DOI:10.11918/j.issn.0367-6234.201705095

复杂事件处理的自适应制造情景识别方法

任 磊,任明仑

(教育部过程优化与智能决策重点实验室(合肥工业大学), 合肥 230009)

摘 要: 制造过程中的任务、自然条件、电力水平等环境因素,制约物体状态及其关系的变化. 智能制造单元需要自适应的对不同情境约束下的事件和复杂情形及时理解判断,提出基于复杂事件处理(Complex event processing, CEP)的情境约束情景识别方法,以实时作出合理的优化决策. 针对忽视情境约束对事件判别的影响,构建基于情境约束的多层次事件模型,给出同生、情境、协同等事件新算子,提出基于事件聚合的制造情景模型与演算过程. 针对情景识别知识库中模式规则生成的不足,通过整合物体数据与环境数据建立映射关联,将感知信息转化为情境事件图谱. 通过综合序数、名义变量等距离计算和自适应熵权法,提出改进的混合聚类方法处理事件图谱实例属性的多样性和关联性,构建知识库以为情景实时识别提供服务支持. 运用 4 个真实数据集和 1 个制造过程仿真数据集进行实验,均验证本文模型和方法的有效性,适用于大规模学习问题,并阐明情境因素能显著提升复杂制造应用中的事件判断、情景识别的准确性.

关键词: 情境约束,数据流,复杂事件,情景识别,混合聚类方法

中图分类号: TP393 **文献标志码**: A **文章编号**: 0367-6234(2017)11-0171-07

Adaptive context based situation identification based on complex event processing

REN Lei, REN Minglun

(Key Laboratory of Process Optimization and Intelligent Decision Making of Ministry of Education, Hefei University of Technology, Hefei 230009, Anhui, China)

Abstract: In the manufacturing process, environmental factors such as tasks, natural conditions, power levels, etc., restrict the state change of objects and their relationships. Intelligent manufacturing units need to adaptively understand and judge events and complex situations under different context constraints, and a adaptive situation identification method based on complex event processing is proposed naturally, to make real-time optimization decision reasonable. In view of the phenomenon that the influence of context constraints on event discrimination is neglected, a context-aware hierarchical event model is built, and new operators of events such as contemporaneity, context and collaboration are given, while manufacturing situation model and aggregation process are proposed. Aiming at the shortage of generating method on situation model in knowledge base, the mapping association between an object and environment data is established firstly, and such sensed information is transformed into context based events. Integrating the distance calculation of ordinal, nominal variable and adaptive entropy weight method, a improved mixed clustering method is put forward to deal with the diversity and relevance of complex event instance attributes, providing service support for real-time situation identification. Finally, 4 real data sets and 1 simulation data set of are employed for manufacturing process. Experiment results have verified the validity and adaption of the proposed model and method in large-scale problem, and expounded that context factors can significantly improve the accuracy of event judgment and situation recognition in complex manufacturing applications.

Keywords: context constraints; data stream; complex event; situation identification; mixed clustering method

物联网、云计算、大数据等新兴信息技术的应用,使得制造单元具有高度智慧性[1],能通过感知、识别当前发生事件而采取最优决策.由于环境的动态性和制造过程的复杂性,需实时关注设备、人员、环境等状态,产生海量的多模态时变数据流. 每一条数据流都反映制造过程某一方面的状态变化,共同刻画了所面临的制造情景.如何快速处理这些大规模数据以理解当前发生情形,实时地把各种不同来源的信息转化为决策利用的知识成为一项基本挑战.CEP 作为一种新兴流数据处理技术[2-3],能够抽取有意义的事件,通过顺序、聚合、因果等事件算子发现事件关系[4],从而识别更高决策价值的复杂事件模式.

很多学者运用 CEP 技术对制造系统底层的流数据处理[5]、事件发现[6]到高层的情景挖掘[7]与实

收稿日期: 2017-05-17
基金项目: 国家自然科学基金(71531008);国家自然科学基金面上项目(71271073)
作者简介: 任 磊(1988—),男,博士研究生;
任明仑(1969—),男,教授,博士生导师
通信作者: 任明仑, renml@hfut.edu.cn

时决策[8]进行广泛研究.但假定周围环境固定不变,忽视任务、时空等情境对物体状态的影响,如同样的设备温度 20 度在 $Task_1$ 时为正常事件,而在 $Task_2$ 时则为异常事件.情境制约着主体对事物的认知,事件更全面的理解、自适应的决策均需要情境支持[9].虽有学者对制造领域情境分类建模,但仍将其作为事物的一种状态属性,并未与其他属性区分,更未分析其对事件判别的制约.现有事件模式生成主要依赖于人工定义[10],难以获取全面知识且容易出现错误和组合爆炸.机器学习方法也只应用于简单场景,涉及的实体、数据维度较少,难以解决制造过程大数据的新问题[14].基于现有 CEP 系统不足[11],将多层次情境考虑到事件理解过程中,提出基于情境约束的自适应情景识别方法,通过改进混合聚类进行模式挖掘和在线实时识别,从而优化任务执行决策活动.

