第 31 卷 第 5 期
2016 年 10 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol. 31 No. 5
Oct. 2016

DOI:10.13364/j.issn.1672-6510.20150209

基于 HMM 的食品安全风险预测方法研究

马永军，刘宇姗，侯阳阳，刘 洋
（天津科技大学计算机科学与信息工程学院，天津 300222）

摘 要：以乳品为例，提出了一种基于隐马尔可夫模型（HMM）的食品安全风险预测方法. 依据 HACCP 对乳品供应链进行分析，找出供应链各个环节的关键控制点，并将其危害影响因素作为 HMM 的量化指标. 分析时将供应链分为 4 个环节，每个环节最终的安全等级概率作为下一环节的初始状态概率分布，利用 HMM 预测供应链风险等级和计算风险值，从而为供应链风险评价提供依据.

关键词：食品安全；HMM 模型；乳品；供应链；风险预测
中图分类号：TP311 **文献标志码**：A **文章编号**：1672-6510（2016）05-0069-05

Food Safety Risk Prediction Based on HMM

MA Yongjun，LIU Yushan，HOU Yangyang，LIU Yang
（College of Computer Science and Information Engineering，Tianjin University of Science & Technology，Tianjin 300222，China）

Abstract：Taking dairy products as the research subjects，a food safety risk assessment method based on the Hidden Markov Model（HMM） was developed. The critical control point for each sector of the supply chain was found out with the method of hazard analysis critical control point. Those critical points can be used as quantitative indicators of HMM. According to the characteristics HMM model，the supply chain was divided into four sectors. The ultimate probability of each sector is the initial state probability distribution of the next part. Finally the level of risk in the supply chain and risk level can be calculated with HMM，which can provide a reference for supply chain risk assessment.

Key words：food safety；HMM；dairy products；supply chain；risk prediction

食品安全是关系国计民生的重大问题，近年来食品安全问题的预警和治理已受到高度关注. 其中，风险预测作为风险预警和治理的基础受到格外重视. 风险预测是利用适当的风险预测工具和方法来确定安全风险等级和优先控制顺序的过程，从而为实施有效的风险管理措施提供决策支持.

目前常用的食品安全风险预测方法主要分为定性和定量两大类，其中定性的方法主要有灰色系统法、德尔菲法、层次分析法等. 灰色系统法[1]是对不确定系统的研究方法；德尔菲法[2]能够充分发挥各位专家的不同意见，评定过程公正；层次分析法[3]是食品安全风险分析与预测中应用最为广泛的方法. 但是它们的共性缺点就是在运算过程中受主观因素影响较大，使得结果的精确度不够，并且不能根据系统状态的变化实时改变风险预测结果[4]. 定量的分析方法主要有贝叶斯网络法、人工神经网络和支持向量机（SVM）等. 贝叶斯网络法[5]是一种基于概率推理的图形化概率网络，能在有限的不确定信息条件下进行学习和推理，但是贝叶斯网络构造繁琐，实际应用时还需反复交叉不断完善，易用性不好. 人工神经网络[6]具有高度的自适应能力，但是学习速度较慢，而且算法容易陷入局部极值. SVM[7]本质上是将问题转换为凸优化问题，可以保证找到全局极值，但在将低维空间向高维空间映射时，又存在计算耗时问题，因此不适合对于大样本数据的分析处理.

食品安全分析和食品供应链密不可分，食品供应

收稿日期：2015–11–16；**修回日期**：2016–04–18
基金项目：教育部规划项目（12YJAZH091）
作者简介：马永军（1970—），男，吉林长春人，教授，yjma@tust.edu.cn.

链主要由原料生产、生产加工、储运和消费 4 个环节组成，食品安全风险存在于食品供应链的各个环节中，而且存在食品安全风险迁移问题，即一个环节的风险会向下一个环节传递. 在目前的方法中，对风险的判断一般是对 4 个环节的风险进行简单平均或加权平均，未充分考虑食品安全风险的形成机理，割裂了 4 个环节之间的风险传递关系.

