# 基于离群点挖掘的 RFID 冷链温控研究①

赵卫东　周尚晨　孙一鸣　(复旦大学 软件学院 上海 200433)

**摘　要**: 用 RFID 技术解决了冷链物流管理中的温度实时监控难题；面对随之而来的数据爆炸问题，结合 RFID 数据挖掘算法和冷链温控实际需求，提出了 RFID 离群点快速挖掘算法 QOD，并采用剪枝策略使算法进一步优化，随后通过实验证明了算法的准确性；最后展望了 RFID 冷链温控研究的未来发展方向。

**关键词**: 冷链物流；温度传感；RFID；离群点挖掘

## Research of Temperature Control Based on Outlier Mining in RFID Cold Chain

ZHAO Wei-Dong, ZHOU Shang-Chen, SUN Yi-Ming

(School of Software, Fudan University, Shanghai 200433, China)

**Abstract:** Firstly, the chief problem of cold-chain logistics management was discussed, and the application of the RFID technology in the temperature monitoring was introduced. Then, the development of RFID data mining algorithms were studied in detail. Furthermore, a quick outlier detection algorithm (QOD) was proposed and the accuracy of QOD was proved by experimental results. Finally, the future of RFID cold chain temperature-controlled research was prospected.

**Keywords:** cold chain logistics; temperature tensor technology; RFID; outlier detection

## 1　引言

2009 年 2 月 25 日，在国务院常务会议上，物流业振兴规划被纳入国家十大产业振兴规划中，其中冷链物流的发展备受关注。中国食品工业协会资料显示，2007 年，全国水果总需求量已达到 7400 万吨，人均需求量为 55.72 公斤。预计到 2010 年，中国水果总需求量将达到 8000 万吨，人均需求量相应达到 57.31 公斤。果品产业生产总量约占世界的 14%，居世界首位。但食品冷链物流工作却不尽如人意，据统计，目前国内各类易腐食品年总产量近 7 亿吨，果蔬等农产品在采摘、运输、储藏等物流环节上的损失率高达 25 %-30 %，经济损失约达 750 亿元，可满足近 2 亿人口的基本营养需求，损耗量居世界首位，而冷链发达国家的果蔬损失率则控制在 5%以下[1]。同时，由于食品保存不当，我国多次出现食物中毒等食品安全问题事件，据有关部门介绍，我国每年约有 24 万人食物中毒，专家估计这个数字尚不到实际发生数的 1/10。由此可见，我国冷链物流还存在技术落后、损耗过大、效率偏低等问题，需要尽快采取合适的办法，解决技术瓶颈，提高质量保障度。

本文首先讨论了我国冷链物流的现状和存在问题，针对温度无法实时监控这一核心问题，引入 RFID 温度控制技术予以解决。然后重点分析了 RFID 应用后所带来数据爆炸问题，通过比较研究各类可行的 RFID 数据挖掘方法，结合冷链温控工作的需要，提出了基于离群点的快速挖掘算法 QOD，并通过实验说明其有效性。最后总结归纳了目前研究的主要问题和发展方向。

## 2　RFID冷链物流

目前冷链物流在国内方兴未艾，相关标准和模式远未成熟，也存在很多问题，尤其是温度控制方法的缺乏直接制约了冷链货物品质的提高。

### 2.1 冷链物流存在问题

近年来，不少学者对冷链物流进行了相关研究。在一系列问题中，温度是冷链物流的关键点，对于温度的监测和预警是控制冷链物流的安全和质量的核心

① 收稿时间:2010-03-14;收到修改稿时间:2010-04-12

所在[2]。现阶段我国冷链物流温控管理最大的技术瓶颈是温度监测技术手段滞后，主要症结是：人工测量和纸面记录；无统一数据系统支持；实时性差、监管脱节；取证困难、无法确定责任；无法进行预警、损失率大等。要解决这一瓶颈，就需要先引入现代的温度监测方法。

