = 0 \\ 1 + Sig(a) & \text{if } Sig(a) > 0 \end{cases}. \quad (10)$$

(14)用改进的相异度度量方法计算相异度矩阵,式(11):

$$wd(x_i, x_j) = \sum_{a \in A} weight(a) \times \delta_a(x_i, x_j). \quad (11)$$

其中,$\delta_a(x_i, x_j)$ 参见步骤(9)所述的定义。该公式的含义为任意 2 个对象之间的距离,当属性值相同时依然为 0,当属性值不同时取属性的权值作为距离计算,这样价值高的属性影响力就更大。

(15)计算隶属度矩阵 $W_{l\times n}$，式(12)：

$$w_{i,l}=\begin{cases}1, & \text{若 } X_i=Z_l;\\ 0, & \text{若 } X_i=Z_h, h\neq l;\\ 1/\sum_{h=1}^{k}\left[\frac{d(Z_l,X_i)}{d(Z_h,X_i)}\right]^{2/(\alpha-1)}, & \text{若 } X_i\neq Z_l, X_i\neq Z_h, 1\leqslant h\leqslant k.\end{cases} \tag{12}$$

其中，k 表示当前数据集划分为 k 个簇，即存在 k 个聚类中心；Z_l 表示当前第 i 个类的聚类中心；Z_h 表示其它类的聚类中心；

(16)根据隶属度更新聚类中心集合 Z，采用属性众数作为聚类中心的新的属性值。即遍历每一个类簇，计算类簇里每一个属性的每一个属性值的总数，用总数最高的属性值替换当前该类簇的聚类中心；

(17)回到步骤(15)重新计算隶属度，根据每个样本的最大隶属度重新归类；如果隶属度不再变化，那么 k 类的聚类已经完成，跳转至步骤(18)；

(18)根据当前隶属度矩阵与相异度矩阵计算聚类准则函数，式(13)：

$$E_c(W,Z)=\sum_{l=1}^{k}\sum_{i=1}^{n}w_{i,l}^{\alpha}wd(X_i,Z_l). \tag{13}$$

其中，n 是数据集的规模，即聚类对象的数量；$Z_l=[z_{l1},z_{l2},\ldots,z_{lm}]$ 是能够代表聚类 l 的向量，即聚类中心；$w_{i,l}\in[0,1]$ 是隶属度矩阵 $W_{l\times n}$ 的一个元素，表示对象 X_i 划分到聚类 l 中的隶属度；$\sum_{l=1}^{k}\varpi_{li}=1$；$wd$ 是改进后的相异度(距离)；$\alpha>1$ 是加权指数；

(19)聚类数量 k 递增 1，并回到步骤(1)，直到 $k=\sqrt{n}$ 为止，聚类准则函数最小的那一轮聚类为最后的聚类结果。

4　实验结果分析

为了验证算法的有效性，利用著名的算法有效性指标正确率 AC(accuracy)、类精度 PC(precision)、召回率 RE(recall) 进行实验比对，式(14) ~ 式(16)：

$$AC=\frac{\sum_{i=1}^{k}a_i}{|U|}, \tag{14}$$

$$PC=\frac{\sum_{i=1}^{k}\left(\frac{a_i}{a_i+b_i}\right)}{k}, \tag{15}$$

$$RE=\frac{\sum_{i=1}^{k}\left(\frac{a_i}{a_i+c_i}\right)}{k}. \tag{16}$$

其中，k 表示数据集当前的聚类数量，$|U|$ 表示整个数据集的对象数量，令 a_i 代表被正确分配到第 i 类的对象数量，令 b_i 代表被错误分配到第 i 类的对象数量，令 c_i 代表被错误排除出第 i 类的对象数量。实验中，除了准备好有效性指标之外，还需要有明确分类结果的测试数据集，因此模型中选择加州大学欧文分校提出的用于机器学习的数据库——UCI 数据库作为实验数据，其包含 335 个数据集，是一个常用的标准测试数据集。该数据集中分别含有数值型数据集与文本型数据集，其中的文本型数据集适用于本模型提出的算法有效性实验。每个数据集提供了一份完整的数据记录、分类属性和分类结果集，实验中将数据集导入聚类分析模型执行改进的算法，进而计算 PC，AC，RE3 项指标值，从而检验算法的有效性。

本文选择 UCI 数据库中的 3 个文本型数据集 Soybean、Zoo 和 Vote 来检验算法。并通过与其它 K-Modes 算法进行 AC，PC，RE 指标值的对比，发现本文所改进的算法具有更高的聚类有效性。实验中选择随机选取初始聚类中心的 K-Modes 算法(Huang's k-modes with random)和基于平均密度选取初始聚类中心的 K-Modes 算法(Huang's k-modes with Xing's method)参与比较[10]。实验结果见表 3 ~ 表 5。

表 3　Soybean 数据集聚类有效性指标表

Tab. 3　Index table for clustering effectiveness of Soybean data set

Validation Measure	Huang's K-Modes with Random	Huang's K-Modes with Xing's Method	Proposed method
AC	0.854 3	1.000 0	1.000 0
PC	0.889 0	1.000 0	1.000 0
RE	0.827 7	1.000 0	1.000 0

