# 传感器数据中事件样本与错误样本的系统化区分框架

崔筱宁[1,2,3,4,5]，赵保华[1,2,3,4]，李青[4,5]，周颢[1,2,3]

(1.中国科学技术大学计算机科学与技术学院，230027，合肥；2.北京邮电大学网络与交换技术国家重点实验室，100876，北京；3.安徽省计算与通讯软件重点实验室，230027，合肥；4.中科大-香港城大联合研究中心互联网服务实验室，215123，江苏苏州；5.香港城市大学电脑科学系，香港)

**摘要**：针对传感器网络中对事件/异常检测的研究在一定程度上忽略了区分数据样本的重要性问题，依据传感器数据的不确定性分析了事件样本和错误样本的相似点和不同点，设计了系统化区分框架，通过节点级时域处理、邻居级空间处理、聚簇级权重排序和网络级决策融合的方法逐层过滤，将原始样本集划分为正常样本集、错误样本集和事件样本集.真实数据集的实验结果显示，所提框架在不同网络质量下对样本的辨识率均在97%以上，可将误报率降低到传统事件/异常检测方法的1/10，且漏报率不超过传统方法.

**关键词**：传感器数据；事件；错误；系统化区分框架

**中图分类号**：TP393 **文献标志码**：A **文章编号**：0253-987X(2010)10-0030-06

## Systematic Distinction of Events and Errors in Sensor Data

CUI Xiaoning[1,2,3,4,5], ZHAO Baohua[1,2,3,4], LI Qing[4,5], ZHOU Hao[1,2,3]

(1. School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China; 2. State Key Laboratory of Networking and Switching Technology, Beijing, 100876, China; 3. Anhui Province Key Laboratory of Software in Computing and Communication, Hefei 230027, China; 4. Joint Research Lab of Excellence, CityU-USTC Advanced Research Institute, Suzhou, Jiangsu 215123, China; 5. Department of Computer Science, City University of Hong Kong, Hong Kong SAR, China)

**Abstract**: Due to neglecting the importance of distinguishing sensor data in event/anomaly detection, similarities and differences among event samples and error samples are analyzed based on the sensor data uncertainty, and a systematic distinction framework is designed to partition the raw data set into event subset, error subset and ordinary subset through node-level temporal processing, neighbor-level spatial processing, cluster-level ranking and network-level decision fusion. Experimental results on real-sensed data show that the framework achieves a distinction ratio as high as 97% in different network cases. Comparisons with traditional methods show that the proposed framework reduces the false-alarm rate to 1/10 of the traditional methods and does not exceed the traditional miss-hit rate.

**Keywords**: sensor data; event; error; systematic distinction framework

传感器网络是利用兼具物理感知和通信传输功能的传感器节点进行环境监测、事件预警的一种自组织网络[1]，事件检测[2]是其中的一个重要应用.由于传感器节点在运行中会出现不可预测的故障，导

收稿日期：2010-01-11. 作者简介：崔筱宁，(1983—)，女，博士生；赵保华(联系人)，男，教授. 基金项目：国家自然科学基金资助项目(60872009，60602016)；国家"863计划"资助项目(2007AA01Z428，2009AA01Z148)；安徽高校省级自然科学研究计划重大项目(ZD2008005-2，ZD200904，JK2009A013，JK2009A025).

致采集到的数据不一定真实准确，使辨别传感器数据中的事件样本和错误样本的难度大大增加，并直接影响传感器数据的网内处理性能. 根据传感器节点资源受限且采样易于出错等特点，传感器数据处理流程的设计原则主要体现在简化操作、容错性强、分布式处理和对数据样本的高辨识率等方面[3-4]. 目前，用于异常检测的主要技术包括直方图统计[5]、核密度估计[6-7]、权重排序[8-9]、依赖关系分析[10]等. 此外，还有一些针对非正常样本在区域检测方面的扩展研究[11-12]. 尽管如此，以上研究工作或将事件检测和错误检测作为两个独立的问题来研究，或将二者混淆，将错误样本与事件样本统一归为异常样本. 从相关综述文献可以看到，目前与传感器数据处理相关的研究工作缺乏对分辨事件数据和错误数据的足够重视[4,13]，因此，本文着眼于设计一个系统化的数据区分框架来解决这一问题.