1 基于情境约束的多层次事件模型

多层次情境制约物体状态变化和事件判别,给出情境约束的事件模型和基于事件聚合的情景演算过程.

1.1 情境数据流

决策单元所关注物体某一属性状态随时间不断变化,形成一条不同时刻下的动态数据序列即数据流 DS:

$$DS = \{v(a_1), v(a_2), \cdots, v(a_n)\}. \tag{1}$$

式中:a 为物体属性,$v(a_i)$ 为第 i 时刻属性值.如图 1 中设备转速曲线即是一条 t_0 到 t_l 时刻的动态数据流.

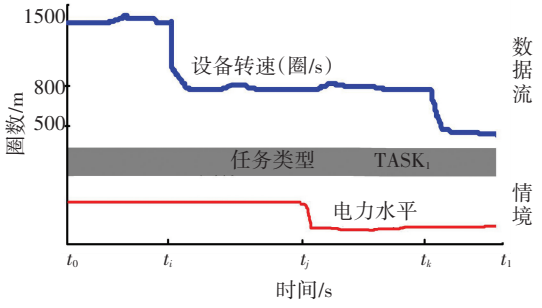


图 1 情境数据流

Fig.1 Context based data stream

智能单元关注事物处于动态环境中,其属性值在不同环境下具有不同分布特征、状态变化模式.物体属性状态随着情境变化而形成时间序列即情境数据流 CSS:

$$CSS = \{\langle v(a_1), ct_1\rangle, \langle v(a_2), ct_2\rangle, \cdots, \langle v(a_n), ct_n\rangle\}. \tag{2}$$

式中:$v(a_j)$ 为第 j 时刻情境约束 ct_j 该物体属性 a 的数值,ct_j 为第 j 时刻情境值.如图 1 中,t_0 到 t_j 的情境为 $(TASK_1, EL_1)$,t_j 到 t_l 的情境为 $(TASK_1, EL_2)$,转速曲线是一条在两类情境下的数据流.两类情境的数据特征存在差异,$(TASK_1, EL_1)$ 下设备转速水平明显高于 $(TASK_1, EL_2)$,原因是电力水平的高低影响运转速度.

具有相同情境约束的不同时刻物体状态数据构成同类情境数据流段 CSE,描述物体状态在同一情境下的数值分布:

$$CSE = \{\langle v(a_1), ct_k\rangle, \langle v(a_2), ct_k\rangle, \cdots, \langle v(a_m), ct_k\rangle\}. \tag{3}$$

式中:$v(a_i)$ 为第 i 时刻情境约束 ct_k 下属性 a 的数值,ct_k 为同类情境约束.图 1 中 t_0 到 t_j 的数据流为同类情境数据流段 CSE_1,t_j 到 t_l 是同类情境数据流段 CSE_2.

情境是描述物体相关行为和状态变化时所处的环境信息,可表示为 CT:

$$CT = (V, T). \tag{4}$$

式中:V 为情境数值,T 为时间集.制造情境包含任务、自然环境、时空信息等.情境具有时变性,不同时刻形成情境序列流 $CTS = (ct_1, ct_2, \cdots, ct_n)$,$ct_i$ 为第 i 时刻情境值.情境具有层次性,高层抽象情境是底层情境的泛化表示,情境层次结构通过有向无环图描述 $H = (CT_H, W_H)$.如图 2,CT_H 为某一层次情境类,$W_H \subseteq CT_H \times CT_H$ 为情境节点层次关系,节点间偏序关系 $\prec_H$ 刻画情境类之间的具体化关系、泛化关系.

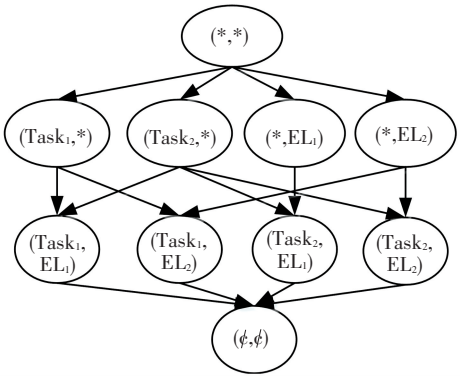


图 2 情境层次结构

Fig.2 Context hierarchy

给定情境节点 CT_i,$CT_i' \subseteq H$,存在 $\forall X \in CSS, CT_i(X) \subset CT_i'(X)$,则 $CT_i \prec_H CT_i'$ 即从 CT_i' 到 CT_i 存在有向边.CT_i' 比 CT_i 更泛化,而 CT_i 比 CT_i' 更具体,如情境 $(Task_1, *)$ 比 $(Task_1, EL_1)$、$(Task_1, EL_2)$ 更抽象.如果没有 $CT_i \neq (\varphi, \cdots, \varphi)$ 且 $CT_i \prec_H CT_i'$,则 CT_i' 为一个具体化类.某情境节点在层次结构中没有更具体化情境时即 $CT_k \subseteq H$,$\neg \exists\ CT_k' \in H$ 满足 $CT_k' \prec_H CT_k$,则 CT_k 为最具体化情境 $\overline{CT}_H$,本节应用最具体化情境.