HMM 模型是序列数据处理和统计学习的一种重要概率模型，具有建模简单、数据计算量小、运行速度快等特点[8]，其在风险预测分析领域的应用较广泛[9]. HMM 模型被用来对网络系统安全进行实时风险预测，表现出了很好的适用性和扩展性[10]. 此外，HMM 模型中的初始状态分布矢量参数，对应初始状态时各风险等级出现的概率，该参数可以有效表示食品供应链各环节对下一环节的影响程度. 因此本文以乳品为例，利用 HMM 模型进行食品安全风险分析与预测.

1　HMM模型

HMM 模型是在马尔可夫链的基础上发展起来的，在 HMM 中观察到的事件是状态的随机函数，该模型是一个双重随机过程，即一个观察状态，一个隐藏状态.

HMM 的数学表达式为

$$\lambda=(N,M,\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{B}) \tag{1}$$

也可以简单表示为

$$\lambda=(\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{B}) \tag{2}$$

其中：N 为 HMM 中状态的个数；M 为 HMM 中对应的观测值个数；$\boldsymbol{\pi}$ 是初始状态的分布矢量；$\boldsymbol{A}$ 为状态转移概率矩阵；$\boldsymbol{B}$ 是观察向量的概率矩阵.

初始状态的分布矢量可表示为

$$\boldsymbol{\pi}=(\pi_1,\pi_2,\cdots,\pi_N) \tag{3}$$

$$0\leqslant\pi_i\leqslant 1,\sum_{i=1}^{N}\pi_i=1$$

状态转移概率矩阵可表示为

$$\boldsymbol{A}=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{bmatrix} \tag{4}$$

元素 a_{ij} 表示从状态 i 转移到状态 j 的概率，即

$$a_{ij}=P\left(a_i\rightarrow a_j\right)$$

$$1\leqslant i\leqslant N,1\leqslant j\leqslant N$$

观察向量的概率矩阵可表示为

$$\boldsymbol{B}=\begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1M} \\ b_{21} & b_{22} & \cdots & b_{2M} \\ \vdots & \vdots & \vdots & \vdots \\ b_{N1} & b_{N2} & \cdots & b_{NM} \end{bmatrix} \tag{5}$$

元素 b_{jk} 表示在 j 状态的情况下，观察状态 k 出现的概率，即

$$b_{jk}=P\left(b_k\mid b_j\right)$$

$$1\leqslant j\leqslant N,1\leqslant k\leqslant M$$

HMM 需要解决 3 个问题：

(1) 评估问题：给定五元组模型 $\lambda=(N,M,\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{B})$ 和观察序列 $\boldsymbol{O}=(O_1,O_2,\cdots,O_T)$，计算这个观察序列出现的概率.

(2) 解码问题：给定五元组模型 $\lambda=(N,M,\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{B})$ 和观察序列 $\boldsymbol{O}=(O_1,O_2,\cdots,O_T)$，求可能性最大的隐藏状态序列.

(3) 学习问题：也叫训练问题. 给定观察值序列 $\boldsymbol{O}=\left(O_1,O_2,\cdots,O_T\right)$，据此确定 HMM 模型 $\lambda=(N,M,\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{B})$. 即如何调整 $\lambda=(N,M,\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{B})$，使得 $P(O\mid\lambda)$ 最大.

上述的 3 个问题分别对应 3 个经典的解决算法[11-12]，分别是前向–后向算法、Viterbi 算法、Baum-Welch 算法.

2　食品安全风险预测方法

食品供应链主要包括原料生产、生产加工、储藏运输和消费 4 个环节，每个环节都会受很多因素的影响. 假设影响因素为

$$\boldsymbol{X}_N=\{X_{N1},X_{N2},\cdots,X_{Ni}\}$$

其中 $\boldsymbol{X}_N\left(1\leqslant N\leqslant 4\right)$ 为风险预测的观测值，X_{Ni} 为第 N 个环节的第 i 个影响因素，通过一定的方法反映出风险预测值和其影响因素之间的关系，数学模型表达式为

$$Y_N=f\left(X_{Ni}\right) \tag{6}$$

$$1\leqslant N\leqslant 4$$

HMM 模型有 2 个随机过程：可见的状态序列，本文指供应链各个环节检测到的数据；系统的真实状态，本文指供应链各环节的真实风险值. 供应链每个环节的 HMM 三元组参数模型包括状态转移概率矩阵 $\boldsymbol{A}$，观测概率矩阵 $\boldsymbol{B}$ 和最初状态分布 $\boldsymbol{\pi}$，字母表示为 $\lambda=(\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{B})$. 除了第一个环节外，其他 3 个环节的参数模型都受上一环节的影响，本环节最终的风险概率直接影响下一环节的初始状态概率分布. 例如，原