## 1　问题分析

本文基于以下基本假设研究事件样本与错误样本的区分问题：①网络采用层次结构，在数据处理时，位于低层的传感器节点将数据就近转发给上层的数据处理中心；②数据处理中心收到的数据不会由于节点通信故障而受损，即只考虑节点采集原始数据过程中产生的错误，不考虑由通信故障造成的数据错误；③传感器网络中不存在由恶意网络攻击造成的虚假数据[14]. 基于上述假设，事件数据与错误数据的区分问题可由定义 1 和定义 2 来描述.

**定义 1**　传感器数据样本 smp＝〈dtype，value，time，location，sID〉，其中 dtype 反映数据样本的物理意义，value 是数值，time 和 location 表明采样时间和地点，sID 是采集该样本的传感器节点编号.

**定义 2**　事件与错误的区分问题(Distinction of Events and Errors，DEE). 在给定区域 S 内的传感器节点在时间 T 内采集到的数据集合记为 $D=\{\text{smp}\}$，DEE 问题是指找出集合 $D$ 的事件数据子集 $D_{\text{event}}$、错误数据子集 $D_{\text{error}}$ 和正常数据子集 $D_{\text{ordinary}}$，其中 $D=D_{\text{event}}\cup D_{\text{error}}\cup D_{\text{ordinary}}$.

根据对传感器错误数据的分类[14]，表 1 给出了错误数据的类型和与之相似的事件数据类型. 由于事件和错误产生原因不同，事件通常不会过于频繁地发生，且事件发生前后的数据模式通常有较大改变，而错误样本在传感器网络中出现频率较高且其出现的时间、地点和数值具有不可预测性，通常表现为随机的数值变化[4].

表 1　易混淆的事件类型和错误类型

| 事件类型 | 错误类型 |
|---|---|
| 偶然事件<br>特征：发生前无任何征兆.<br>示例：突发性心率异常事件. | 离散型错误<br>特征：孤立数据样本值. |
| 瞬时事件<br>特征：突然发生并持续较短时间.<br>示例：车祸事件. | 尖峰型错误<br>特征：短时间内数据变化率超出预期. |
| 持续事件<br>特征：事件持续一定的时间并占据一定的区域.<br>示例：烟雾扩散. | 阻塞型错误<br>特征：在超出预期长度时间段内的数据样本序列出现连续的 0 值或 0 变化率的情况. |

## 2　系统化区分框架

根据事件样本与错误样本的相似点及不同点，本文从数据挖掘的角度提出一个系统化的区分框架来解决 DEE 问题. 该框架的设计基于以下假设. ①传感器网络依簇划分，由一个局部处理中心负责处理簇内的传感器数据，各局部处理中心都可(经由一定路径)与远程基站联络. ②一个事件可在至少一个簇内被发现. 至少 $m$($m$ 为整数且 $m>k/2$)个传感器节点可以几乎同时发现同一事件，其中 $k$ 是网络中一个传感器节点的平均单跳邻居数量. ③$k$ 个邻居节点在同一时间单元内产生同类型错误样本的概率非常小，设为 $\varepsilon$，其中 $\varepsilon$ 为趋近于 0 的小正数.

判断异常数据样本的方法有直方图频率比较法[5]、核密度估计法[6-7]或权重排序法[8-9]等. 这些方法或依赖网络中的全局数据样本信息进行异常判别，或没有系统化地考虑网络结构. 本文试图整合上述方法中的优势，同时规避其不足. 基于前述假设，系统化区分框架的流程图如图 1 所示.

该框架包含 4 个处理步骤，分别代表 4 个级别的数据处理粒度：节点级、节点-邻居级、聚簇级和网络级. 数据处理的重点是利用异常样本出现频率和数据模式变化规律来区分其中的事件样本和错误样本. 考虑到传感器节点的计算能力[15]和能量限制[16]，以及被监测事件要求的响应时间，不同级别采用不同的数据处理方法，见图 1.