表 4　Zoo 数据集聚类有效性指标表

Tab. 4　Index table for clustering effectiveness of Zoo data set

Validation Measure	Huang's K-Modes with Random	Huang's K-Modes with Xing's Method	Proposed method
AC	0.840 3	0.920 8	0.940 6
PC	0.824 1	0.881 9	0.910 4
RE	0.633 2	0.785 7	0.842 9

表 5 Vote 数据集聚类有效性指标表

Tab. 5 Index table for clustering effectiveness of Vote data set

Validation Measure	Huang's K-Modes with Random	Huang's K-Modes with Xing's Method	Proposed method
AC	0.836 2	0.839 7	0.857 6
PC	0.822 6	0.862 4	0.961 3
RE	0.731 7	0.794 1	0.802 5

为了更直观展示 3 种算法在各实验数据集上的有效性指标对比，将 *AC*、*PC* 和 *RE* 在不同数据集上分别求出平均值生成折线图如图 1 所示。

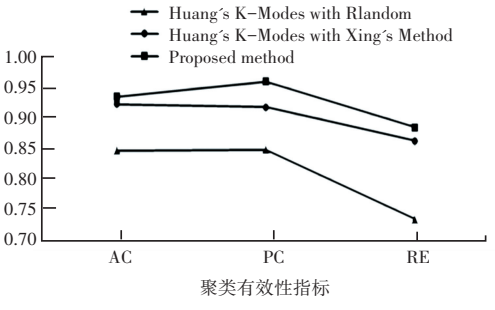

图 1 聚类有效性指标平均值对比图

Fig. 1 Comparison chart of average value of cluster validity index

本次研究中，将改进的聚类算法应用于通过爬虫系统爬取获得的食品安全数据集进行规律挖掘，经过对初始数据集的清洗整理，最后筛选出2 751条有效记录。利用上述算法中对最大聚类数的计算，实验中需要进行 52 轮聚类，并分别计算聚类准则函数值，取值最小的那一轮聚类为最后的聚类结果。聚类完成后每个类簇的聚类中心，即每个类中最具有代表性的食品安全事件已经被挖掘出来见表 6。需要特别说明的是，由于该数据集网站“掷出窗外”近年来暂无更新，因此安全事件发生日期均是若干年前；同时，被曝光的大多数安全事件中缺乏关于食品品牌的具体数据，因此在预处理数据时均已标记为“未知品牌”。

5 结束语

经过 UCI 文本数据集的实验测试，本文改进的基于信息熵的模糊 K-Modes 算法与传统的算法相比在聚类正确率、类精度和召回率 3 项有效性指标上均表现出更加优越的特性。

虽然国家越来越重视百姓的餐饮安全，但问题屡禁不止。利用数据挖掘算法对已经采集到的食品安全事件数据集进行聚类分析，找到每个聚类中心，这些聚类模式代表着数据集中最值得关注的焦点事件，从而引起监管部门的重视以及消费者的警惕。

表 6 食品安全数据集聚类中心

Tab. 6 Clustering center of food safety data set

日期	地区	食品品牌	食品种类	不安全因素
2006	北京	未知品牌	零食	添加剂
2011	北京	未知品牌	饮品	不合格
2011	广东	未知品牌	肉类	致癌物
2012	广东	未知品牌	粮油	不合格
2011	山东	未知品牌	水产	添加剂
2012	山东	未知品牌	水果	农药
2009	浙江	未知品牌	饮品	添加剂
2012	江苏	未知品牌	零食	添加剂
2011	上海	未知品牌	零食	添加剂
2012	上海	未知品牌	饮品	添加剂
2012	福建	未知品牌	肉类	不合格
2012	福建	未知品牌	饮品	不合格
2011	四川	未知品牌	粮油	不合格
2013	湖南	未知品牌	粮油	不合格
2009	湖北	未知品牌	粮油	添加剂

在下一步研究应用中，继续关注食品安全领域的相关事件，尽可能获取到最新的数据进行数据挖掘，以便对当下的食品质量监管提供参考性建议。

参考文献

[1] 国务院：《“十三五”国家食品安全规划》[EB/OL]. (2018-11-29). http://www.miit.gov.cn/n1146290/n1146392/c5497473/content.html.

[2] 余学军. “互联网+”时代食品安全智慧监管策略研究[J]. 食品工业, 2018, 67(8): 244-246.

[3] 李清光，李勇强，牛亮云，等. 中国食品安全事件空间分布特点与变化趋势[J]. 经济地理，2016, 36(3): 9-16.

[4] 葛成恩. 掷出窗外的食品安全[J]. 品牌，2012, (12): 71.

[5] HUANG Zhexue. Extensions to the k-means algorithm for clustering large data sets with categorical values[J]. Data Mining and Knowledge Discovery，1998, 2: 283-304.

[6] HUANG Zhexue, Michael K.Ng. A Fuzzy K-Modes Algorithm for Clustering Categorical Data[J]. IEEE Transactions on Fuzzy Systems, 1999, 4(7): 108-110.

[7] 程慧平，程玉清. 基于 AHP 与信息熵的个人云存储安全风险评估[J]. 情报科学，2018，36(7): 145-151.

[8] 李根强，刘莎，张亚楠，等. 信息熵理论视角下网络集群行为主体的观点演化研究[J]. 情报科学，2020，38(1): 42-47.