### 2.2 RFID 温度监测

目前正在研究的温度监测设备主要有以下三类，SD 卡温度记录仪，iButton 温度记录器和 RFID 冷链无线温度监控。但前两者都无法实时监控。

RFID 冷链无线温度监控系统，采用有源 RFID 技术，通过温度传感器实时获取温度数据，然后通过在运输车辆、船舶上设置 GPRS、CDMA 等实时传输装置，对目标物品进行实时监测。不仅能回溯物品损坏原因。还可以及时抢救部分物品。目前 RFID 的应用还集中在贵重商品或敏感领域。虽然在短时间内，RFID 成本过高和标准不统一等问题使其难以在冷链物流行业广泛应用，液体和金属制品对 RFID 信号的干扰、数据安全和隐私保护、RFID 标签的回收与再利用等问题也依然困扰着研究人员。但 RFID 大容量、传输范围广、装配灵活的特性决定了它非常适合解决冷链温控难题，随着 RFID 技术的发展，成本和标准问题肯定会得到解决。因此学界需要未雨绸缪，对 RFID 冷链温控开始深入的研究。

### 2.3 冷链温度数据特点

使用 RFID 技术可以大幅提高冷链物流的效率，而对 RFID 数据进行适当的分析，将会进一步提高冷链的监管和预警能力。

RFID 冷链数据具有以下特点：数据量大，半结构化，响应时间短；数据实时到达；到达顺序不受挖掘工具控制；数据量大且无法预知；数据很快会被覆盖，处理后难以再次提取[3]。由于这些特点，传统的统计方法无法提供及时的处理，因此需要引入数据挖掘方法对其进行分析。

## 3 RFID冷链数据挖掘

RFID 冷链数据的收集和挖掘过程必须同时进行，但挖掘目标相对稳定，因此需要建立快速挖掘模型来响应用户的实时查询。在 RFID 冷链的数据挖掘过程中，需要关注以下方面：内存的限制、系统的实时响应、单次扫描、结果的近似性和算法的自适应性[4]。

### 3.1 常用挖掘方法的优劣

对 RFID 数据进行挖掘通常有三类方法：聚类、分类和频繁模式挖掘。

传统的聚类方法已经不能适用于动态的 RFID 冷链温度数据，算法只能适用于纯数值属性数据，不能适用于多属性的数据，这也是许多传统聚类算法存在的问题。

当前 RFID 温度数据分类过程主要面临两个问题。一是数据在内存中的表示问题：内存大小是有限的，而不断出现的数据必须得到实时的处理。二是概念漂移问题：即从训练数据中所需要学习的概念是随时间不断变化的，并且这种概念的变化程度以及漂移的具体位置都是未知的。RFID 数据的内存表示问题可以通过使用增量学习的方法，训练一个概要数据结构来实现。而概念漂移问题，至今仍然没有理想的解决方法。目前比较有效的解决方案是使用固定大小的滑动窗口来装载 RFID 数据流中的数据，分块构建或更新所构建的分类模型，并以最新的分类模型来响应用户的分类要求。

频繁模式挖掘是数据挖掘的一个重要研究分支，频繁模式挖掘的关键步骤就是如何快速准确地对项(集)进行频度计数，以获得满足最小支持度要求的频繁项(集)。概括地说，现有的适合 RFID 冷链温度数据频繁模式挖掘算法大致可分为如下两大类[5]：基于概率误差区间的近似算法和基于确定误差区间的近似算法。RFID 数据流频繁模式的挖掘算法通常会面临空间和时间的问题。然后，两者之间存在矛盾，即当某个算法空间复杂度较低的时候通常会浪费大量的时间作为代价；而时间复杂度较低的时候通常会占用太多的存储空间。因此两者之间的平衡问题成了数据流频繁模式挖掘的难点。

综上所述，聚类、分类和频繁模式挖掘方法对于 RFID 冷链数据的分析都难以令人满意，因此需要结合 RFID 冷链温度数据挖掘的需求来选取更合适的方法。

离群点挖掘方法

在冷链温度数据挖掘中，挖掘目的是找出那些与集中的大多数温度数据有显著差异的数据，也就是说，罕见事件通常比常规事件更有吸引力。因此可以采用离群点挖掘方法来监控温度的异常波动。