料生产环节通过 HMM 模型计算，最终得出各个安全状态出现的概率分布为 $\boldsymbol{\delta}=\{\delta_1,\delta_2,\cdots,\delta_i\}$（$i$ 为真实状态的个数），那么下一环节，即生产加工环节的初始状态分布概率 $\boldsymbol{\pi}=\boldsymbol{\delta}$.

设有 A_1 — A_5 共 5 个状态，表示供应链风险等级，其中 A_1 表示正常安全状态，供应链受威胁的概率为 $P=0$；A_2、A_3、A_4 表示危险等级逐级加深，对应供应链受威胁的概率分别为 $0<P\leqslant 0.2$（低风险）、$0.2<P\leqslant 0.5$（中级风险）、$0.5<P\leqslant 0.8$（中高级风险）；A_5 表示重大危险状况，供应链受威胁的概率为 $0.8<P\leqslant 1$.

这里的 HMM 为离散 HMM，在建立离散的 HMM 模型系统时，可以假定：系统状态包含系统的所有信息，并且在当前状态下的观察是独立的. 则 HMM 模型的状态转移模型可以表示为图 1. 从一个节点移动到另一个节点，表示系统在源节点显示的状态，会转变成目的节点的状态，该模型图是一个完全连接图，表明任意安全状态都有转变为其他任意安全状态的可能.

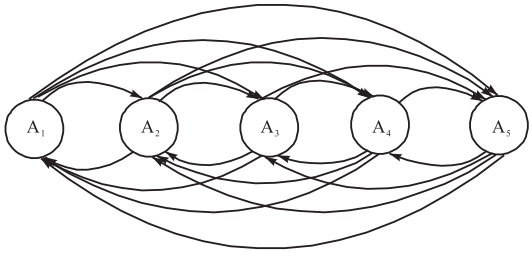

图 1　HMM 状态转移模型

Fig. 1　Transfer model of HMM state

据以上分析，提出图 2 所示的风险预测流程.

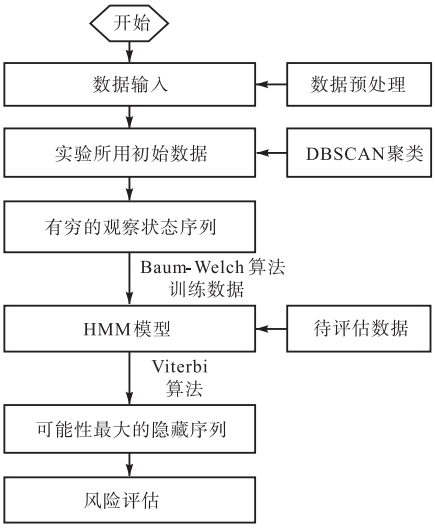

图 2　风险预测流程

Fig. 2　Risk assessment process

在用 HMM 模型进行风险预测时，首先对数据进行预处理，采用的归一化映射为

$$f: x \to y = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{7}$$

其中 $x, y \in \mathbf{R}^n$. 归一化后的原始数据在[0，1]范围内，即 $y_i \in [0,1]$，$i=1,2,\cdots,n$. 然后，将归一化的数据通过 DBSCAN 聚类去除噪音点，得到初始数据. DBSCAN 算法不需要事先知道要形成的簇类的数量，可以发现任意形状的簇类，并且能够识别出噪声点，适用于类别数未知的聚类问题[13]. 把聚类之后的数据通过 Baum-Welch 算法训练得到对应的 HMM 三元组模型后，即可通过 Viterbi 算法对供应链进行相应的风险预测.

3　实验分析

在食品安全方面，乳品一直受高度关注. 所以本文以乳品为例进行分析. 乳品在从农场到餐桌的过程经历了 4 个环节：原奶生产、乳品加工、储藏运输和消费. HACCP[14]（hazard analysis critical control point）表示危害分析和关键控制点，它是对食品加工过程中可能发生的食源性危害进行识别和评估，进而采取一定控制的食品安全控制体系. 利用 HACCP 确定关键控制点的判断流程如图 3 所示.

