食品溯源时序数据的函数型聚类分析

高 嵘，王 强，罗 东，秦志光

(电子科技大学计算机科学与工程学院 成都 610054)

【摘要】借助函数型数据分析的方法和思想，在一个一般性的框架下讨论对猪肉链时序数据的聚类问题，提出了函数型聚类分析在食品追溯时序数据分析上的应用方法：把时序数据看成一个完整的关于时间的函数对象，而非个体观测值的简单排列，将离散数据转化为函数数据；用基函数展开系数向量的距离代替原函数之间的距离，减少了大量数值积分，简化了运算。通过对复杂的、时序性强的溯源数据进行函数型聚类分析，把复杂离散的数据聚类为连续的分类信息，使得溯源数据的可用性极大增强，可以为决策者和进一步的分析提供科学依据。

关 键 词 聚类分析； 食品溯源； 射频识别； 时序数据

中图分类号 TP391 **文献标识码** A doi:10.3969/j.issn.1001-0548.2012.04.016

Functional Cluster Analysis of Time Series Data in Food Traceability

GAO Rong, WANG Qiang, LUO Dong, and QIN Zhi-guang

(School of Computer Science and Engineering, University of Electronic Science and Technology of China Chengdu 610054)

Abstract A general scheme of time series data in the food supply chain is proposed. By using functional cluster analysis, the time series data are treated as a complete object of a time function, rather than a simple arrangement of individual observations. In this scheme, the discrete data are transformed into functional data while the distance of the origin function is replaced with the distance of the expansion coefficients' vector of the base function. In this way, the system can reduce the large number of numerical integration and simplify the calculation. Experiments indicate that the availability of the traceability data is enhanced significantly after discrete data are clustered into sequential information.

Key words cluster analysis; food traceability; radio frequency identification (RFID); time series data

随着社会分工的逐步细化、食品生产方式的专业化，食品供应链条的复杂化，食品由原料生产到最终消费的中间环节变得越来越多，使得食品安全涉及生产、加工、存储、运输、销售等整个食品供应链。食品从生产到消费中间环节的增加，客观上增加了引发食品安全问题的概率，食品及原料的时间特性以及食品链的复杂度都导致了安全问题分析追查的复杂程度。

目前国内外为促进食品安全投入运行的追溯系统为数不少。文献[1-2]分析了食品安全的问题并提出追溯方法。文献[3-4]提出了构建基于RFID技术的动物食品安全可溯源系统软、硬件实现方案。但由于食品数据的时效性和复杂性，以往食品溯源应用只能完成对食品及其原料本身相关数据的收集，以及一些直观的简单追溯应用。如一块市场出售的猪肉发现了微生物污染或重金属超标，传统追溯系统可以追查出该块猪肉来自哪里，但并不能解释微生物污染或重金属超标的原因，不能针对出现的问题做出深入分析并为决策部门提供相应建议。因为在生猪长达几个月的生长过程以及屠宰、运输、销售过程中，水质、饲料成分、周边环境都可能随着时间变化，传统系统无法将大量的离散数据进行有效挖掘。

本文通过对一个基于RFID技术的猪肉供应链追溯系统大量一手的数据提取，借助函数型数据分析(functional data analysis，FDA)[5-7]的基本方法和思想，在一个一般性的框架下讨论对猪肉链时序数据的聚类问题。通过数据的挖掘对生猪养殖、屠宰、流通、消费环节质量安全进行有效深入的统计、分析、监管、追踪；通过一个统一数据中心及监管平台，将相应数据结果最终发布到基于Web技术的公共网络平台。

收稿日期：2010－07－16； 修回日期：2010－10－23
基金项目：四川省科技计划公益性项目(07GF001-003)；国家科技部2010年中小型企业创新基金(10C26225123015)
作者简介：高嵘(1971－)，女，博士生，主要从事电子商务、有限自动机、食品追溯方面的研究.