RFID 冷链数据的特点决定了在构建一个数据流挖掘模型时，变化到达之前的数据可能偏离那些与不

再保持的特征相关的模型[5,6]。目前绝大多数的研究集中在通过丢弃旧数据或给其较小的权重来匹配变化分布的算法[5]。

文献[7]提出了检测数据流什么时候发生变化的方法，采用参照窗口(reference window)和滑动窗口(sliding window)。每当一个新数据点出现时，滑动窗口向前滑动一次，而参照窗口当且仅当检测到数据流中出现变化时才进行更新。

杨宜东等利用动态网格对空间中的稠密和稀疏区域进行划分，过滤处于稠密区域的大量主题数据，而对于稀疏区域中的候选离群点，采用近似方法计算其离群度，具有较高的离群度的数据就作为离群点输出[8]。

周晓云等提出针对数据流特点提出基于加权频繁模式离群因子(WFPOF)的高维数据流离群点检测算法FODFP-Stream[9]。该算法通过动态发现和维护频繁模式来计算离群度，能有效处理高维类别属性数据流，并进一步扩展到数值属性和混合属性数据流，通过数据衰减系数的设定，可以有效地处理数据流数据中的概念漂移问题。

上述离群点检测方法都能够对 RFID 冷链温度数据进行动态分析，但也存在资源占用过多、响应速度不够、算法复杂度太大等问题。本文借鉴已有研究成果，结合冷链物流温度监控的特点，提出了一种新的离群点快速挖掘算法 QOD。

## 4　离群点快速挖掘算法QOD

### 4.1 相关定义和性质

设对象集 $X=\{x_1,x_2,\ldots,x_n\}$，$c_c$ 表示在指定条件 c 下与环境相关的关系。

定义 4.1(局部邻居). 对象 $x_i$ 的局部邻居是指与对象 $x_i$ 在指定条件 c 下，存在与环境相关的关系 $c_c$ 的对象。

定义 4.2(局部邻域). 对象 $x_i$ 的局部邻域 $N(x_i)$ 是指对象 $x_i$ 的所有局部邻居的集合。

定义 4.3(邻居权值). 设 $x_i,x_j\in X$，且 $x_j\in N(x_i)$，$x_i$ 的第 j 个邻居的邻居权值定义为：

$$W_{ij}=\sum_{p=1}^{q}a_p\cdot\frac{G_{pj}}{\sum_{r=1}^{N(xi)}G_{pr}} \tag{1}$$

其中 q 表示确定权值的最大环境特性数目，$G_{pr}$ 表示对象 $x_i$ 的第 r 个邻居的环境特性 $G_p$ 的值，$a_p$ 是确定环境特性 $G_p$ 重要性的因子，且 $\sum_{p=1}^{q}a_p=1$ 如果环境属性在对象 $x_i$ 与其局部邻居 $x_j$ 为中的影响被忽略(即 q=0)，那么指定权值为 $\frac{1}{|N(xi)|}$，相当于同等考虑对象 $x_i$ 局部邻居的影响。

利用归一化技术对数据集进行归一化处理，设 max 和 min 是固有属性的最大值和最小值，$f(x_i)$ 是对象 $x_i$ 的固有属性值。设

$$F(x_i)=\frac{f(x_i)-\min}{\max-\min} \tag{2}$$

以保证 $F(x_i)\in[0,1]$。

由上述邻居权值定义，并结合欧氏距离公式，给出加权距离的定义。

定义 4.4(加权距离). 设 $x_i,x_j\in X$ 且 $x_j\in N(x_i)$，$x_i$ 和 $x_j$ 的固有属性是 $f(x_i)$ 和 $f(x_j)$，归一化属性是 $f(x_i)$ 和 $f(x_j)$，且 $F(x_i),F(x_j)\in[0,1]$。则 $x_i$ 和 $x_j$ 之间的距离为：

$$D(x_i,x_j,W_{ij})=\sqrt{\sum_{k=1}^{n}W_{ij}\cdot[F(x_i)-F(x_j)]^2} \tag{3}$$

定义 4.5(局部离群度). 对象 $x_i$ 的局部离群度表示为：

$$LO(x_i)=\frac{1}{|N(xi)|}\sum_{i=1}^{|N(xi)|}D(xi,xj,Wij),x_j\in N(x_i) \tag{4}$$