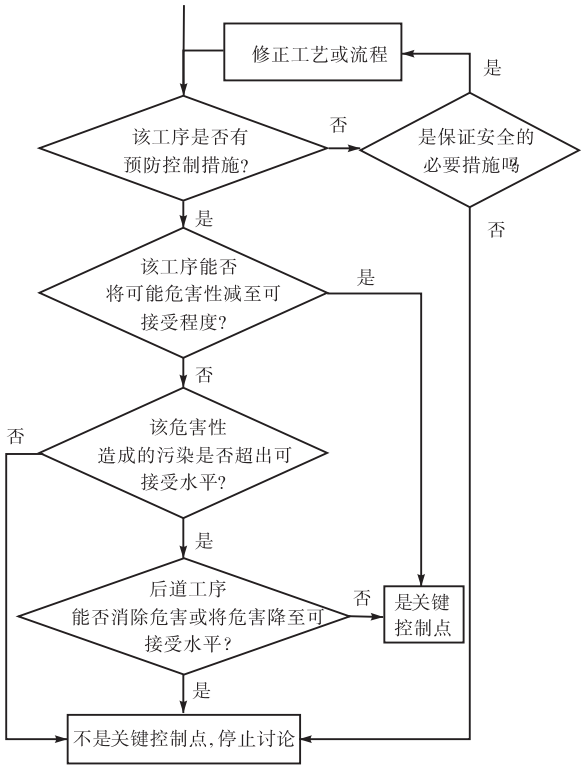

图 3　确定关键控制点的判断树

Fig. 3　Critical control point decision tree

根据 HACCP 体系并结合乳品供应链的 4 个环节，对各个环节中的生物性、物理性和化学性的潜在危害进行分析，确定了关键控制点. 以第一环节为例，原奶生产环节的关键点选择见表 1.

由于乳品一般都难以长久保持并容易受到微生物污染，供应链上任何环节的潜在危险因素都会影响食品质量安全. 乳品加工环节是整个供应链最重要的部分，负责将收集到的原奶加工完成，储运和消费环节的风险因素也基本来源于供应链的前几个阶段. 因此，乳品生产供应链的各个环节紧密联系，对每一个环节的检查监管都要格外重视. 最后一个环节的风险值即为整条供应链的风险大小.

表 1　乳品原奶生产环节关键控制点分析

Tab. 1　Critical control point analysis of raw material production

关键点	潜在危害	危害是否显著	关键控制点
饲料	生物性:细菌	否	是
	化学性:化学物质含量	是	
	物理性:紫外线、温度	否	
运输到牧场	生物性:细菌	是	否
	物理性:温度	否	
养殖	生物性:细菌及病毒	是	是
	化学性:药物感染	是	
	物理性:紫外线照射	否	
运输到食品加工厂	生物性:细菌	是	是
	物理性:温度	是	

通过对天津某乳品企业的调研，采集大量数据，并建立了乳品追溯系统，利用 HMM 模型对乳品安全风险进行预测.

乳品供应链各环节的隐藏状态有 5 个：A_1—A_5，分别代表 5 个安全等级. 对归一化的数据通过 DBSCAN 聚类进行聚类处理后，共得到 4 类，也就是系统的 4 个观察状态：V_1—V_4；利用 HMM 模型依次对 4 个环节进行参数评估和风险概率计算，每个环节最终的风险概率直接影响下一环节的初始状态概率分布.

在训练 HMM 模型时，利用 Baum-Welch 算法，通过让概率 $P(\boldsymbol{O}\mid\lambda)$ 达到局部最大值得到 HMM 的参数模型.