1 食品时序数据函数型聚类分析

由于食品的生产加工销售是一个动态过程，质量形态随着时间变化，大部分数据中心的数据是一类时序数据，如冷鲜猪肉细菌指数、屠宰车间环境指数、运输车辆温度以及位置数据、市场各接触面菌落总数以及销售数据等，反映了属性值在时间顺序上的特征。

FDA的基本特征是把时序数据看成一个完整的关于时间的函数对象，而非个体观测值的简单排列，它所针对和处理的对象是函数，而不再是以数据表等形式出现的离散数据。使用函数型数据分析的优点有[8]：

1) 利用平滑的曲线对原始数据进行清洗，可以消除一定程度的观测误差；

2) 针对函数的积分、微分等运算可以提供丰富的分析工具；

3) 可以处理不等时间间隔取样的问题，且不同环节的取样时间可以不必相等；

4) 一旦将原始函数用特定的基函数展开(basis expansion)，则不同的基函数展开系数就捕捉了该曲线的几乎所有信息，因此，大多数情况下，FDA最终均可以简化为直接针对基函数展开系数分析，从而大大提高了运算速度。

只要将原始函数在标准正交基函数上展开，则基函数展开系数向量之间的欧式距离与原始函数之间的欧式距离是一致的；如果在非正交的基函数上展开，则只需对系数向量之间的距离进行一定修正，依然可以得到一致的结论。在这个一般框架下，大量传统的聚类方法均可以直接应用到时序数据聚类分析。本文还将这一方法推广到了多变量函数型数据，解决了多变量时序数据的聚类问题。

1.1 基函数展开

FDA总是将一系列函数(曲线)作为研究对象，而实际上只能得到函数在有限时点上的取值，即原始的时序数据。因此，FDA的首要工作是将离散的时序数据转变成连续且光滑的函数形式。

假设 $x^i(t)(i=1,2,\cdots,n)$ 是选取的n个猪肉链函数对象，而得到 $x^i(t)$ 的T_i个观察值 $y^i=(y_1^i,y_2^i,\cdots,y_{T_i}^i)'$。由于有数据误差的存在，所以有以下模型：

$$y_j^i=x^i(t_j^i)+\varepsilon^i(t_j^i) \qquad i=1,2,\cdots,n,\ j=1,2,\cdots,T_i \quad (1)$$

为了计算 $x^i(t)$，首先需要将其在一组基函数 $\Phi(t)=\{\phi_1(t),\phi_2(t),\cdots,\phi_k(t)\}'$ 上展开，即将 $x^i(t)$ 表示成基函数的线性组合：

$$x^i(t)=\sum_{k=1}^{k}c_k^i\phi_k(t) \qquad i=1,2,\cdots,n \qquad (2)$$

矩阵形式为：

$$x^i(t)=c^{i'}\Phi(t),\ \boldsymbol{c}^i=(c_1^i,c_2^i,\cdots,c_k^i)' \qquad i=1,2,\cdots,n$$

在每个时间点 t_j^i 将式(2)代入式(1)，再应用最小二乘法则，得到函数 $x^i(t)$ 的基展开系数向量为：

$$\boldsymbol{c}^i=\arg\min_{c^i}\sum_{j=1}^{T_i}X_i^2\left[y_j^i-\sum_{k=1}^{k}c_k^i\phi_k(t_j^i)\right]^2=(\boldsymbol{B}^{i'}\boldsymbol{B}^i)^{-1}\boldsymbol{B}^iy^i$$

式中，矩阵 $\boldsymbol{B}^i=(\phi_k(t_j^i))_{T\times K}$ 中的元素是第k个基函数在时间点 t_j^i 上的取值。这里每一个函数对象 $x^i(t)$ 都是利用各自的观测向量独立进行评估，因此并不要求在相同的时间点采集数据。同时，一旦给定基函数，则函数集合 $\{x^1(t),x^2(t),\cdots,x^n(t)\}$ 的信息就被系数向量集合 $\{c^1,c^2,\cdots,c^n\}$ 唯一地反映出来。

1.2 函数型数据聚类分析

聚类对象间的距离度量指标是聚类分析的首要问题，衡量距离的指标有很多种，其中欧式距离具有最优良的数学性质，使用它作为函数聚类主要的相似性度量指标。对于给定的两个函数$x(t)$和$z(t)$，其欧式距离为：

$$D_{xz}=\int_0^T(x(t)-z(t))^2\mathrm{d}t \qquad (3)$$

但如果直接使用式(3)聚类过程，整个猪肉产品链的数据需要大量数值积分，这将导致算法时间复杂度增加。为了简化运算，将$x(t)$和$z(t)$用相同的K维基函数 $\Phi(t)$ 展开，用$\boldsymbol{x}$和$\boldsymbol{z}$分别表示$x(t)$和$z(t)$的基函数展开系数向量，则有：