1.2　基于情境约束的事件和情景模型

事件是系统关注物体有决策意义的行为和状态变化，如温度异常、转速波动等.简单事件是系统中最小的不可分割的物体状态的瞬时事件 e：

$$e=(\mathrm{ID},\mathrm{type},\mathrm{time},v(a)). \tag{5}$$

式中：ID 为标识符，type 为事件类型，time 为发生时间，$v(a)$ 为属性值.如设备温度在 t_0 时刻 100 度为温度异常简单事件 $e_1=(123,温度,t_0,100>50)$.

1.2.1　情境事件

由于情境制约着事件的理解，现有事件模型难以适应动态环境的决策需求.通过添加情境因素，拓展构造情境事件 cte：

$$\mathrm{cte}=(\mathrm{ID},\mathrm{type},\mathrm{time},v(a),\mathrm{ct}). \tag{6}$$

式中：ct 为情境约束，同一状态值在不同情境下具有差别的意义.如设备温度 t_0 时 100 度，为 TASK_1 下正常事件，TASK_1 要求温度(50,150).而 t_2 时 100 度，为 TASK_2 下异常事件，TASK_2 要求温度(50,80).

情境数据流中每个数值对应一个简单情境事件，按照时间顺序形成情境事件流 CTE：

$$\mathrm{CTE}=(\mathrm{type},\mathrm{state}(\mathrm{cte}),\mathrm{ct}). \tag{7}$$

式中：type 为事件类型，state(cte) 为所有事件集，ct 为情境约束集.图 1 中运转曲线对应一条情境事件流，同样的设备转速在两个情境下分别为异常和正常事件.

1.2.2　基于事件聚合的情景模型

情景是系统所关注的特定环境下发生的多种关联事件集合，即通过简单事件聚合的复杂事件模式.如 TASK_1 中 t_0 时刻设备出现温度、速度异常事件，则聚合为故障情景.后文中情景与复杂事件是相同概念 SA：

$$\mathrm{SA}=(\mathrm{type},\mathrm{state}(\mathrm{CTE}),F). \tag{8}$$

式中：type 为情景类型，state(CTE) 为多种情境事件集，F 为情境事件算子集.制造情景是制造过程中某一时刻所有物体状态变化，是多种事件的同时发生而聚合成的复杂事件，只需同生算子而不考虑事件间的顺序、选择、因果关系等.

同生算子 $\Uparrow$：给定事件 E_1 和 E_2，如两者同时发生，则表示为 $E_1\Uparrow E_2$. 如设备故障情景，是通过温度、转速异常事件的同生运算得到.由于同生算子并不能保障两个事件的情境约束相同，将两个没有关联的事件聚合到一起并没有决策意义，出现风发牛不相及现象.

针对情境对事件聚合的约束，提出情境算子 $\Downarrow$：给定事件 E_1 和 E_2，如两者发生在相同情境约束下，表示为 $E_1\Downarrow E_2$. 两事件同时满足同生和情境算子得到协同算子 $\Updownarrow$：给定事件 E_1 和 E_2，如两者同时发生在相同的情境约束下，则为 $E_1\Updownarrow E_2$.

多条事件流协同聚合形成一条新的复杂事件图谱流即情景流 $\mathrm{SAS}=(\mathrm{sa}_1,\mathrm{sa}_2,\cdots,\mathrm{sa}_n)$，整个演算过程如图 3 所示.给定两个简单事件流 E_1 和 E_2，分别表示机器温度和震动事件，需要对持续到达的两个子事件按照业务规则进行组合，通过同生算子、情境算子得到复合事件图谱，最终运用函数 Z 进行情景判别，确认当前情景模式并给出相关描述.

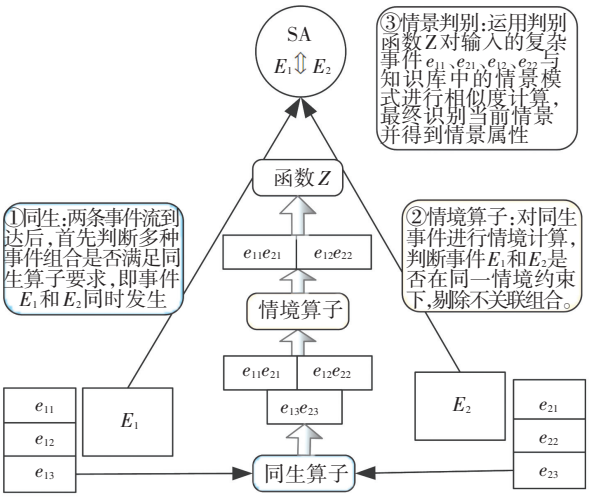


图 3　事件聚合过程

Fig.3　Event polymerization process

2　基于复杂事件处理的自适应情景识别方法

2.1　基于复杂事件处理的情景识别过程

制造情景识别包括事件与情境抽取、事件图谱构建、模式匹配和情景确认四个阶段，如图 4.根据领域知识对环境数据分类，构建多个情境类，假定情境类已经获得.分析不同情境类下物体状态分布规律，界定事件类型并构建复杂事件图谱.通过线下的混合聚类对大量事件实例学习，形成情景知识库.最终将实时感知数据与知识库中事件模式匹配识别当前情景，具体步骤如下：