其中，$N(xi)$ 表示对象 $x_i$ 的局部邻居的个数，局部邻域 $N(xi)$ 中的最大局部邻域离群度记为 $MaxLO(N_+(x_i))$，其中 $N_+(x_i)=N(x_i)\cup\{x_i\}$

定义 4.6(全局近似离群度). 全局近似离群度是固有属性在整个数据集中的离群度的近似平均值，表示为：

$$GAD=\sqrt{\sum_{k=1}^{n}\overline{LO_{ek}}^2} \tag{5}$$

其中 $\overline{LO_e}$ 表示每一维子空间 e 中所有对象的局部离群度平均值

定义 4.7(局部近似离群度). 局部近似离群度是固有属性在对象 $x_i$ 局部邻域 $N(xi)$ 上离群度的近似平均值，表示为：

$$LAD=\sqrt{\sum_{k=1}^{n}\overline{\sum_{i=1}^{N}LO(x_i)_{ek}}^2} \tag{6}$$

其中 $LO(x_i)_{ek}$ 表示每一维子空间 e 中对象 $x_i$ 的局部离群度。

定义 4.8(局部离群度因子). 对象 $x_i$ 的局部离群度因子定义为：

$$LOF(x_i)=\frac{LO(x_i)+e}{LO(x_j)+e}, x_j\in N(x_i) \quad (7)$$

设 $e$ 为足够小的正数，以避免计算中分母为 0。最终获得离群点，需将有对象的 LOF 值从大到小排列，而当分子分母同加上一个非常小的正数 $e$ 时，不会改变 LOF 原有顺序。实验中取

$e=\min(\min\{LO(x_i)\neq 0, x_i\in X\}, MIN)$，其中 MIN 为程序可计算最小精度值。

上述计算中，所有固定属性均归一化到 $[0,1]$ 区间，且 $W_{ij}\le 1$，故有 $\frac{e}{1+e}\le LOF(x_i)\le\frac{1+e}{e}$，$e$ 的取值决定 $LOF(x_i)$ 的取值范围。

当对象的局部离群度为 0 时，表示对象与其邻域对象的固有属性值相同，$LOF(x_i)=0$；

当对象的离群度与邻域对象的平均离群度相同时，表示对象的固有属性在有规律地变化，$LOF(x_i)=1$。

故当 $LOF(x_i)\le 1$ 时，对象正常；当 $LOF(x_i)\ge 1$ 时，对象开始离群；随着 $LOF(x_i)$ 值的增大，其离群程度也增大。

定义 4.9 (离群点). 给定 n 个对象的数据集 X，需要检测的离群点数目为 m，则计算所得 $LOF(x_i)$ 最大的前 m 个对象就是离群点。

在 RFID 冷链物流体系中，全局近似离群度 GAD 代表某条物流线路中温度的正常波动范围。相应的，局部近似离群度 LAD 就代表某段线路里温度的波动范围。离群点则代表了出现异常温度的情况，快速挖掘离群点有助于迅速发现冷链物流的故障点，减少损失。

根据基于正态分布的离群点的定义：若对象 $y_i\in Y$，$\left|\frac{y_i-m}{s}\right|\ge 3$，则 $y_i$ 为离群点，即偏离平均值 $m$ 超过 $3s$ 的数据点就是离群点[10]。相反，如果某个数据点的离群值低于相应邻域近似平均离群值的 1/3，则认为这样的数据点不可能是离群点，即如果对象 $y_i\in Y$，$\left|\frac{y_i-m}{s}\right|<C, C=\frac{1}{3}$，则认为对象 $y_i$ 不可能是离群点。同样地，如果这个数据点的整个局部邻域中最大局部离群值 $MaxLO(N_+(x_i))$ 低于整个数据集中近似平均离群值的 $\frac{1}{3}$，通常认为这样的邻域 $N_+(x_i)$ 不可能包含离群点。由此得出两个性质。

