基于数据场的粗糙聚类算法

李 学 苗夺谦 冯琴荣
(同济大学计算机科学与技术系 上海 201804)

摘 要 聚类分析是数据挖掘的研究热点。传统的聚类算法都是把一个对象精确地划分到一个聚类簇中,类别之间的界限是非常精确的。随着Web挖掘技术的发展,精确地划分每个对象的聚类算法面临着巨大的挑战。根据数据场理论和经典粗糙集理论所具有处理不精确与不确定性数据的特性,提出一种新的基于数据场的粗糙聚类算法,该粗糙聚类算法采用势值作为对象的划分依据,避免传统粗糙聚类算法一贯采用基于欧氏距离的划分方法。算法首先通过对数据对象进行粗分然后再不断迭代细分,直至形成稳定的聚类簇。实验分析过程中,把提出的算法与粗糙K-means算法和粗糙K-medoids算法进行了比较,结果表明该算法在交叉数据集上具有较好的聚类效果,而且收敛速度较快。

关键词 粗糙聚类,数据场,势值,Davies-bouldin指标

Rough Clustering Algorithm Based on Data Field

LI Xue MIAO Duo-qian FENG Qin-rong
(Department of Computer Science and Technology, Tongji University, Shanghai 201804, China)

Abstract Clustering analysis is the hotspot in Data mining, all the conventional clustering algorithms precisely put the each object into one cluster, the bounders between clusters are precise, as the development of the Web mining, clustering algorithms that precisely divide each object face great challenges. Based on the data field theory and classic rough set theory's character that processes the uncertainty and imprecise data, a novel rough clustering algorithm based on data field was proposed, it divides the objects through computing potential value, which avoids the conventional rough clustering partition method based on euclidean distance. The approach iterates from rough to un-rough incessantly till the stable clusters form. At the experimental analysis process, we compared the algorithm that we proposed with rough K-means algorithm and rough K-medoids algorithm, the result shows the algorithm that we proposed has better clusters on the crossed datasets and fast convergence.

Keywords Rough clustering, Data field, Potential value, Davies-bouldin index

1 引言

聚类算法是数据挖掘的研究热点。传统的聚类算法都是将一个对象精确地划分到一个聚类簇中,聚类簇之间的界限是非常精确的。随着Web挖掘技术的迅速发展,保持聚类簇之间界限的严格精确性面临着巨大的挑战,聚类簇之间出现了模糊的界限。Joshi和Krishnapuram认为在Web挖掘中聚类簇之间存在交集,对象不是精确地划分到每个聚类簇中,而是以不同的程度隶属于每个类[1]。粗糙集理论正是处理这种不精确与不确定数据的有效工具。P. Lingras等人首次将粗糙集理论引入到数据聚类中来,其主要思想是将聚类簇看成是一个不确定的集合,通过粗糙集的上近似和下近似的概念来描述聚类簇的轮廓。粗糙聚类(Rough Clustering,简称RC)算法应用到多个领域,而且取得了非常好的效果[2,3]。但是,目前的RC算法在对象归属划分时,考虑的都是基于点与点之间的欧氏距离[4,5],忽视了待划分对象与聚类簇内其他对象的作用关系。而数据场理论认为空间中的对象是受到场源作用的,对象在场源作用下产生势值,拥有相同势值的对象形成等势面,等势面以局部最大势值形成抱团簇。

本文根据数据场理论的相关知识,将数据场理论应用到RC中来,提出了一种新的RC算法。这种算法依据待划分对象的势值大小来进行划分,划分到聚类簇下近似中的对象肯定属于这个类。划分到聚类簇上近似中的对象可能属于这个类,也有可能不属于这个类,这样从初次粗分到不断迭代细分,直至形成稳定的聚类簇。

本文结构组织如下:第2节介绍RC算法研究背景,包括目前主要应用的粗糙K-means算法和粗糙K-medoids算法,并对二者进行比较;第3节首先简要介绍一下数据场理论相关知识,接着给出基于粗糙模型的数据场势函数定义,然后给出基于数据场的RC算法(Rough Clustering Algorithm based on Data Field)具体步骤;第4节是聚类算法实验结果比较分析,将本文提出的算法与粗糙K-means算法和粗糙K-medoids算法进行比较,算法采用Rough Davies-Bouldin(RDB)来度量聚类效果,同时也对3种RC算法的收敛速度进行了

到稿日期:2008-03-20 本文受国家自然科学基金资助项目(60475019, 60775036),2006年博士学科点专项科研基金(20060247039)资助。
李 学(1984—),男,硕士生,研究方向为数据挖掘、粗糙集,E-mail:lixue0501@hotmail.com;苗夺谦 教授,博士生导师,主要研究方向为粗糙集、数据挖掘、粒度计算;冯琴荣 博士生,研究方向为粒度计算。

比较；最后是总结。

2 RC算法的研究背景

聚类是一种常见的数学分析工具。传统的聚类算法都是将每个对象归属到一个确定的类别信息中，类别之间的界限是非常精确的。但是随着Web挖掘技术的发展，这种聚类算法面临着巨大的挑战。鉴于粗糙集具有处理不精确和不确定数据的优点，P. Lingras等人首次将粗糙集理论引入到数据聚类中来，而且取得了很好的应用效果。目前基于RC的算法主要有两种：粗糙K-means算法和粗糙K-medoids算法。

2.1 粗糙K-means算法

粗糙K-means算法将对象划分到聚类簇的下近似和上近似，通过下近似和上近似计算新的聚类中心，然后重新划分对象。此过程反复进行，直到形成稳定的聚类结果。整个过程主要包括对象划分和聚类中心调整两个步骤。