定义供应链所有特征观察值序列前向变量[11]为

$$a(i)=P(O_1,O_2,\cdots,O_t,q_t=\theta_t\mid\lambda) \tag{8}$$

$$1\leqslant t\leqslant T$$

由式(8)可知，供应链所有特征观察值序列概率为

$$P(O\mid\lambda)=\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(i) \tag{9}$$

$$1\leqslant t\leqslant T-1$$

通过式(9)求取 HMM 的三元组参数模型 $\lambda=(\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{B})$，再由重估公式[15]可得

$$\overline{\pi_i}=\xi_1(i) \tag{10}$$

$$\overline{a_{ij}}=\sum_{t-1}^{T-1}\xi_t(i,j)/\sum_{t=1}^{T-1}\xi_t(i) \tag{11}$$

$$\overline{b_{ij}}=\sum_{\substack{t-1\\O_t=V_k}}^{T}\xi_t(j)/\sum_{t=1}^{T}\xi_t(j) \tag{12}$$

输出概率 $P(\boldsymbol{O}\mid\lambda)$ 会随着重估次数的增加越来越大，直至参数 π_i、a_{ij}、b_{ij} 收敛或算法达到停止条件为止.

计算得到第一环节的初始状态分布概率为 $\boldsymbol{\pi}=\{0.4612,0.1132,0.2139,0.0547,0.1570\}$，状态转移矩阵为

$$\boldsymbol{A}=\begin{bmatrix}0.1028 & 0.0367 & 0.1244 & 0.1914 & 0.5447\\ 0.2126 & 0.2620 & 0.2173 & 0.0103 & 0.2978\\ 0.1134 & 0.2874 & 0.2337 & 0.2749 & 0.0906\\ 0.3884 & 0.0649 & 0.2494 & 0.1342 & 0.1632\\ 0.2178 & 0.2031 & 0.3187 & 0.2210 & 0.0395\end{bmatrix}$$

观测向量概率矩阵为

$$\boldsymbol{B}=\begin{bmatrix}0.0174 & 0.7426 & 0.1228 & 0.1173\\ 0.2169 & 0.1672 & 0.3590 & 0.2569\\ 0.5773 & 0.1112 & 0.1560 & 0.1555\\ 0.0330 & 0.2189 & 0.3475 & 0.4006\\ 0.3838 & 0.1947 & 0.1524 & 0.2692\end{bmatrix}$$

有了 HMM 的三元组模型，就可以对相关数据进行解码分析. 在此给定一组观察序列 $\{O_2,O_1,O_3,O_4,O_4,O_3,O_2,O_2,O_1,O_1,O_4,O_3\}$，观察序列的状态分布如图 4 所示.

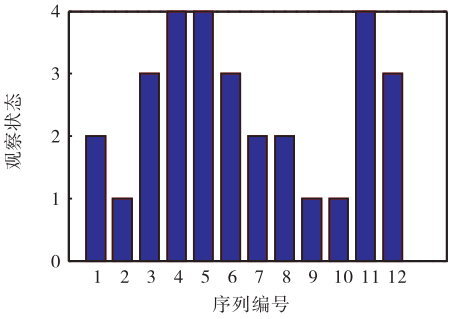

图 4　观察状态序列

Fig. 4　Sequence of observed state

继而利用 Viterbi 算法求解观察序列的最优路径，得到供应链最有可能的真实状态. 首先将 π_i 与所有特征观察值转移概率 b_i 相乘，得到初始化路径

$\delta_t(i)$.

$$\delta_t(i)=\pi_i b_i(O_1),\psi_t(i)=0 \quad (13)$$

$$1\leqslant i\leqslant N$$

将初始化的路径 $\delta_t(i)$ 与状态转移概率矩阵的元素 a_{ij} 相乘，取最大值与所有特征观察值转移概率 b_j 相乘，得到当前路径最大值 $\delta_t(j)$.

$$\delta_t(j)=\max_{1\leqslant i\leqslant N}\left[\delta_{t-1}(i)a_{ij}\right]b_j(O_t) \quad (14)$$

$$2\leqslant t\leqslant T,1\leqslant j\leqslant N$$

$$\psi_t(j)=\arg\max_{1\leqslant i\leqslant N}\left[\delta_{t-1}(i)a_{ij}\right] \quad (15)$$

$$2\leqslant t\leqslant T,1\leqslant j\leqslant N$$

根据式(14)、(15)结果，可求出对应的食品安全各环节真实状态的最大概率 P^* 和最大概率对应状态序列 q_t^* 分别为

$$P^*=\max_{1\leqslant i\leqslant N}\delta_T(i) \quad (16)$$

$$q_t^*=\psi_{t+1}\left(q_{t+1}^*\right) \quad (17)$$

$$1\leqslant t\leqslant T+1$$

因此，可以评估出该环节对应的真实状态 $\{T_1,T_2,T_1,T_2,T_4,T_4,T_1,T_2,T_4,T_2,T_5,T_2\}$.