性质 4.1. 存在对象 $x_i\in X$，若 $MaxLO(N_+(x_i))<C\cdot GAD$，则局部邻域 $N_+(x_i)$ 中都不可能包含候选离群点，故整个局部邻域 $N_+(x_i)$ 可从数据集中剪枝。

性质 4.2. 若在对象 $x_i\in X$ 的局部邻域 $N(x_i)$ 中，有 $LO(x_i)<C\cdot GAD$，则对象 $x_i$ 不可能成为离群点，故对象 $x_i$ 可从数据集中剪枝。

运用这两条性质，将数据集中一部分不可能成为离群点的数据对象从数据集中剔除，这就减少了空间离群度因子的计算量，从而提高算法的效率。

由此表明，在 RFID 冷链物流应用中，若某一段温度数据的最大偏离值都在允许的范围之内，则可以认为这一段数据都正常，其中没有离群点。

### 4.2 算法描述

上文给出了离群点快速挖掘算法需要的相关定义和性质，下面给出 QOD 算法的具体流程。

输入：对象集 $X=\{x_1, x_2,..., x_n\}$，固有属性函数为 $f(x_i)$，$c_c$ 表示在指定条件 c 下与环境相关的关系

输出：离群点集。

快速离群点检测算法 QOD 的算法过程：

步骤 1. 数据归一化

运用式(2)将 $f(x_i)$ 归一化，保证 $F(x_i)\in[0,1]$；

步骤 2. 计算加权距离

运用式(3)计算对象与其局部邻居的加权距离；

步骤 3. 计算局部离群度

运用式(4)计算对象与其局部邻域中对象的离群度，同时将每个局部邻域中的最大局部离群度值放入相应的最大局部邻域离群度中 $MaxLO(N_+(x_i))$；

步骤 4. 剪枝当前对象

根据性质 4.1 和性质 4.2，分别对数据集中的局部邻域和逐个对象进行剪枝

步骤 5. 计算局部离群度因子并输出离群点

对剪枝后遗留的数据集，运用式(7) 计算局部离群度因子 $LOF(x_i)$ ;并将所求的各点的 $LOF(x_i)$ 值按降序排列，并输出前 m 个对象作为离群点。

算法的总复杂度为 $O(kn\log_s n)$ 。

## 5　实验结果比较

本文使用的数据集为 Columbia 大学冷链运输项目实验数据集，该数据集中测试样本 8124 组。整个实验从中随机选出 500 组数据，运行在主频为 P4 2.4GHz、内存 1GB、硬盘 160 GB、操作系统为 Windows XP sp2 的主机上。基于此环境，对整个算法进行相关性能测试。

表 1　不同算法对测试错误率的比较

| 算法类别 | 测试错误率% |
|---|---|
| 本文提出的算法 | 5.77（±0.63） |
| 基于滑动窗口的算法 | 18.30（±0.65） |
| 基于动态网格划分算法 | 22.79（±0.61） |
| FODFP-Stream 算法 | 11.23（±0.66） |

表 1 给出了分别使用不同的离群点检测算法和本文提出的算法对 Columbia 数据集进行学习和离群点分离的结果。本文提出 QOD 算法的测试分离离群点的错误率明显低于其他离群点检测算法的测试错误率。

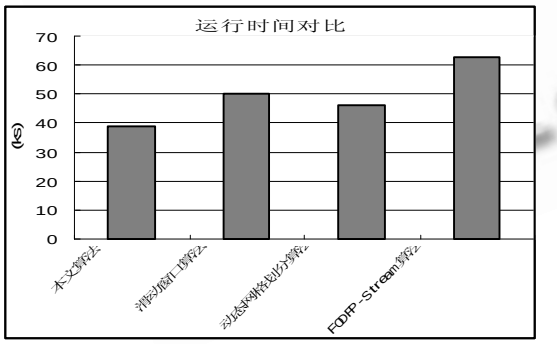


图 1　不同算法平均运行时间的比较

图 1 给出了面对相同数据集，分别使用不同的离群点检测算法和本文提出的 QOD 算法平均执行时间的比较。本文提出算法运行时间略低于其他离群点检测算法，说明算法在保证正确率的前提下，并没有损失速度。