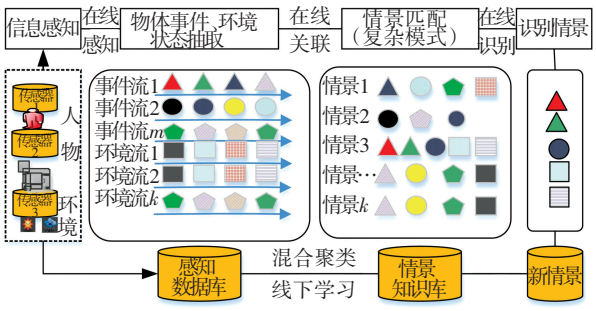


图 4　基于复杂事件处理的情景识别过程

Fig.4　Situation identification based on complex event processing

1)事件、情境抽取:参照相关规则模型,对感知到的人、物、环境数据进行简单事件和情境判别;

2)复杂事件图谱构建:根据 2.2 节事件聚合算子建立事件与情境关联,聚合得到事件图谱实例;

3)情景匹配:计算当前事件实例与知识库中情景模式的相似度,基于贴近原则选择满足阀值的情景;

4)情景确认:分析匹配情景特征,实施应对方案.如未满足阀值,则将当前实例作为新情景存贮到知识库,更新情景知识.

2.2 事件图谱的构建

根据领域知识将感知到的多维数据转化为简单事件,形成事件流和事件矩阵,通过聚合算子构建复杂事件图谱以界定所面临的情形.

2.2.1 数据矩阵

给定某一制造加工场景,感知的物体状态检测变量数为 m, 情境变量为 k,n 个时刻获取海量数据为 $\boldsymbol{OB}$ 和 $\boldsymbol{CT}$ 两个时空数据矩阵:

$$\boldsymbol{OB}=\begin{vmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{vmatrix},$$

$$\boldsymbol{CT}=\begin{vmatrix} ct_{11} & ct_{11} & \cdots & ct_{1k} \\ ct_{21} & ct_{22} & \cdots & ct_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ ct_{n1} & ct_{n2} & \cdots & ct_{nk} \end{vmatrix}. \tag{9}$$

式中:$\boldsymbol{OB}$ 中第 i 行为第 i 时刻物体状态各个变量的值,第 j 列为第 j 个变量在不同时刻的值. $\boldsymbol{CT}$ 中第 i 行为第 i 时刻各个情境因素的值,第 j 列为第 j 个情境因素在不同时刻下的值.

情境约束类通过专家知识界定,共获得 p 个情境类为 $\mathrm{CT}=(\mathrm{CT}_1,\mathrm{CT}_2,\cdots,\mathrm{CT}_p)$. 通过建立“物-环境”的数据关联,原有两个数据矩阵合并为一个情境数据矩阵 $\boldsymbol{CTX}$:

$$\boldsymbol{CTX}=\begin{vmatrix} x_{11} & x_{12} & \cdots & x_{1m} & \mathrm{CT}_1 \\ x_{21} & x_{22} & \cdots & x_{2m} & \mathrm{CT}_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} & \mathrm{CT}_p \end{vmatrix}. \tag{10}$$

$\boldsymbol{CTX}$ 中第 i 行为第 i 时刻在情境约束 CT_i 下物体状态各个变量值, x_{ij} 为第 i 时刻在情境约束 CT_i 下物体第 j 个变量值.

2.2.2 事件图谱

$\boldsymbol{CTX}$ 中某一变量具有连续值域,这些值间虽然差距较大,但在特定情境约束下具有相同意义,如设备正常转速分布在(800,1 500). 基于制造知识和统计分析,将变量的波动区间映射为不同事件类型如正常、异常事件等[12].假定第 i 个变量在情境约束 CT_k 下正常、异常波动分别为 $(x_i^{\max},x_i^{\min})_{\mathrm{nor}}^{\mathrm{CT}_k}$、$(x_i^{\max},x_i^{\min})_{\mathrm{abn}}^{\mathrm{CT}_k}$, 则事件的映射关系:

$$x_i=\begin{cases} E_{\mathrm{nor}}, & \text{if} \quad [x_i\in(x_i^{\max},x_i^{\min})_{\mathrm{nor}}^{\mathrm{CT_k}}]; \\ E_{\mathrm{abn}}, & \text{if} \quad [x_i\in(x_i^{\max},x_i^{\min})_{\mathrm{abn}}^{\mathrm{CT_k}}]. \end{cases} \tag{11}$$

将 CTX 中的每一个数值均映射为情境约束事件,则得到复杂事件图谱 $\boldsymbol{TEQ}$:

$$\boldsymbol{TEQ}=\begin{vmatrix} e_{11} & e_{12} & \cdots & e_{1m}\ CT_1 \\ e_{21} & e_{22} & \cdots & e_{2m}\ CT_2 \\ \vdots & \vdots & \vdots & \vdots \\ e_{n1} & e_{n2} & \cdots & e_{nm}\ CT_p \end{vmatrix}. \tag{12}$$

$\boldsymbol{TEQ}$ 中第 i 行表示第 i 时刻在情境约束 CT_i 下出现的事件集合,反映 CT_i 下所关注的情景状态, x_{ij} 为第 i 时刻在情境约束 CT_i 下物体第 j 个变量的事件类型.