(1)节点级　在传感器节点内部主要对节点缓存的历史数据进行时域处理，若某样本值或其变化率超出物理阈值(例如，温度值低于热力学零度)则显示为错误样本，归入 $D_{\text{error}}$ 集合. 此外，时域处理还

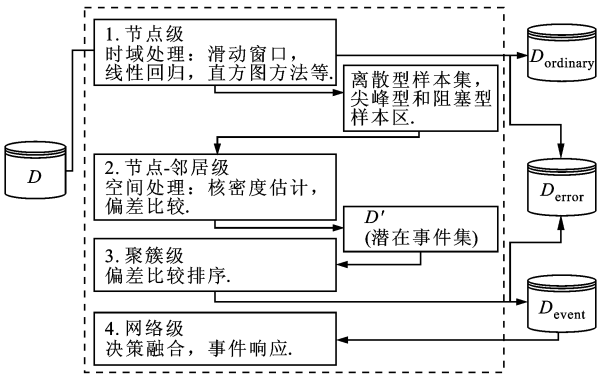


图 1 系统化区分框架基本流程图

标记出节点缓存数据序列中的离散型错误样本集合，以及尖峰型和阻塞型错误样本的区间，用于后续处理.

本文采用基于固定宽度滑动窗口的线性回归方法对节点内部数据进行时域处理. 根据传感器节点的缓存大小 $l$ 设置滑动窗口的宽度 $w(w\leqslant l)$，根据采样频率 $s$ 设置滑动窗口的滑动速度 $v$，保证在单位时间 $t$ 内传感器节点上的数据更新量不超过滑动窗口处理的数据量，即 $st\leqslant wv$. 按照表 2 中采用的各种区分度量标准来计算和比较预测样本值与对应真实采样值的差别，最终反映样本序列在时域上的数据模式. 由于事件和错误都会引起数据模式的显著变化，从而导致预测值与实测值之间存在较大的误差，据此可将相关样本标记出来进行后续处理，其余样本则归入集合 $D_{ordinary}$.

(2)节点-邻居级 在传感器节点及其单跳邻居之间进行空间处理，比较它们标记出的离散型错误样本集以及尖峰型和阻塞型错误区间是否一致.

传感器网络是采样和通讯链路均易出错的不可靠网络，且有一定的能量制约. 一方面，邻近节点的空间相关性较高，通过节点与其单跳邻居的协作，可提高采样准确性；另一方面，单跳通讯比多跳通讯的链路代价更低，为确保通讯链路的可靠性和节省能量，采用节点与单跳邻居交换传感器数据.

基于本章假设，若有足够比例(实验中将此比例设为大于或等于 50%)的邻居节点在同一时间单元内报告了同样类型的数据异常，则该异常很可能说明网络中发生了被监测事件. 这些异常样本被归入潜在事件集合 $D'$，其他异常样本则归入 $D_{error}$ 集合.

(3)聚簇级 聚簇级的处理工作通常由具有稳定能量供应的节点来完成，本文称之为局部处理中心，这些节点不具备环境数据采集的功能，只具备信息通信和计算的功能.

局部处理中心参照 $D_{ordinary}$ 集合中的正常数据来评估集合 $D'$ 中的样本，最终从 $D'$ 中选出事件样本并构成集合 $D_{event}$. 由于假设簇内任一传感器节点与其单跳邻居节点同时产生类似错误样本的可能性很小，本文采用基于偏差比较排序的策略来评估 $D'$ 中的样本. 在获得 $D_{event}$ 之后，集合 $D'\backslash D_{event}$ 中的样本并入集合 $D_{error}$.

聚簇性能的好坏取决于聚簇规模的大小和数据样本的分布情况，需要根据具体应用场景在实验中不断摸索来设定，本文设定的聚簇规模为 10 个邻近节点，即大约每 10 个邻近节点的数据交由一个局部处理中心来处理. 对簇内各节点数据样本按采样时间计算样本值的偏差并排序，以偏差较小的数据样本值作为网络级决策的依据.

(4)网络级 基站获得 $D_{event}$ 集合并进行决策融合[17]，即根据事件信息采取相应的措施. 此级别不涉及事件样本与错误样本的区分，故不在本文展开讨论.

## 3 性能评估

本文引入由 Intel Berkeley Research Lab 提供的传感器实测数据集[18]，向原始数据注入一定比例的错误样本和事件样本进行性能评估. 实验的 3 个步骤分别对应于系统化区分框架中节点级、节点-邻居级和聚簇级数据处理. 不失一般性，系统化区分框架不对节点采样时的错误概率和事件样本值的变化规律进行限定.