$$D_{xz}=\int(x(t)-z(t))^2\mathrm{d}t=\int(x'\Phi(t)-z'\Phi(t))^2\mathrm{d}t=$$

$$\int((\boldsymbol{x}-\boldsymbol{z})'\Phi(t))^2\mathrm{d}t=\int((\boldsymbol{x}-\boldsymbol{z})'\Phi(t)\Phi'(t)(\boldsymbol{x}-\boldsymbol{z}))\mathrm{d}t=$$

$$(\boldsymbol{x}-\boldsymbol{z})'\int(\Phi(t)\Phi'(t))\mathrm{d}t(\boldsymbol{x}-\boldsymbol{z})$$

令K阶方阵 $\boldsymbol{W}=\int(\Phi(t)\Phi'(t))\mathrm{d}t$，可得 $D_{xz}=(\boldsymbol{x}-\boldsymbol{z})'\boldsymbol{W}(\boldsymbol{x}-\boldsymbol{z})$。如果基函数是标准正交基，矩阵$\boldsymbol{W}$就退化成单位阵，这时函数之间的距离就变成系数向量之间的欧式距离。如果基函数非正交，D_{xz}可以被理解为系数向量之间以基函数的协差阵为权重的加权欧式距离。

这样就得到一个解决猪肉产品链各个环节时序数据聚类问题的一般框架。其一般性在以下两点体现：1) 原始时序数据能利用任意基函数展开，无关该基函数是正交还是非正交；2) 任何基于欧式距离的聚类方法(如k-means、CURE、BIRCH等)都能被

应用到时序数据聚类分析中。

由于系统在现实运行中，大量的时序数据超过两个变量，每一个多变量时序对应的函数对象由多个函数构成，如猪肉的屠宰间环境函数可以由胴体表面细菌、胴体分割面监控、车间温度等多函数构成；猪肉的重金属指标函数可以由饲料成分变化、养殖场水质变化、不同部位猪肉重金属含量等函数构成。上面的讨论结果使式(3)可以很容易扩展到多变量情形。

若函数 $x_l(t)$ 和 $z_l(t)(l=1,2,\cdots,p)$ 分别表示 p 维多变量函数 $x(t)$ 和 $z(t)$ 的第一个变量所代表的函数，则可定义函数 $x(t)$ 和 $z(t)$ 之间的欧式距离为：

$$D_{xz}=\int_0^T\sum_{l=1}^{p}(x_l(t)-z_l(t))^2\,\mathrm{d}t$$

用 $\boldsymbol{x}_l$、$\boldsymbol{z}_l$ 分别表示将用相同基函数展开的系数向量，可得：

$$D_{xz}=\sum_{l=1}^{p}((\boldsymbol{x}_l-\boldsymbol{z}_l)'\boldsymbol{W}(\boldsymbol{x}_l-\boldsymbol{z}_l)) \tag{4}$$

2　实际应用

选取2010年6月1日至2010年6月10日10天的具有连续监控数据的猪肉供应链各工序接触面及胴体表面微生物污染状况数据，通过对猪肉供应链中屠宰、运输、销售等环节微生物污染状况的监测，分析微生物污染因素。采用单位面积的细菌总数来衡量微生物污染指标，随着时间的变化，湿度和温度对猪肉表面和各接触面的影响导致细菌总数不断变化。本文总共选取13种接触面的菌落总数时序数据，其中包括刀具、工人的手、屠宰车间空气、冷却车间空气、屠宰用水、托盘、冷藏车空气、分割台面、包装材料、包装车间空气等。原始数据因篇幅原因未列出，可在文献[9]中查询。