根据同生、情境和协同聚合算子,进一步剔除 TEQ 进中不符合业务逻辑和运算规则的事件,从而得到最终的情境约束复杂事件图谱 $CTEQ$:

$$\boldsymbol{CTEQ}=\begin{vmatrix} e_{11}^{'} & e_{12}^{'} & \cdots & e_{1m}^{'} \\ e_{21}^{'} & e_{22}^{'} & \cdots & e_{2m}^{'} \\ \vdots & \vdots & \cdots & \vdots \\ e_{n1}^{'} & e_{n2}^{'} & \cdots & e_{nm}^{'} \end{vmatrix}. \tag{13}$$

2.3 改进混合聚类方法

$\boldsymbol{CTEQ}$ 是一个由多条事件流聚合构成的事件图谱,出现现有混合聚类没有的数值类型,且变量在不同模式中具有差异权重,以往方法只对数值、名义两类变量的权重基于主观经验分配[13].本文通过统一的距离公式,运用自适应熵权改进混合聚类方法.给定 n 个情景对象集 $X=(X_1,X_2,\cdots,X_n)$, 每个对象是一组事件集合,混合聚类的目标就是将 X 划分为 k 簇,每个簇为一个情景模式,以最小化簇间距离为目标函数:

$$Z_{\mathrm{OF}}=\sum_{l=1}^{k}\sum_{i=1}^{n}u_{il}d(x_i,Q_l). \tag{14}$$

x_i 为第 i 个对象, Q_l 为第 l 个簇的中心点, $d(x_i,Q_l)$ 为第 i 个对象到第 l 个簇中心的距离.

2.3.1 改进距离计算

距离 d 可通过相似度计算即 $d=\mathrm{SIM}(x_i,Q_j)=\mathrm{sim}(\mathrm{CT})\sum_{i=1}^{k}w_i\mathrm{sim}(E)$, 如果 CT 不同,则事件相似度为零. CT 为名义值,通过布尔运算 $\mathrm{d}(\mathrm{CT}_i,\mathrm{CT}_j)=\begin{cases}1,\mathrm{CT}_i=\mathrm{CT}_j;\\0,\mathrm{CT}_i\neq\mathrm{CT}_j.\end{cases}$ 求得. 重点分析序数、语义型值的相

似度，得到统一的事件距离公式：

$$\mathrm{sim}(E) = (d_{\mathrm{sq}} + d_{\mathrm{nm}} + d_{\mathrm{sem}}). \tag{15}$$

d_{sq} 为序数距离，d_{nm} 为名义距离，d_{sem} 为语义距离.

(a)序数距离：序数集合 $L=(l_1,l_2,\cdots,l_g)$，l_i 是语言序数值，g 是序数值数，语言序数 l_r 和 l_s 的心理亲近度为 π_{rs}，∇为映射 π：$L^2\rightarrow\nabla$，$\nabla=(\pi_{rs}\mid r,s\in(1,2,\cdots,g))$ 为所有语言序数心理亲近度的集合[14]，即 $\pi=(\pi^*_{12},\pi^*_{23},\cdots,\pi^*_{n-1,n})$，$\pi^*_{i-1,i}>0$. 根据已知序数亲近度向量计算不同序数的距离相似度.

(b)名义距离：考虑共现概率，相对分类属性 A_j 属性 A_i 的两值 x 和 y 距离用 $\max d^{ij}(x,y)=P_i(z\mid x))+P_i(z^c\mid y)-1$ 计算[15]. z 为值域 $\mathrm{Dom}(A_j)$ 子集，z^c 为 z 的补集，$P_i(z\mid x)$ 表示属性 i 值为 x 的数据对象在属性 j 上的值属于 z 的条件概率，$P_i(z^c\mid y))$ 表示属性 i 值为 y 的数据对象在属性 j 上的值属于 z^c 的条件概率.

(c)语义距离：语义结构树中，如 x 处在从 y 到根节点路径上，则 x 为 y 的层次祖先.如 x 和 y 处在同一层次，两点是相等的即 $x\equiv y$. $\mathrm{LCA}(x,y)$ 为 x 和 y 最低层次的共同祖先，任意两点的共同点为 $\mathrm{LCP}(x,y)$[16]：

$$\mathrm{LCP}(x,y)=\begin{cases}x\text{ 或 }y,x\equiv y;\\ x,\text{如果 }x\text{ 是 }y\text{ 的祖先;}\\ y,\text{如果 }y\text{ 是 }x\text{ 的祖先;}\\ \mathrm{LCA}(x,y),\text{其他.}\end{cases} \tag{16}$$

语义层次树中任意两节点之间的距离通过 $D(x,y)=d_x+d_y-2d_{\mathrm{LCP}(x,y)}$ 得到，d_x、d_y、$d_{\mathrm{LCP}(x,y)}$ 为节点 x、y、$\mathrm{LCP}(x,y)$ 位置到根节点的层次距离.

