doi: 10.3969/j.issn.1005-3026.2016.12.002

一种面向不确定数据流的聚类算法

韩东红[1], 王 坤[1], 邵崇雷[2], 马 畅[1]
(1. 东北大学 计算机科学与工程学院, 辽宁 沈阳 110169; 2. 沈阳理工大学 机械工程学院, 辽宁 沈阳 110159)

摘 要: 作为大数据的重要组成,产生于传感器、移动电话设备、社交网络等的不确定流数据因其具有流速可变、规模宏大、单遍扫描及不确定性等特点,传统聚类算法不能满足用户高效实时的查询要求. 首先利用MBR (minimum bounding rectangle) 描述不确定元组的分布特性,并提出一种基于期望距离的不确定数据流聚类算法,计算期望距离范围的上下界剪枝距离较远的簇以减少计算量;其次针对簇内元组的分布特征提出了簇MBR的概念,提出一种基于空间位置关系的聚类算法,根据不确定元组MBR和簇MBR的空间位置关系排除距离不确定元组较远的簇,从而提高聚类算法效率;最后在合成数据集和真实数据集进行实验,结果验证了所提出算法的有效性和高效性.

关 键 词: 不确定数据流;聚类;大数据;数据挖掘;最小边界矩形
中图分类号: TP 391 **文献标志码**: A **文章编号**: 1005-3026(2016)12-1677-06

A Cluster Algorithm for Uncertain Data Stream

HAN Dong-hong[1], *WANG Kun*[1], *SHAO Chong-lei*[2], *MA Chang*[1]
(1. School of Computer Science & Engineering, Northeastern University, Shenyang 110169, China; 2. School of Mechanical Engineering, Shenyang Ligong University, Shenyang 110159, China. Corresponding author: HAN Dong-hong, E-mail: handonghong@cse.neu.edu.cn)

Abstract: As an important component of big data generated in the sensor, mobile phone devices, social networks etc., uncertain streaming data have many characteristics, such as variable rate, large-scale, single-pass scanning, and uncertainty. Traditional clustering algorithms cannot meet efficient real-time inquiry requirements for the users. Firstly, MBR (minimum bounding rectangle) was used to describe the distribution characteristics of uncertain tuples. And then, a clustering algorithm based on expected distance was proposed for uncertain data stream. The bounds of expected distance range to filter the clusters with far distance can be calculated. Secondly, cluster MBR concept based on the distribution of the tuples in a cluster was presented. Then, a clustering algorithm was given, which excludes the clusters far from the uncertain tuple by the spatial location relationship between uncertainty tuple MBR and clusters MBR, thereby increasing the efficiency of clustering algorithm. Finally, experiments running on synthetic datasets and real datasets verify that the proposed algorithms are effective and efficient.

Key words: uncertain data stream; cluster; big data; data mining; MBR (minimum bounding rectangle)

近年来,社交网络、移动电话应用、电子商务网站等产生的数据呈指数级增长,面向大数据的分析和处理技术的研究方兴未艾. 作为大数据的一种数据模型,广泛存在于实时监控系统、基于位置的服务(LBS)[1]、传感器网络[2]、射频识别电子标签(RFID)[3-4]等的不确定数据流,因其具有数据量规模宏大、高速、单遍扫描及不确定性等特征,需设计高效实时的增量算法对其进行聚类分析.

数据的不确定性是由数据采集及传输、数据集成等原因产生的,可分为元组级不确定性和属性级不确定性[5-6]. 数据流环境下聚类分析面临

收稿日期: 2015-08-28
基金项目: 国家自然科学基金资助项目(61173029,61332006,61672144).
作者简介: 韩东红(1968-),女,河北平山人,东北大学副教授.

的主要挑战是对源源不断到来的数据进行实时高效处理,不确定性的引入增加了解决这一挑战的难度.