实验首先测试了表 2 中哪种区分度量标准能最为准确地反映实际样本数值的变化情况，并将该度量标准用于步骤 1 中的样本值及变化率预测和步骤 2、3 中的样本区分. 实验输入数据集取自 4 个邻近节点(文献[18]中编号为 1、2、3 和 33 的节点)，平均每个节点 2 000 个温度数据样本，由于存在漏采样或过采样的情况，综合 4 个节点的采样情况，整合后共有 2 787 个原始数据样本. 各种度量值均以滑动窗口为单位进行计算.

图 2 比较了这些度量值的精确度. 与其他度量标准相比，无论在微弱变化还是剧烈变化的样本区间，相关系数都能够最为精确地反映原始样本值的变化情况，例如，在[2 989,3 633]和[4 740,5 566]这两个时间区间内，原始数据变化较小，相关系数也几乎无波动；在[3 689,3 904]时间区间内，原始样本值急速增大，相关系数值也相应大幅下降(呈反向

关系）. 其他度量标准或多或少表现出与原始数据变化不一致的情形，例如，在时间区间[5 063，5 154]

表 2　实验参数配置

| 步骤和标准 | 参数 |
|---|---|
| 全局配置 | 数据量：每个节点 200 个温度样本（含注入的错误和事件样本）；<br>错误出现频率≤50%；<br>事件发生频率≤1%；<br>错误/事件样本的取值范围：[−30，100]. |
| 步骤 1（节点级） | 滑动窗口宽度：40 个样本；<br>滑动速率：50%. |
| 步骤 2（节点-邻居级） | 平均每个节点有 3 个单跳邻居. |
| 步骤 3（聚簇级） | 平均聚簇规模：10 个邻近节点. |
| 区分度量标准 | 相关系数、平均绝对误差、绝对均方根误差、相对误差、相对方根误差. |
| 性能评估标准 | $r_1$：把错误样本当作正常样本的比例；<br>$r_2$：把错误样本当作事件样本的比例；<br>$r_3$：把正常样本当作事件样本的比例；<br>$r_4$：把正常样本当作错误样本的比例；<br>$r_5$：把事件样本当作正常样本的比例；<br>$r_6$：把事件样本当作错误样本的比例；<br>$r_7$：正确区分事件、错误和正常样本的比例；<br>满足 $r_1+r_2+r_3+r_4+r_5+r_6+r_7=1$.<br>误报率=$r_2+r_3$；漏报率=$r_5+r_6$. |

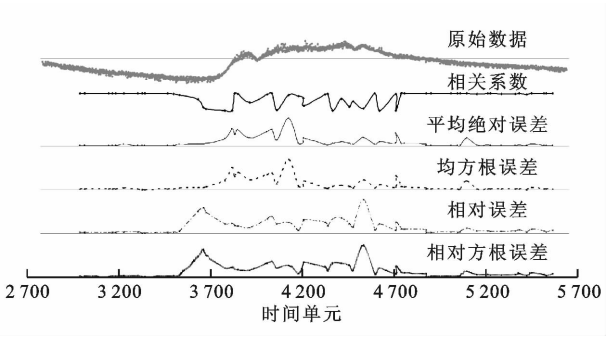


图 2　不同度量标准下的预测精确度对比

范围内，原始样本值基本无变化，除相关系数保持稳定之外，其他 4 种度量值都表现出一定程度的波动. 因此，实验采用相关系数作为差别度量标准.

表 2 中的错误出现频率是指错误样本数量占样本总数量的比例，它反映了网络中原始数据质量的好坏；事件发生频率是指事件样本数量占样本总数量的比例. 实验中将错误出现频率设置为 10%～50%，将事件发生频率设置为不超过 1%，这与现实中错误出现频率远高于事件发生频率的事实一致. 表 2 中 $r_7$ 即辨识率，反映了经过每个处理步骤正确分辨出的样本数量在总样本数量中所占的比例.