2.3.2　事件权重分配

每个事件类型在情景模式聚类中重要性不同，通过自适应熵权法分配事件权重[17].从 i 事件类型看，聚类后事件 l 样本落在第 j 情景中心的概率为 $P^k_{il}=\dfrac{\sum_{j=1}^{N}\mu^k_{ijl}}{\sum_{i=1}^{C}\sum_{j=1}^{N}\mu^k_{ijl}}$，以完全无序为参考初始状态，事件类型 i 的熵为 $H^k_{il}=-\dfrac{1}{K\ln C}\sum_{i=1}^{C}\sum_{l=1}^{L}P^k_{il}\ln P^k_{il}$，则整个系统熵为 $H=\sqrt{\sum_{k=1}^{K}(w_k^2H_K^2)}$. 由于 $\min(\cdot)^{1/2}$ 与 $\min(\cdot)$ 等效，则得到 $E(H)=\sum_{k=1}^{K}(w_k^2H_K^2)$，从而得出 $w_k=\dfrac{\lambda}{2H_k}$. 由于 $\dfrac{\lambda}{2H_k}$ 与 $\dfrac{1}{H_k}$ 等效，将事件类型权重为 $w_k=\dfrac{1}{H_k}$. H_k 为分类后事件类型的无序程度，$1-H_k$ 分类后事件类型有序程度，得到调整的权重为 $w_k=\dfrac{1-H_k}{(K-\sum_{k=1}^{K}H_k)H_k}$，进一步可到 $w_k\leftarrow\dfrac{w_k}{\sum_{k=1}^{K}w_k}$ 为当前聚类下事件类型的最优化权重.

2.3.3　改进聚类算法步骤：

Step 1　计算数据对象 x_i 之间的距离，得出 k 阀值的邻居集；

Step 2　选择中心性最大点作为初始簇类点 $\mathrm{Cen}(x_i)=\dfrac{\mid\mathrm{Nb}(x_i)\mid}{\mid\mathrm{Nbmax}\mid}$，$\mathrm{Nb}(x_i)$ 是对象 x_i 的 k 邻居集，Nbmax 为最大邻居数；

Step 3　计算该对象到所有簇中心的距离 $d(x_i,Q_l)$ 和簇间距离 $B(W,Q)$，将其分配到目标函数最小的簇中，更新相应簇的中心和属性权重；

Step 4　当所有对象都被分配后，重新计算数据对象到当前簇中心间的距离 $d(x_i,Q_l')$；

Step 5　重复 Step4，直到迭代过程中没有改变簇的数据对象存在；

Step 6　得到满足目标的 k 个簇 $\mathrm{SA}=(\mathrm{SA}_1,\mathrm{SA}_2,\cdots,\mathrm{SA}_k)$.

2.4　基于模式匹配的情景识别算法

知识库共聚类得到 k 个情景 $\mathrm{SA}=(\mathrm{SA}_1,\mathrm{SA}_2,\cdots,\mathrm{SA}_k)$，$\mathrm{SA}_i$ 为第 i 个情景模式，簇中心为向量 $\mathrm{SA}_i^*=(e_1^*,e_2^*,\cdots,e_m^*,\mathrm{CT}_i)$. 当前感知数据 $X_{t0}=(x_1,x_2,\cdots,x_m,\mathrm{CT}_{t0})$，为识别 X_{t0} 为哪一种情景，将其转化为 $\mathrm{CTE}_{t0}=(e_1,e_2,\cdots,e_m,\mathrm{CT}_{t0})$，然后与知识库中情景匹配，选择相似度最大情景模式作为识别结果：

$$Z=\mathrm{agrmax}(\mathrm{SIM}(\mathrm{SA}_1^*,\mathrm{CTE}_{t0}),\mathrm{SIM}(\mathrm{SA}_2^*,\mathrm{CTE}_{t0}),\cdots,\mathrm{SIM}(\mathrm{SA}_k^*,\mathrm{CTE}_{t0})). \tag{17}$$

3　实验分析

为验证本文改进混合聚类和情景识别方法的效果，首先利用已有的 4 个真实数据集进行实验对比，然后通过仿真制造过程数据集，进一步检测本文方法的优势.所有实验环境为 Windows10 操作系统，仿真开发环境为 Microsoft Visual C++ 2010，硬件是 CPU 为 Intel Core I5 2.6 GHz、内存为 4 GB 的鸿基台式机，算法通过 Matalabe 编码实现.

表 1　数据集描述

Tab.1　The description of data sets

数据集	数值属性数	分类属性数	对象个数	聚类数
Zoo	1	14	101	7
Heart	5	8	270	2
Flag	10	18	194	8
German Credit	7	13	1 000	2

3.1 真实数据集

从 UCI 学习库中找到 Zoo、Heart、Flag、German Credit4 个数集,如表 1.与 K-Prototype、DBSCAN、K-Center 算法比较,通过聚类准确率 r、平均聚类纯度 purity 两指标验证算法有效性.实验结果如表 2,在 Zoo 数据集上,本文算法比 K-Center 在聚类准确率 r、平均聚类纯度 purity 上效果稍差外,比 K-Prototype、DBSCAN 均高,可能原因是 Zoo 包含分类属性较多,K-Center 在处理该类型数据具有优势.在其他三个数据集上,本文算法均优于 K-Prototype、DBSCAN、K-Center 算法,效果优良.