真实状态序列的分布如图 5 所示.

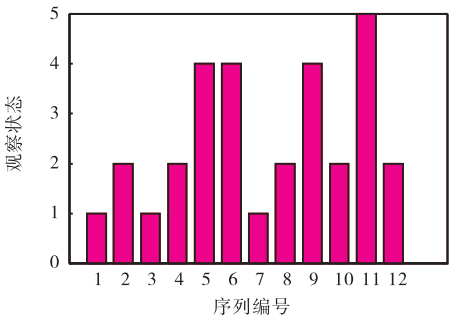

图 5　真实状态序列

Fig. 5　Sequence of true state

由结果可知，系统在 $t=5,6,9,11$ 时的风险级别较大，在 $t=1,3,7$ 时刻，风险级别较小. 通过 Viterbi 算法可以计算出原奶生产环节在最后时刻处在各状态的概率为 $\boldsymbol{\delta}=\{0.1672,0.5342,0.1745,0.0723,0.0518\}$，则在乳品生产环节中，初始状态分布概率就为 $\boldsymbol{\pi}=\{0.1672,0.5342,0.1745,0.0723,0.0518\}$.

参照对第一环节的计算过程，依次利用 Baum-Welch 算法对各环节进行参数计算，然后利用 Viterbi 算法计算系统的真实状态.

最后，计算得到最终环节——消费环节各安全等级的概率为 $\boldsymbol{\pi}=\{0.1547,0.6073,0.0645,0.1043,0.0692\}$.

假设在 A_1 — A_5 不同安全等级的情况下，系统对应的开销，即风险对系统的影响分值为 $C=\{0,5,10,15,20\}$. 参考风险评价模型的典型代表 OCTAVE[10]给出一个计算风险值的公式

$$R_t=\sum_i^N\alpha_t(i)c(i) \quad (19)$$

其中：R_t 表示在 t 时刻系统所处的总体风险值；$\alpha_t(i)$ 为 t 时刻系统处在安全状态 A_i 的概率；N 是安全状态的数目；$c(i)$ 是与状态 A_i 关联的开销.

根据式(19)，可以计算出乳品的最终风险值 $R=6.049\,5$.

因此，可以利用上述方法对乳品各环节的不同时刻进行风险预测与开销分析，有助于生产者和决策者更全面了解乳品供应链各环节的风险态势，及时查出问题发生的缘由，根据已有流程或者规定采取相应措施，并不断更新和完善风险规避的方法，最大程度减少乳品安全对群众的影响.

4　结　语

本文利用 HMM 模型对乳品质量安全风险进行分析与预测. 该方法根据食品安全风险的形成机理，考虑了 4 个环节之间的风险传递关系以及各个环节风险的动态性，结合定性与定量评价，可以更精确地描述乳品风险的大小. 该风险预测模型除了可用于对乳品进行评估外，还可以扩展应用到食品安全的其他领域. 今后还需通过更多实验对各环节风险传递的影响因子进行优化，改进 HMM 模型的参数训练方法，以提高参数模型的训练速度和评估准确度.

参考文献：

[1] 黄秋婷，丁怡，邱佩丽. 基于灰色残差修正模型的食品安全监测预警分析方法研究[J]. 现代农业科技，2013(5)：304–305.

[2] Soon J M，Davies W P，Chadd S A，et al. A Delphi-based approach to developing and validating a farm food safety risk assessment tool by experts[J]. Expert Systems with Applications，2012，39(9)：8325–8336.

[3] 安珺. 基于层次分析法的乳品质量安全预警系统研究[D]. 哈尔滨：东北农业大学，2012.

[4] 孔大维. 基于供应链管理的食品可追溯系统的构建研究[D]. 成都：成都理工大学，2013.

[5] 张丽，滕飞，王鹏. 基于贝叶斯网络的食品供应链风险评价研究[J]. 食品研究与开发，2014(18)：179–182.

（下转第 78 页）