图1是13种接触面的菌落总数拟合曲线，图中大量曲线交织在一起，很难直接从图中看出有用信息。

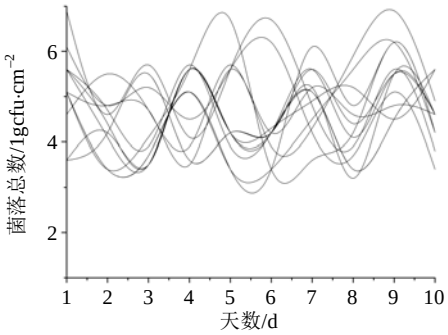

图1　接触面菌落总数拟合曲线

依据式(4)采用欧式距离作为函数间相似性度量，以k-means为例说明本文的聚类方法。图2显示了以欧式距离作为相似性度量时将13个接触面数据聚为4类的结果。从图中可以看出，这里的聚类结果基本上是将菌落变化水平相近的接触面聚在了一起(这一点可以从各个子图的纵坐标看出)，聚类结果很好地捕捉了菌落总数在形态上的相似性。由此可以给决策者和进一步的分析提供依据。

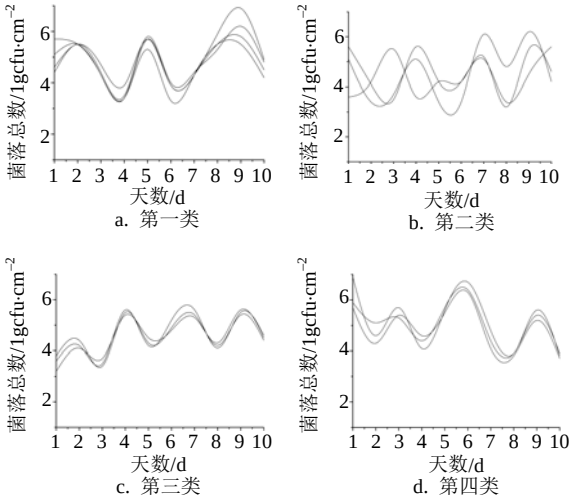

图2　依据欧式距离进行聚类的结果

如果样本量较小，还可以采用系统聚类法进行聚类，该方法可以通过聚类树形图直观观察聚类结果。具体步骤为：1) 将一阶导函数分别用相同基函数展开，得到基展开系数向量；2) 根据式(4)计算个体间距离矩阵；3) 以该距离矩阵为基础进行系统聚类。在本例中，类间距离使用Ward法度量。图3为最终所得聚类树形图。通过该图可判断出运输环节冷藏车和托盘菌落总数异常。

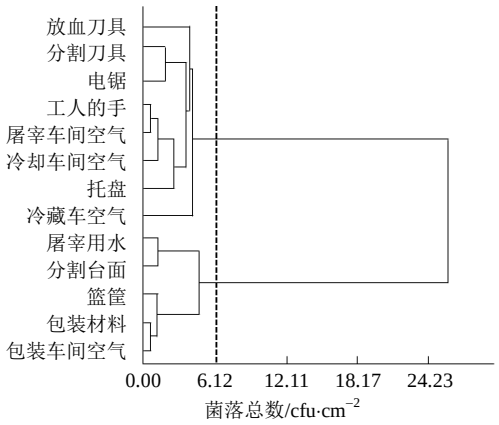

图3　系统聚类树图

(下转第591页)

LIU Yun-long, CHEN Jun-liang. A new method for software fault tolerance and its application[J]. Journal of Beijing University of Posts and Telecommunications, 1998, 21(1): 23-28.

[2] 范守文, 黄洪钟, 杨玻玻. 机电产品的容错纠错设计系统及其基本框架研究[J]. 计算机集成制造系统, 2007, 13(7): 1275-1281
FAN Shou-wen, HUANG Hong-zhong, YANG Bo-bo. Fault tolerance & fault rectification design system & its framework for electromechanical products[J]. Computer Integrated Manufacturing Systems, 2007, 13(7): 1275-1281.

[3] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报, 2000, 26(1): 32-42.
ZHANG Xue-gong. Intrduction to statistical learning theory and support vector machines[J]. Acta Automatica Sinica, 2000, 26(1): 32-42.

[4] 马笑潇, 黄席樾, 柴毅. 基于支持向量机的故障过程趋势预测研究[J]. 系统仿真学报, 2002, 14(11): 1548-1551.
MA Xiao-xiao, HUANG Xi-yue, CHAI Yi. Fault process trend prediction based on support vector machines[J]. Journal of System Simulation, 2002, 14(11): 1548-1551.

[5] 宋国杰, 唐世渭, 杨冬青. 数据流中异常模式的提取与趋势监测[J]. 计算机研究与发展, 2004, 41(10): 1754-1759.
SONG Guo-jie, TANG Shi-wei, YANG Dong-qing. Extraction and trend detection of unusual patterns over data streams[J]. Journal of Computer Research and Development, 2004, 41(10): 1754-1759.