表 2 4 类算法结果比较

Tab.2 Comparison results of 4 algorithms %

数据集	本文算法		K-Prototype		DBSCAN		K-Center	
	r	purity	r	purity	r	purity	r	purity
Zoo	0.863	0.923	0.628	0.783	0.747	0.765	0.906	0.946
Heart	0.842	0.975	0.577	0.736	0.752	0.894	0.545	0.741
Flag	0.746	0.876	0.513	0.782	0.556	0.823	0.624	0.824
German Credit	0.841	0.927	0.710	0.845	0.703	0.835	0.710	0.876

3.2 仿真制造过程

对奇瑞智能化生产车间调研,以某一零部件加工情景进行建模实验,分析智能设备如何在复杂动态环境下对事件聚类学习、在线实时识别. 关注事件为代加工产品到达、设备运行状态(温度、运转速度、震动频率),所处情境约束包括任务类型、电力水平. 参照实际情况生成 1 000 个时刻的数据分布,图 5 中共有两类情境约束为 $CT_1(Task_1, EL_1)$、$CT_2(Task_1, EL_2)$,4 类情景分别为 0 到 200 时刻的一般正常情景、200 到 500 时刻的一般异常情景,500 到 800 时刻的低压正常情景、800 到 1 000 时刻的低压异常情景.

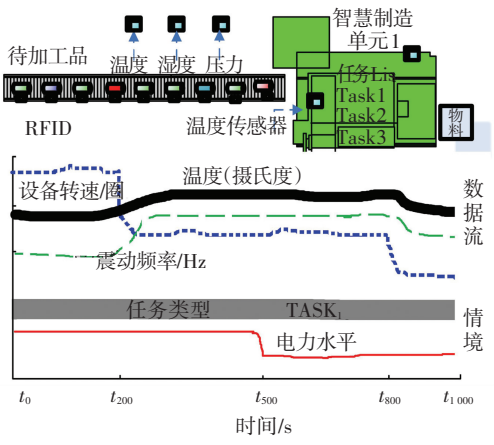


图 5 情境数据流

Fig.5 Context based data stream

将生成数据转化不同类型事件,构建情境事件图谱.运用改进混合聚类进行模式学习,与 K-Prototype、K-Center 比较,结果如表 3.本文算法在 r、purity 上均优于另外两种算法,具有良好的聚类效果.同时,由于考虑属性权重和不同数值类型,在时间上相对其他算法略差,但差别较小,仍具有较好的效率.

表 3 算法结果比较

Tab.3 Comparison results of 3 algorithms

算法	制造情景		
	准确率 r /%	聚类纯度 purity /%	时间 Time /ms
本文方法	98.5	99.87	986
K-Prototype	87.6	95.62	875
K-Center	90.4	92.51	763

考虑外部环境对事件聚类的影响,对有、无情境约束两类数据集应用改进算法聚类分析如表 4. 在 200 到 500、500 到 800 两个时间段,物体在温度、震动等上具有相似数值和分布模式.不考虑任务类型和电力水平,则将它们聚成一类则只得到 3 个情景模式.与实际情形出现偏差,有一类情景没有识别出来,导致 r、purity 较低.实际上,两个时间段由于电力水平不同,分别为正常电压下的异常情景和低压下的正常情景.考虑情景约束,得到 4 个情景模式与实际相符,显著提升识别的精度.

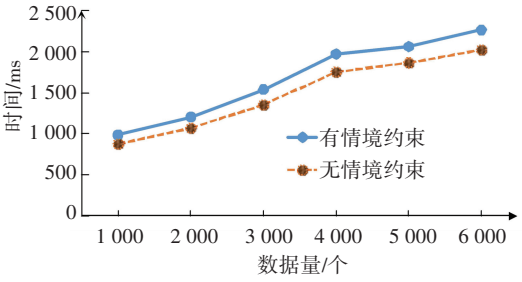
表 4 情境约束下的聚类结果比较

Tab.4 Comparison results of situation clustering in context constraint

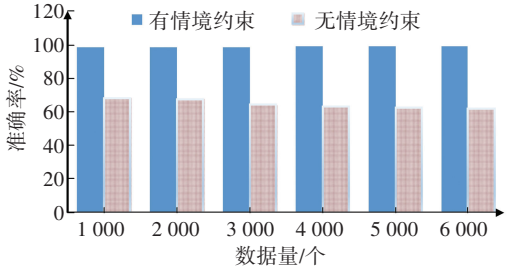
算法	制造情景		
	准确率 r /%	聚类纯度 purity /%	聚类数目
有情境约束	98.5	99.87	4
无情景约束	68.4	70.21	3

将原有 1 000 数据实例扩展到 2 000、3 000、4 000、5 000、6 000 5 种情形,情景数目仍为 4,进行实验 3.如图 6(a),本文算法不论有无情境约束,时间均随数据增长而增加但仍在接受范围内,具有良好的适应性. 图 6(b)中有情境约束时算法需计算情境相似度,增加计算复杂度,整体时间略大于无情境

约束.有情境约束的准确率随数据增长基本稳定，整体准确率明显高于无情境约束.无情境约束呈现下降趋势，可能由于聚类知识重叠，出现具有相同状态值但不同情境的实例.