Aggarwal 等[7]提出了最早的聚类演化数据流的双层框架结构——CluStream 算法,将聚类过程分为在线聚类和离线聚类两部分. Aggarwal 等在 2008 年提出了 UMicro 算法[8]以解决属性级不确定数据流的聚类问题. 文献[9]提出一种基于投影空间的不确定数据流聚类算法 UPStream. 针对离散概率模型中的元组级不确定性问题,文献[10]提出基于信息熵的不确定数据流聚类算法 LuMicro. 文献[11]提出通过利用不确定聚类特征的指数直方图来获取最新数据的分布特征的方法,采用双层架构模型对不确定数据流进行聚类. 文献[12]考虑不确定元组期望值和不确定性,提出基于 Voronoi 图的聚类算法以减少滑动窗口中期望距离计算量. 文献[13]引入动态更新以适应数据变化的免疫模型,提出对元组级不确定流数据进行聚类的 IUMicro 算法. 文献[14]提出 UIDMicro 聚类算法,分别用标准差和区间数来表示不确定流数据,同时采用 two-layer 聚类窗口模型和动态聚类模型更新策略对不确定数据流聚类. 文献[15]提出 UDSSC 算法使用滑动窗口机制接收新数据,引入子空间簇生成策略和新型离群点机制. 文献[16]提出基于不确定数据流聚类的动态通信距离评估方法.

本文侧重研究离散概率模型表示的元组级不确定数据流的聚类算法,旨在提高算法执行效率.

1 不确定数据流聚类算法

1.1 基于期望距离的不确定数据流聚类算法

1.1.1 相关定义

定义 1 不确定数据流. 若干随时间变化的 d 维不确定元组构成不确定数据流 S,$S=\{(X_1,t_1),(X_2,t_2),\cdots,(X_i,t_i),\cdots\}$,其中 X_i 是一个 d 维的元组,由 k_i 个可能的实例组成,$X_i=(<x_{i_1},p_{i_1}>,<x_{i_2},p_{i_2}>,\cdots,<x_{i_{k_i}},p_{i_{k_i}}>)$,$p_{i_j}$表示第 j 个实例出现的概率,且 $0<p_{i_j}\leqslant 1$,$\sum_{j=1}^{k_i}p_{i_j}=1$,$t_i$ 表示元组 X_i 到达的时间戳.

定义 2 期望距离. X_i 和簇心 c_j 的期望距离为 X_i 内所有可能实例到 c_j 距离的期望之和,即

$$\mathrm{ED}(X_i,c_j)=\sum_{s=1}^{k_i}d(x_{i_s},c_j)p_{i_s}. \tag{1}$$

定义 3 不确定元组的 MBR. 即包含元组内所有可能实例的最小边界矩形,用以下 d 维向量分别表示 MBR 的下界和上界:

$\boldsymbol{lower}=(\min(x_i^{(1)}),\min(x_i^{(2)}),\cdots,\min(x_i^{(d)}))$,

$\boldsymbol{upper}=(\max(x_i^{(1)}),\max(x_i^{(2)}),\cdots,\max(x_i^{(d)}))$.

其中:$\min(x_i^{(j)})$表示 X_i 在第 j 维上的最小值;$\max(x_i^{(j)})$表示 X_i 在第 j 维上的最大值.

定义 4 微簇的聚类特征结构. 包含 d 维的 X_i 的微簇聚类特征结构用$(2\times d+2)$元组$(\mathbf{CF}_1,\mathbf{CF}_2,t,n)$表示,其中 $\mathbf{CF}_1$ 是 d 维向量,为每个不确定元组期望值的线性和,第 q 项存储内容为 $\sum_{i=1}^{n}\sum_{j=1}^{k_i}x_{i_j}^{(q)}p_{i_j}$,$k_i$ 是第 i 个不确定元组内实例数目;$\mathbf{CF}_2$ 是 d 维向量,为不确定元组期望值平方和;t 表示微簇最后更新时间;n 表示微簇内不确定元组个数.

引理 1 当微簇 C 加入新不确定元组 $<X_i,t>$,微簇聚类特征结构的各项均可增量更新(证明略).

定理 1 不确定元组 X_i 的 MBR 内至少存在一个确定点 x',使得 x'到簇心 c_j 的距离与 X_i 到簇心 c_j 的期望距离相等,即 $d(x',c_j)=\mathrm{ED}(X_i,c_j)$.