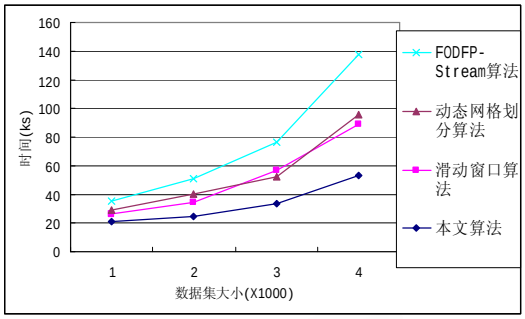


图 2　不同算法随数据集变化的时间性能比较

图 2 给出了面对同步增长的数据集，不同的离群点检测算法和本文提出的算法平均执行时间的比较。本文提出的 QOD 算法运行时间增长速度低于其他离群点检测算法，说明该算法具有良好的特性，在一定范围内比较稳定。

QOD 算法针对冷链温度数据相对连续、维度单一的特点，在计算时简洁直接，不用考虑窗口滑动、降维近似等问题，引入近似离群度剪枝后进一步提高了算法速度，因此在精度和速度上都优于传统离群点挖掘算法。

## 6　总结

冷链物流是未来食品物流领域一个经济增长点，采用数据挖掘方法处理实时 RFID 温度数据，可以有效地对冷链温度实现全自动实时监控。本文所提出的 QOD 算法，利用全局近似离群度和局部近似离群度的两个性质来实现对 RFID 温度数据集的剪枝，实验结果证明算法在处理冷链温度数据时，具有减少用户依赖度，降低算法复杂度和提高精度的优点，有效解决了使用 RFID 技术所带来的海量数据处理问题。

在今后的研究中，需要对算法所检测出的离群点进行解释以及结合领域知识构造规则库，利用规则库对离群数据进行深入分析也是下一步的研究工作，为了更好地使人们理解检测出的离群数据，算法的可视化研究也是一个重要的方向、检测过程中与用户的交互、检测结果的可视化更能帮助用户加深对数据的理解，提高算法的准确性。

### 参考文献

1 鲍长生.冷链物流运营管理研究[硕士学位论文].上海:同济大学, 2007.

(下转第 175 页)

(上接第 170 页)

2 阎君.食品冷链物流市场化研究[硕士学位论文].北京:北京交通大学, 2007.

3 金澈清,钱卫宁,周傲英.流数据分析与管理综述.软件学报, 2004,15(8):1172－1181.

4 潘云鹤,王金龙,徐从富.数据流频繁模式挖掘研究进展.自动化学报, 2007,32(4):594-602.

5 Hulten G, Spencer L, Dom Ingos P. Mining time changing data streams. Proc of SIGKDD'01. New York: ACM Press, 2001:97－106.

6 Aggarwal CC, Han Jiawei, Wang Jianyong, et al. A framework for clustering evolving data streams. Proc of VLDB'03. Berlin: VLDB Endowment, 2003: 81－92.

7 Kifer D, Ben2dayav ID S, Gehrke J. Detecting change in data streams. Proc of VLDB'04. Toronto: VLDB Endowment, 2004:180－191.

8 杨宜东,孙志挥,朱玉全,等.基于动态网格的数据流离群点快速检测算法.软件学报, 2006,17(8):1796－1803.

9 周晓云,孙志挥,张柏礼,等.高维类别属性数据流离群点快速挖掘算法.软件学报, 2007,18 (4):933－942.

10 Liu Ying, Alan P. Sprague. Outlier detection and evaluation by net work flow. International Journal of Computer Applications in Technology, 2008,33(2-3), 237－246.