表 3 给出了实验统计结果. 随着错误出现频率的增加，$r_7$ 在步骤 1 和步骤 2 都有所下降，但步骤 3 纠正了大部分误判结果，且在错误出现频率 10%～50%范围内 $r_7$ 都保持在 97%以上，体现出系统化区

表 3　各种错误出现频率和各步骤中的处理性能比较　　%

| 错误率 | 步骤 | $r_7$ | $r_6$ | $r_5$ | $r_4$ | $r_3$ | $r_2$ | $r_1$ |
|---|---|---|---|---|---|---|---|---|
| 10 | 1 | 91.39 | 0.00 | 0.00 | 0.00 | 0.00 | 6.11 | 2.50 |
| | 2 | 96.94 | 0.28 | 0.00 | 0.00 | 0.00 | 0.28 | 2.50 |
| | 3 | 100.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 20 | 1 | 80.83 | 0.00 | 0.28 | 0.00 | 1.39 | 7.78 | 9.72 |
| | 2 | 88.33 | 0.28 | 0.28 | 1.39 | 0.00 | 0.00 | 9.72 |
| | 3 | 99.44 | 0.00 | 0.56 | 0.00 | 0.00 | 0.00 | 0.00 |
| 30 | 1 | 73.06 | 0.00 | 0.83 | 0.00 | 1.94 | 6.67 | 17.50 |
| | 2 | 79.44 | 0.28 | 0.83 | 1.94 | 0.00 | 0.00 | 17.50 |
| | 3 | 99.44 | 0.00 | 0.56 | 0.00 | 0.00 | 0.00 | 0.00 |
| 40 | 1 | 67.22 | 0.00 | 1.11 | 0.00 | 0.56 | 9.72 | 21.39 |
| | 2 | 75.28 | 0.56 | 1.11 | 0.56 | 0.00 | 1.11 | 21.39 |
| | 3 | 97.22 | 0.00 | 1.67 | 0.00 | 1.11 | 0.00 | 0.00 |
| 50 | 1 | 56.39 | 0.00 | 0.56 | 0.00 | 1.94 | 10.56 | 30.56 |
| | 2 | 65.83 | 0.00 | 0.56 | 1.39 | 0.56 | 1.11 | 30.56 |
| | 3 | 97.78 | 0.00 | 0.56 | 0.00 | 1.67 | 0.00 | 0.00 |

分方法具有较好的容错性. 同时,表 3 还显示出,在不同网络质量情况下,系统化区分框架中每一步的处理都使 $r_7$ 有所增加,这也体现出该框架性能的稳定性.

表 2 中还给出了事件误报率和漏报率的计算方法. 传统的事件检测方法常将错误混淆为事件,导致误报率较高,而传统的异常检测方法常将事件误认为错误,导致漏报率较高. 图 3 和图 4 进一步给出本文提出的针对 DEE 问题的系统化区分框架与传统的事件检测和异常检测方案在误报率和漏报率方面的性能比较,可见在各种不同的网络质量情况下,本文设计的系统化区分框架与传统的事件检测相比可显著降低误报率,且漏报率也明显低于传统异常检测的漏报率.

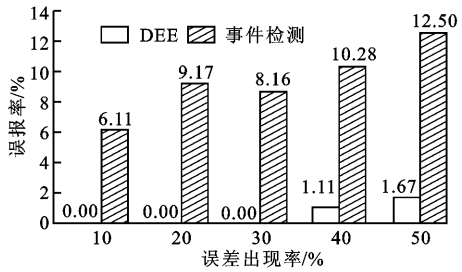


图 3 DEE 与事件检测的误报率比较

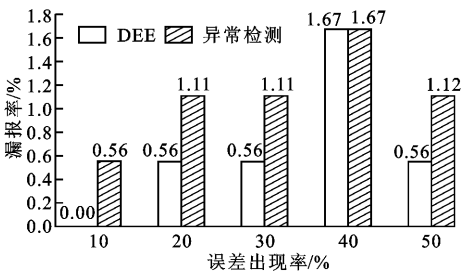


图 4 DEE 与异常检测的漏报率比较

## 4 结 论

不同于传统的事件检测和异常检测问题,本文提出了区分传感器数据中的事件样本和错误样本这一问题,并根据节点自身特点设计了一个多层次的系统化处理结构来过滤传感器数据中的错误样本和提取出事件样本. 在真实数据集上的实验验证结果表明:相关系数是判别预测样本值和实测样本值之间差别时较为合适的度量标准,且通过网络中的系统化数据处理,数据样本的误报率和漏报率与传统的事件/异常检测相比有较大幅度的降低.

**参考文献:**

[1] 孙利民, 李建中, 陈渝, 等. 无线传感器网络[M]. 北京: 清华大学出版社, 2005.