定理 2 若 y_i 为 X_i 的 MBR 的几何中心,对角线长度为 $2r$,$\mathrm{ED}(X_i,c_j)$为 X_i 到点 c_j 的期望距离,则

$$d(y_i,c_j)-r\leqslant\mathrm{ED}(X_i,c_j)\leqslant d(y_i,c_j)+r. \tag{2}$$

1.1.2 算法描述

本文提出聚类不确定数据流的 EDMicro 算法,使用定义 4 的聚类特征结构在线维护微簇,并使用算法 1 作为在线聚类的处理流程,其在线微簇作为输出参与后续的离线宏聚类. 行 7 中,若不确定元组 X_i 到 c_j 的期望距离小于 Threshold,则当前点属于该微簇,否则 X_i 属于一个新微簇或是离群点,Threshold 的设置方法同文献[10]. 若 X_i 属于一个新微簇或为离群点,则为其建立一个新微簇. 如果当前微簇的个数小于 n_{micro},直接新建微簇,否则删除当前微簇集中最久未更新的微簇. EDMicro 算法的伪代码见算法 2.

算法 1 不确定数据流在线聚类算法

输入:最大微簇数目 n_{micro},不确定数据流 S

输出:微簇的集合 C

1. REPEAT
2. 从 S 中读入新元组 $<X_i,t_i>$;
3. IF X_i 是第一个元组
4. 直接为 X_i 创建一个新的微簇;
5. ELSE
6. CALL 某种算法找到距离 X_i 最近的微

簇 c_j；
7. IF ED(X_i,c_j)≤Threshold
8. X_i 属于簇 C_j 并更新微簇的概要信息；
9. ELSE
10. IF (|C|≥n_{micro})
11. 将最久未更新的微簇删除；
12. 对应 X_i 创建一个新微簇；
13. ELSE
14. 直接创建一个以 X_i 为中心的新微簇；
15. ENDIF
16. ENDIF
17. ENDIF
18. UNTIL stream end

算法 2　基于期望距离的 EDMicro 算法
输入：微簇集合 C 及不确定流数据 X_i
输出：距离 X_i 最近的微簇
1. 初始化候选簇集合 Q；
2. Q←C；
3. 计算不确定流数据 X_i 的 MBR 的对角线一半值；
4. FOR k = 1 to |Q|
5. 计算该元组 MBR 的中心点和与 C_k 簇心 c_k 的距离；
6. 计算 X_i 和 C_k 的期望距离 ED 的上界及下界；
7. IF (upper_ED < min_upper_ED)
8. min_upper_ED = upper_ED；
9. ENDIF
10. ENDFOR
11. FOR k = 1 to |Q|
12. IF (lower_ED(X_i,C_k) > min_upper_ED)
13. 将 C_k 从候选簇集合 Q 中删除；
14. ENDIF
15. ENDFOR
16. FOR k = 1 to |Q|
17. 计算 X_i 和候选集合 Q 中剩余微簇的期望距离；
18. ENDFOR
19. RETURN 具有最小期望距离的微簇.

1.2　基于空间位置关系的不确定数据流聚类算法

1.2.1　相关定义

定义 5　微簇的 MBR. 包含微簇内所有不确定元组期望值的最小边界矩形，以两个 d 维向量分别表示 MBR 的下界和上界：
$\mathbf{lower}=(\min(x_i^{(1)}),\min(x_i^{(2)}),\cdots,\min(x_i^{(d)}))$，
$\mathbf{upper}=(\max(x_i^{(1)}),\max(x_i^{(2)}),\cdots,\max(x_i^{(d)}))$.
其中：$\min(x_i^{(j)})$ 表示 MBR 在第 j 维上的最小值；$\max(x_i^{(j)})$ 表示 MBR 在第 j 维上的最大值.

定义 6　含 MBR 的微簇聚类特征结构. 包含 d 维不确定元组的微簇可用 $(4\times d+2)$ 的元组 $(\mathbf{CF}_1,\mathbf{CF}_2,\mathbf{lower},\mathbf{upper},t,n)$ 表示特征结构，其中：$\mathbf{CF}_1$ 表示各不确定元组期望值的线性和，即每一维存储不确定元组对应维的期望值的和，它是 d 维向量；$\mathbf{CF}_2$ 也是 d 维向量，表示各不确定元组期望值的平方和；**lower** 是 d 维向量，表示微簇的 MBR 的下界，其第 q 项为 $\min(x_i^{(q)})$；**upper** 是 d 维向量，表示微簇的 MBR 的上界，其第 q 项为 $\max(x^{(q)})$；t 表示微簇最后更新的时间；n 为微簇内不确定元组的个数.

引理 2　当新的不确定元组加入微簇，包含 MBR 信息的微簇聚类特征结构的各项均可增量更新.