[6] 王晶, 靳其冰, 曹柳林. 面向多输入输出系统的支持向量机回归[J]. 清华大学学报(自然科学版), 2007, 47(S2): 1737-1741.
WANG Jing, JIN Qi-bing, CAO Liu-lin. Support vector regression algorithm for multi-input multi-output systems[J]. Journal of Tsinghua University(Science and Technology), 2007, 47(S2): 1737-1741.

[7] 阎威武, 邵惠鹤. 支持向量机和最小二乘支持向量机的比较及应用研究[J]. 控制与决策, 2003, 18(3): 358-360.
YAN Wei-wu, SHAO Hui-he. Application of support vector machines and least squares support vector machines to heart disease diagnoses[J]. Control and Decision, 2003, 18(3): 358-360.

[8] 张浩然, 汪晓东. 回归最小二乘支持向量机的增量和在线式学习方法[J]. 计算机学报, 2006, 29(3): 400-406.
ZHANG Hao-ran, WANG Xiao-dong. Incremental and online learning algorithm for regression least squares support vector machine[J]. Chinese Journal of Computers, 2006, 29(3): 400-406.

编　辑　黄　莘

(上接第563页)

3　结　论

函数型数据聚类分析将食品追溯系统中原始时序数据利用某种基函数展开，得到一系列系数向量；采用一定的聚类方法，依据加权或未加权的欧式距离，对系数向量进行聚类。传统食品追溯只能解决简单追踪问题，通过函数型数据聚类分析可以从动态角度描述时序数据的类别，大大扩展了食品追溯应用的广度和深度。以大量一手数据为基础的数据分析挖掘，使普通消费者有便捷的手段对食品生产加工全工程的信息可查、明白消费，也方便政府掌握食品生产和加工的行业状况，并对食品安全事件第一时间采取正确行动。

参　考　文　献

[1] 张立钢. 建立食品安全追溯系统, 有效解决农村食品安全问题[J]. 食品安全导刊, 2009(5): 30-31.
ZHANG Li-gang. Building food safety traceability system to effectively address food safety issues in rural areas[J]. China Food Safety Magazine, 2009(5): 30-31.

[2] 卜庆婧, 梁婧晶. 建立食品安全追溯系统, 保障食品质量安全[J]. 食品安全导刊, 2009(9): 73.
PU Qing-jing, LIANG Jing-jin. Building food safety traceability system to ensure the safety of the food quality[J]. China Food Safety Magazine, 2009(9): 73.

[3] 赵金燕, 陶琳丽, 高士争, 等. 基于RFID技术的动物食品安全可溯源系统研究[J]. 云南农业大学学报, 2008, 23(4): 528-531.
ZHAO Jin-yan, TAO Lin-li, GAO Shi-zhen, et al. ttudies on animal food safety traceability system using RFID technology[J]. Journal of Yunnan Agricultural Unversity, 2008, 23(4): 528-531.

[4] CHEN R S. Using RFID technology in produce traceability[C]//Proceedings of the 10th WSEAS international conference on Mathematical methods, computational techniques and intelligent systems. Greece: [s.n.], 2008: 421-425.

[5] ESCABIASA M. Principal component estimation of functional logistic regression: discussion of two different approaches[J]. Journal of Nonparametric Statistics, 2004, 16(2): 365-384.

[6] AGRAWAL R. Mining association rules between sets of items in large databasespresented[C]//Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data. Washington, D C, USA: [s.n], 1993: 151-154.

[7] JAMES G M. Generalized linear models with functional predictors[J]. Journal of the Royal Statistical Society, 2002, 64(1): 411-432.

[8] JAMES G M, SUGAR C A. Clustering for sparsely sampled functional data[J]. Journal of the American Statistical Association, 2003, 98(2): 397-408.

[9] 成都博宇科技有限公司. 成都市生猪产品质量安全可追溯信息平台官方网站[DB/OL]. [2010-05-27]. http://www.cdspsy.com.
Chengdu Boyoi Technology Co, LTD. The quality information of chengdu porktrace[DB/OL]. [2010-05-27]. http://www.cdspsy.com.

编　辑　漆　蓉