[2] SHEPHERD D, KUMAR S. Microsensor applications [M]// SUNDARARAJA S L, RICHARD R B. Distributed sensor networks. London, UK: Chapman & Hall/CRC Press, 2005:11-27.

[3] BUONADONNA P, GAY D, HELLERSTEIN J M, et al. TASK: sensor network in a box, IRB-TR-04-021[R]. Inter Research Berkeley, USA: IEEE, 2005: 133-144.

[4] ZHANG Yang, MERATNIA N, HAVINGA P. Why general outlier detection techniques do not suffice for wireless sensor networks[C]// Intelligent Techniques for Warehousing and Mining Sensor Network Data. Hershey, Pennsylvania, USA: IGI Global, 2009:136-158.

[5] SHENG Bo, LI Qun, MAO Weizhen, et al. Outlier detection in sensor networks[C]// Proceedings of the ACM International Symposium on Mobile Ad Hoc Networking and Computing. New York, USA: ACM, 2007:219-228.

[6] PALPANAS T, PAPADOPOULOS D, KALOGERAKI V, et al. Distributed deviation detection in sensor networks [J]. SIGMOD Record, 2003, 32(4):77-82.

[7] SUBRAMANIAM S, PALPANAS T, PAPADOPOULOS D, et al. Online outlier detection in sensor data using non-parametric models[C]// Proceedings of 32nd International Confere3nce on Very Large Data Bases. Seoul, Korea: VLDB Endowment, 2006:187-198.

[8] BRANCH J W, SZYMANSKI B, GIANNELLA C, et al. In-network outlier detection in wireless sensor networks[C]// Proceedings of 26th International Conference on Distributed Computing Systems. Piscataway, NJ, USA: 2006:51-51.

[9] OTEY M E, GHOTING A, PARTHASARATHY S. Fast distributed outlier detection in mixed-attribute data sets[J]. International Journal of Data Mining and Knowledge Discovery, 2006, 12 (2/3):203-228.

[10] JANAKIRAM D, REDDY V, KUMAR A. Outlier detection in wireless sensor networks using Bayesian belief networks[C]// 1st International Conference on Communication System Software and Middleware. Piscataway, NJ, USA: IEEE, 2006:1-6.

[11] FRANKE C, GERTZ M. ORDEN: outlier region detection and exploration in sensor networks[C]// Proceedings of ACM Special Interest Group on Manage-

ment of Data. New York, USA: ACM, 2009: 1075-1078.

[12] DING Min, CHEN Dechang, XING Kai, et al. Localized fault-tolerant event boundary detection in sensor networks [C]// Proceedings of 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Piscataway, NJ, USA: IEEE, 2005, 2: 902-913.

[13] CHANDOLA V, BANERJEE A, KUMAR V. Anomaly detection: a survey[J]. ACM Computing Surveys, 2009, 41(3): 1-72.

[14] NI K, RAMANATHAN N, CHEHADE M N H, et al. Sensor network data fault types[J/OL]. [2010-01-01]. http: // 166.111.120.94/acm/Content Loader.nsp? view = % 2F1530000% 2F1525863% 2Fa25 - ni.pdf.

[15] 陈磊，赵保华. 一种高信道利用率的分簇传感器网络数据聚合协议[J]. 西安交通大学学报，2009，43(2)：38-42.
CHEN Lei, ZHAO Baohua. A highly-channel-utilized data aggregation protocol for clustered sensor networks [J]. Journal of Xi′an Jiaotong University, 2009, 43(2): 38-42.

[16] 赵保华，李培龙，高存皓，等. 延长实时无线传感器网络生命周期的能量平衡路由[J]. 西安交通大学学报，2007，41(10)：1137-1140.
ZHAO Baohua, LI Peilong, GAO Cunhao, et al. Prolonging lifecycle of energy balance routing in real-time wireless sensor networks[J]. Journal of Xi′an Jiaotong University, 2007, 41(10): 1137-1140.

[17] KOKAR M M, WEYMAN J, TOMASIK J A. Data vs. decision fusion in the category theory framework [C]// 4th International Conference on Information Fusion, [S. l.]: Taylor & Francis, 2001: 3-15.

[18] Madden S. Intel lab data [EB/OL]. (2004-06-02) [2010-01-01]. http: // db. csail. mit. edu/labdata/labdata. html.

（编辑　武红江）