1.2.2　MBR 的空间位置关系

以二维数据为例，不确定元组 X_i 的 MBR 和微簇 MBR 的位置关系包括：

包含 X_i 的 MBR 和微簇的 MBR，其中一个落在另一个内部；

相交 X_i 的 MBR 和微簇的 MBR 的空间位置部分重合但无包含关系；

相离 X_i 的 MBR 和所有微簇的 MBR 都没有共同区域. 意味着 X_i 距离所有微簇均较远，需计算 X_i 和所有微簇的期望距离.

1.2.3　边界情况的优化

通常与 X_i 相交或者包含 X_i 的微簇都是距离 X_i 较近的微簇，与 X_i 相离的微簇都与之较远. 但某些特殊情况下，与 X_i 相离的微簇也可能距离 X_i 更近. 本文给出一种启发式解决方法，即对 X_i 的 MBR 放大以便能够与距其较近的微簇相交或者包含. 可以看出，放大后的 MBR 与其较近范围内微簇的 MBR 均有相交或包含的关系. 参数 τ 控制 X_i 的 MBR 的放大倍数，其合理取值本文将在实验部分给出.

1.2.4　算法描述

聚类过程同样使用算法 1 的处理流程，SRMicro 算法描述见算法 3，集合 Q_1,Q_2,Q_3 分别存放包含、相交、相离的微簇索引.

算法 3　SRMicro 算法
输入：微簇集合 C，不确定元组 X_i
输出：距离 X_i 最近的微簇

```
1.  创建 3 个候选簇集合 Q1,Q2,Q3;
2.  Q1←NULL,Q2←NULL,Q3←NULL;
3.  FOR k = 1 to |C|
4.    判断微簇 Ck 的 MBR 和 Xi 的 MBR 之间
      的位置关系;
5.    SWITCH
6.    {CASE 包含:Ck 加入到候选集合 Q1 中;
      BREAK;
7.     CASE 相交:Ck 加入到候选集合 Q2 中;
       BREAK;
8.     CASE 相离:Ck 加入到候选集合 Q3 中;
       BREAK;}
9.  ENDFOR
10.  IF  (|Q1|≥1)
11.   FOR  j = 1 to |Q1|
12.      计算 Xi 和 Q1 中所有微簇的期望距离
         ED(Xi,cj);
13.   ENDFOR
14.   Return argminc ED(Xi,cj);
15.  ELSE IF (|Q2|≥1)
16.   FOR  j = 1 to |Q2|
17.     计算 Xi 和 Q2 中所有微簇的期望距离
        ED(Xi,cj);
18.   ENDFOR
19.   Return argminc ED(Xi,cj);
20.   ELSE
21.     FOR  j = 1 to |Q3|
22.      计算 Xi 和 Q3 中所有微簇的期望距离
         ED(Xi,cj);
23.     ENDFOR
24.     Return argminc ED(Xi,cj);
25.   ENDIF
26.  ENDIF
```

2 实验结果分析

对本文提出的算法与 LuMicro 算法[10]进行比较. 实验所用计算机内存 2 GB DDRII, CPU 为 Intel(R) Core(TM)2 Duo E8400 @3.00 GHz,操作系统采用 Microsoft Windows XP SP3,开发环境为 Microsoft Visual Studio 2010,编程语言选择 C++.

2.1 数据集及参数设置

为评估算法性能,采用两个真实数据集分别是 KDD-CUP'99 网络入侵检测数据集和美国联邦森林数据集(Forest CoverType),合成数据集采用人工的方式生成. 算法中使用的默认参数设置如表 1 所示.

表 1 默认参数设置

Table 1 Default parameter setting

参数	默认值	含义
N	500 000	数据流最大尺寸
d	0	元组维度
α	3	半径阈值
τ	1	MBR 放大倍数
β	10	不确定元组内的最大实例数目
n_{micro}	500	微簇最大个数

2.2 算法性能分析

2.2.1 效率和有效性

图 1 给出三种算法在不同数据集上的聚类时间,EDMicro 算法运行时间要远低于 LuMicro 算法. 图 2 给出了三种算法以纯度表示的实验结果,本文所提出算法的聚类纯度均高于 LuMicro 算法.