(a)计算时间



(b)识别准确率

图 6　算法的时间和准确率

Fig.6　The time and precision rate of the proposed algorithm

4　结　论

1)考虑情境因素对物体状态变化的影响，构造基于多层次情境约束的事件模型，提出基于同生、情境、协同等事件算子的情景聚合过程.

2)建立物体状态与环境信息关联，界定事件的数据区间，构建复杂事件图谱实例.

3)整合名义、序数和语义距离计算，运用自适应熵权法分配事件权重，提出改进的混合聚类方法进行情景模式学习.

4)真实和仿真数据集实验均证明本文方法的优势，验证情境约束促进事件判断、情景识别的效果.上述结论为动态环境下的复杂事件处理与情景决策，提供了理论基础和应用参考.

参考文献

[1] LI B H, HOU B C, YU W T, et al. Applications of artificial intelligence in intelligent manufacturing: a review[J]. Frontiers of Information Technology & Electronic Engineering, 2017, 18(1):86-96.

[2] TERROSO-SAENZ F, VALDES-VELA M, SKARMETA-GOMEZ A F. A complex event processing approach to detect abnormal behaviors in the marine environment[J]. Information Systems Frontiers, 2016, 18(4):765-780.

[3] BRUNS R, DUNKEL J, MASBRUCH H, et al. Intelligent M2M: Complex event processing for machine to machine communication [J]. Expert Systems with Applications, 2015, 42(3): 1235-1246.

[4]赵会群，孙晶，杨岩坤，等. 复杂事件模式检测与 CEP 测试数据生成算法研究[J]. 计算机学报，2017,44(1):256-272.
ZHAO Huiqun, SUN Jing, YANG Yankun, et al. Study on algorithm of checking event pattern and generating test data for complex event processing[J].Chinese Journal of Computers, 2017,44(1):256-272.

[5] O'DONOVAN P, LEAHY K, BRUTON K, et al. Big data in manufacturing: a systematic mapping study[J]. Journal of Big Data, 2015, 2(1):1-22.

[6] CUGOLA G, MARGARA A. Complex event processing with T-REX [J]. Journal of Systems & Software, 2012, 85(8):1709-1728.

[7] CHONGWATPOL J, SHARDA R. RFID-enabled track and traceability in job-shop scheduling environment[J]. European Journal of Operational Research, 2013, 227(3): 453-463.

[8] YU M, LU R, HONG S H. A real-time decision model for industrial load management in a smart grid[J]. Applied Energy, 2016, 183 (12) :1488-1497.

[9] CADENAS J T, MAR, N, N, et al. Context-aware fuzzy databases [J]. Applied Soft Computing, 2014, 25(C):215-233.

[10]ZHU A, GAO R, UCHIDA S. Could scene context be beneficial for scene text detection? [J]. Pattern Recognition, 2016, 58(10): 204-215.

[11]LEE O J, JUNG J E. Sequence Clustering-based automated rule generation for adaptive complex event processing[J]. Future Generation Computer Systems, 2016,66(2):100-109.

[12]孙锴，高建民，高智勇，等. 基于故障图谱的企业级故障模式识别方法[J]. 计算机集成制造系统，2015，21(2)：519-527.
SUN Kai, GAO Jianmin, GAO Zhiyong, et al. Plant-wide fault pattern recognition method based on fault-spectrum[J]. Computer Integrated Manufacturing Systems, 2015, 21(2):519-527.

[13]PATHAK A, PAL N R. Clustering of mixed data by integrating fuzzy, probabilistic, and collaborative clustering framework[J]. International Journal of Fuzzy Systems, 2016, 18(3): 339-348.

[14]GARE, A-LAPRESTA J, LUIS, et al. Ordinal proximity measures in the context of unbalanced qualitative scales and some applications to consensus and clustering[J]. Applied Soft Computing, 2015, 35 (C):864-872.

[15]NOORBEHBAHANI F, MOUSAVI S R, MIRZAEI A. An incremental mixed data clustering method using a new distance measure [J]. Soft Computing, 2015, 19(3):731-743.

[16]HSU C C, HUANG Y P. Incremental clustering of mixed data based on distance hierarchy[J].Expert Systems with Applications: an International Journal, 2008, 35(3):1177-1185.

[17]黄海新，孔畅，于海斌，等. 自适应特征熵权模糊 C 均值聚类算法的研究[J]. 系统工程理论与实践，2016，36(1):219-223.
HUANG Haixin, KONG Chang, YU Haibin, et al. Research on adaptive entropy weight fuzzy c-means clustering algorithm[J].Systems Engineering-Theory & Practice, 2016, 36(1):219-223.

(编辑　苗秀芝)