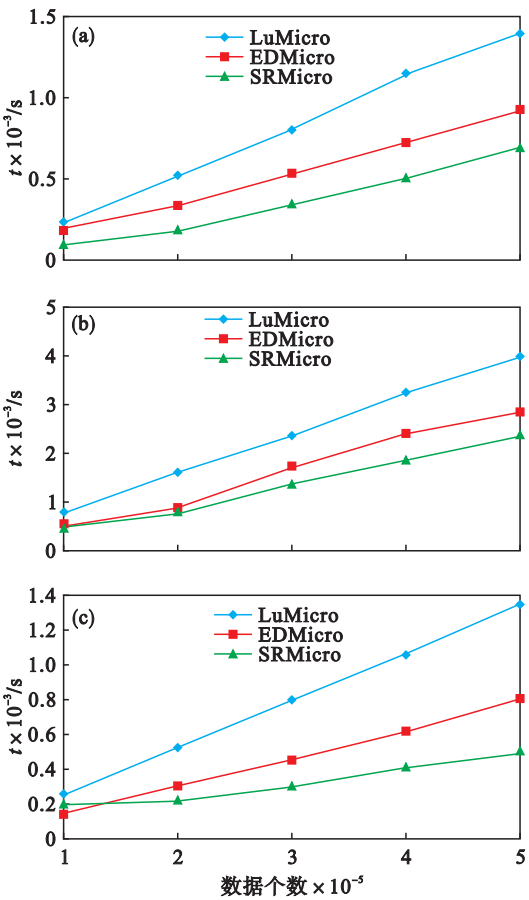

图 1 聚类时间

Fig. 1 Clustering time

(a)—合成数据集;(b)—网络入侵数据集;(c)—森林数据集.

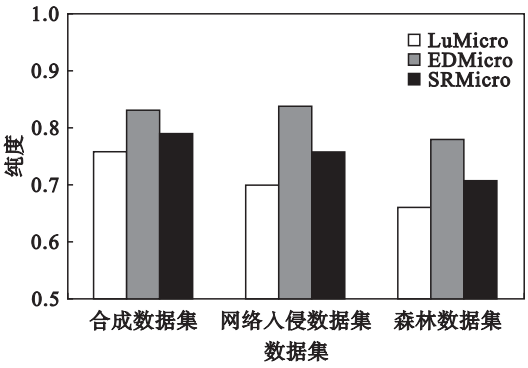

图 2　聚类纯度

Fig. 2　Purity of clustering

2.2.2　参数影响

图 3 给出了 β 分别取 10,20,30,40 时,本文所提出算法与 LuMicro 算法执行时间的对比. 算法执行时间都随着 β 值的增长而增加,EDMicro 算法时间增长要慢于 LuMicro 算法, SRMicro 算法的执行时间缓慢增长,这是因为比较的代价要远小于浮点运算的代价.

图 4 显示了参数 τ 对聚类结果的影响. 本文通过放大不确定元组的 MBR,使其尽量与距离较近的微簇有交集,以保证算法的准确度. 参数 α 表示半径阈值,图 5 显示了 α 在不同值的情况下,算法的聚类纯度与时间.

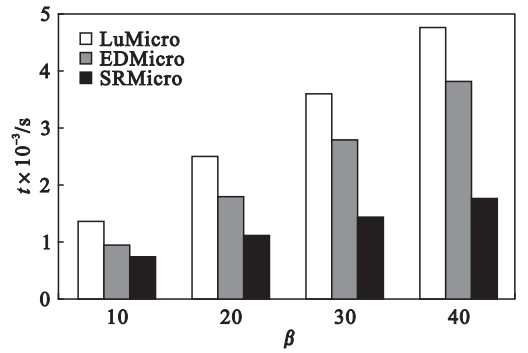

图 3　不同实例数量对执行时间的影响

Fig. 3　Influence on running time with different numbers of instances

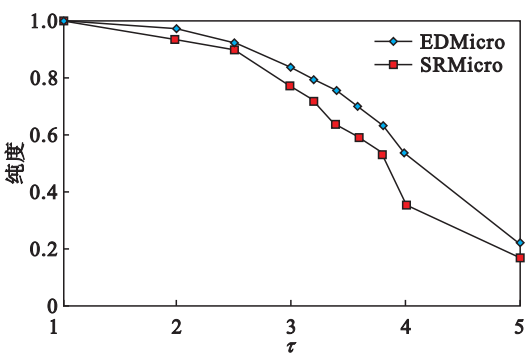

图 4　不同 τ 值对聚类结果的影响

Fig. 4　Influence on cluster purity with different τ

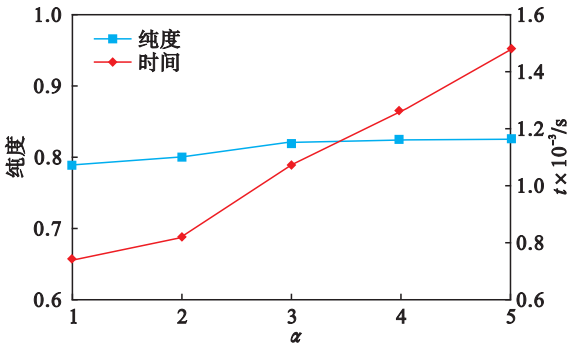

图 5　α 的影响

Fig. 5　Influence of parameter α

2.2.3　可扩展性

图 6 显示了本文提出算法和 SRMicro 算法随维度变化时的聚类处理时间. 可以看出维数增加,EDMicro 算法的聚类时间随之线性增长,并且 EDMicro 算法与 SRMicro 算法的差别随之增大. 图 7 显示出改变微簇数目时,EDMicro 算法和 SRMicro 算法的聚类时间对比.

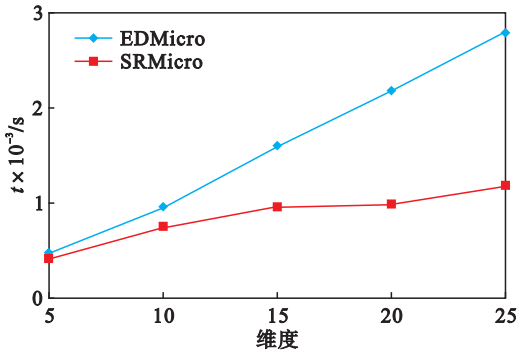

图 6　维度扩展

Fig. 6　Dimension expansion

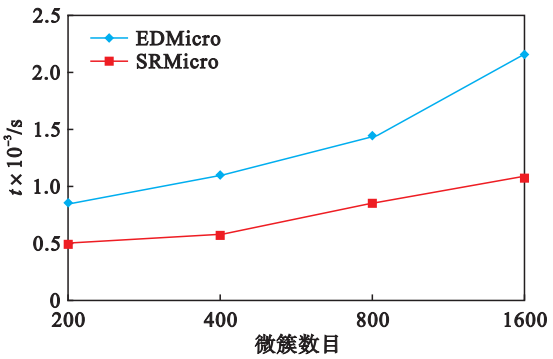

图 7　不同微簇个数对执行时间的影响

Fig. 7　Influence on execution time with different numbers of micro-clusters

3　结　　论

本文提出两种不确定数据流聚类的 EDMicro 算法和 SRMicro 算法. 前一种算法中实例分布特征用不确定元组 MBR 表示,不确定元组和簇心

期望距离的取值范围通过计算期望距离进行推导,过滤距离不确定元组较远的微簇从而减少计算代价. 后者提出簇的 MBR 的概念,从不确定元组 MBR 和簇的 MBR 的空间位置关系,将距离待处理的不确定元组较远的簇过滤. 大量实验验证了本文提出算法的有效性和高效性,并具有良好的可扩展性.

参考文献:

[1] Liu L. From data privacy to location privacy: models and algorithms [C]//Proceeding of the 33rd International Conference on Very Large Data Bases. Vienna, 2007: 1429 – 1430.

[2] Deshpande A, Guestrin C, Madden S, et al. Model-driven data acquisition in sensor networks[C]// Proceeding of the 30th International Conference on Very Large Data Bases. Toronto, 2004: 588 – 599.

[3] Gu Y, Yu G, Zhang T C. RFID complex event processing techniques [J]. *Journal of Frontiers of Computer Science and Technology*, 2007, 1(3): 255 – 267.

[4] Jeffery S R, Garofalakis M N, Franklin M J. Adaptive cleaning for RFID data streams[C]//Proceeding of the 32nd International Conference on Very Large Data Bases. Seoul, 2006: 163 – 174.

[5] 周傲英,金澈清,王国仁,等. 不确定性数据管理技术研究综述[J]. 计算机学报,2009,32(1):1 – 16.
(Zhou Ao-ying, Jin Che-qing, Wang Guo-ren, et al. Summary of research on uncertain data management technology[J]. *Chinese Journal of Computers*, 2009, 32(1): 1 – 16.)

[6] Sarma A D, Benjelloun O, Halevy A Y, et al. Working models for uncertain data [C]//Proceeding of the 22nd International Conference on Data Engineering. Washington DC, 2006: 145 – 157.

[7] Aggarwal C C, Han J W, Wang J, et al. A framework for clustering evolving data streams[C]//Proceeding of the 29th International Conference on Very Large Data Bases. Berlin, 2003: 81 – 92.

[8] Aggarwal C C, Yu P S. A framework for clustering uncertain data streams [C]//Proceeding of the 24th International Conference on Data Engineering. Cancun, 2008: 150 – 159.

[9] Aggarwal C C. On high dimensional projected clustering of uncertain data streams [C]//Proceedings of the 25th International Conference on Data Engineering. Shanghai, 2009: 1152 – 1154.

[10] Zhang C, Gao M, Zhou A Y. Tracking high quality clusters over uncertain data streams [C] //Proceeding of the 25th International Conference on Data Engineering. Shanghai, 2009: 1641 – 1648.

[11] Huang G Y, Liang D P, Ren J D, et al. An algorithm for clustering uncertain data streams over sliding windows[C]// Proceeding of the 6th International Conference on Digital Content, Multimedia Technology and Its Applications. Seoul, 2009: 173 – 177.

[12] Cao K Y, Wang G R, Han D H, et al. A framework for high-quality clustering uncertain data stream over sliding windows [C]// Proceeding of the 13th International Conference on Web-Age Information Management. Harbin, 2012: 308 – 313.

[13] 肖丹萍,叶东毅. 基于免疫原理的不确定数据流聚类算法[J]. 模式识别与人工智能,2012,25(5):826 – 834.
(Xiao Dan-ping, Ye Dong-yi. An algorithm of uncertain data stream cluster based on immune principle [J]. *Pattern Recognition and Artificial Intelligence*, 2012, 25(5): 826 – 834.)

[14] 罗清华,彭宇,彭喜元. 一种多维不确定数据流聚类算法[J]. 仪器仪表学报,2013,34(6):1330 – 1337.
(Luo Qing-hua, Peng Yu, Peng Xi-yuan. Multi-dimensional uncertain data stream clustering algorithm [J]. *Chinese Journal of Scientific Instrument*, 2013, 34(6): 1330 – 1337.)

[15] 胡德敏,余星. 一种不确定数据流子空间聚类算法[J]. 计算机应用研究,2014,31(9):2606 – 2608.
(Hu De-min, Yu Xing. Subspace clustering algorithm for uncertain data stream [J]. *Application Research of Computers*, 2014, 31(9): 2606 – 2608.)

[16] Luo Q H, Yan X Z, Li J B, et al. A dynamic distance estimation using uncertain data stream clustering in mobile wireless sensor networks [J]. *Measurement*, 2014, 55(9): 423 – 433.

(上接第 1676 页)

[2] Qin S J. An overview of subspace identification[J]. *Computer & Chemical Engineering*, 2006, 30(10/11/12): 1502 – 1513.

[3] Gustafsson T. Recursive system identification using instrumental variable subspace tracking[C]// Proceedings of the 11th IFAC Symposium on System Identification. Fukuoka, 1997.

[4] Oku H, Kimura H. Recursive 4SID algorithms using gradient type subspace tracking[J]. *Automatica*, 2002, 38(6): 1035 – 1043.

[5] Houtzager I, van Wingerden J W, Verhaegen M. Recursive predictor-based subspace identification with application to the real-time closed-loop tracking of flutter [J]. *IEEE Transactions on Control Systems Technology*, 2012, 20(4): 934 – 949.

[6] Choi S W, Martin E B, Morris A J, et al. Adaptive multivariate statistical process control for monitoring time-varying processes [J]. *Industrial & Engineering Chemistry Research*, 2006, 45(9): 3108 – 3118.

[7] Ding S X, Zhang P, Naik A, et al. Subspace method aided data-driven design of fault detection and isolation systems [J]. *Journal of Process Control*, 2009, 19(9): 1496 – 1510.

[8] Naik A S, Yin S, Ding S X, et al. Recursive identification algorithms to design fault detection systems [J]. *Journal of Process Control*, 2010, 20(8): 957 – 965.