粗糙K-means算法利用点之间距离的差值将 N 个对象划分到 K 个聚类簇的上下近似中，粗糙K-means算法聚类中心调整公式表示为：

如果 $\overline{Am}_i-\underline{Am}_i\neq\Phi$，

$$m_i=w_{lower}\frac{\sum\limits_{v_j\in\underline{Am}_i}v_j}{|\underline{Am}_i|}+w_{upper}\frac{\sum\limits_{v_j\in\overline{Am}_i-\underline{Am}_i}v_j}{|\overline{Am}_i-\underline{Am}_i|}$$

否则：$m_i=\frac{\sum\limits_{v_j\in\underline{Am}_i}v_j}{|\overline{Am}_i|}$

粗糙K-means算法存在明显的局限性，即粗糙K-means算法对聚类簇交界的对象不能很好地分类，同时在重新计算聚类中心以后，所有对象要根据新的聚类中心重新进行划分，因而迭代时间长、收敛速度慢。鉴于粗糙K-means算法存在的不足，G. Peters在处理噪声数据方面对粗糙K-means方法进行了改进[6,7]，S. Mitra采用遗传算法优化粗糙K-means算法的参数[8]。在文献[9]中，K. Vogts等人也采用遗传算法对粗糙K-means算法进行了改进。

2.2 粗糙K-medoids算法

K-medoids算法是Kaufman提出的[10]。与传统K-means算法不同的是，K-medoids是用真实存在的对象来代表聚类的质心。K-medoids算法以CPC(Compactness of the clustering)作为算法迭代终止的度量准则。

$$CPC=\sum_{k=1}^{K}CPC(C_k)$$

其中 $CPC(C_k)=\sum\limits_{x_n\in C_k}d(x_n,m_k)$

Peters等人在粗糙K-means算法和经典K-medoids算法的基础上提出了粗糙K-medoids算法。粗糙K-medoids就是引入粗糙集理论上下近似的概念，对经典K-medoids算法的度量准则CPC进行重新定义，即RCPC(Rough Compactness of the Clustering)：

$$RCPC=\sum_{k=1}^{K}RCPC(C_k)$$

其中 $RCPC(C_K)=w_l\sum\limits_{x_n\in\underline{C}_k}d(x_n,m_k)+w_b\sum\limits_{x_n\in\overline{C}_k-\underline{C}_k}d(x_n,m_k)$

2.3 二者比较

粗糙K-medoids算法相对于粗糙K-means算法来说，它的优点是采用真实存在的数据对象作为聚类的代表点，因此对离群数据比较敏感，同时引入RCPC作为判定聚类算法终止的条件，因此用户可以根据自身需求制定评判准则；它的缺点是高质量代表点的选择相对较为困难，数据分布的细微变化会对聚类结果产生很大影响。而粗糙K-means算法采用均值作为聚类簇的代表点，更能反映聚类簇的基本分布情况，两者算法各有利弊。但是，不论是粗糙K-means算法还是粗糙K-medoids算法，其迭代次数较多，收敛速度较慢；对象划分的依赖准则是基于点与点的欧氏距离，忽视待划分对象与聚类簇内其他数据对象的关系。针对粗糙K-means算法和粗糙K-medoids算法存在的以上问题，我们提出一种新的RC算法，该RC算法在聚类代表选择上既考虑了对象均值，也考虑了真实存在的对象，同时收敛速度较快。

3 基于数据场的RC算法

3.1 数据场的介绍

场的概念是1837年英国物理学家法拉第提出的，是用于描述物质间粒子的非接触相互作用。随着认知物理学的发展，人们将其抽象为一个数学概念，用来描述某个物理量或者数学函数在空间中的分布。李德毅院士在传统物理场的基础上，提出了基于数据对象的数据场理论[11]。数据场理论是把N维空间中的对象看成是有相互作用的，就是在没有外力的作用下，对象也能相互吸引而相向运动。数据场理论的引入是数据挖掘领域中的一个突破，因为传统的数据挖掘算法中只考虑对象之间一对一的映射关系，忽视了一对多或者多对一关系。而数据场理论克服了这样的问题，因为它把空间中某点的状态看成是其他对象共同作用的结果。

数据场理论采用势函数来表示对象间的相互关系，通过计算空间中任意对象的势值来判定对象的归属情况，不像粗糙K-means算法和粗糙K-medoids算法那样只计算任意对象到聚类中心或质心的欧氏距离来判定对象的归属。根据数据场理论，处于一个等势面上的所有对象具有相同的势值，最终归属一个类别中的对象以局部最大势值形成抱团簇。

考虑到短程场作用更有利于数据分布的聚簇特性，本文采用具有良好性质的高斯函数来定义数据场的标量势。鉴于数据场理论具有很好的离群检测功能，其在图像处理和入侵检测领域都取得了很好的应用[12]。

3.2 势函数相关概念

势函数是数据场理论的核心概念。下面我们结合粗糙模型给出势函数的相关定义。

定义1 设 U 是一个论域，$X=\{X_1,X_2,\cdots,X_K\}$ 是论域 U 上的一个划分，则任意对象 $x\in X_i$ 的势函数定义如下：

$$p(x,X_i)=\sum_{x_j\in X_i}\exp(-\frac{||x-x_j||^2}{2\sigma^2})$$

其中 σ 用于控制对象间的相互作用，称为影响因子；$||x-x_j||$ 表示对象 x 与对象 x_j 的距离。

定义2 设 U 是一个论域，$X=\{X_1,X_2,\cdots X_K\}$ 是论域 U 上的一个覆盖，对任意给定的 $X_i\in X$，则任意对象 $x\in X_i$ 的势函数定义成如下的分段函数形式：

(1) 当 $\overline{B}(X_i)-\underline{B}(X_i)\neq\phi\wedge\underline{B}(X_i)\neq\Phi$，$p(x,X_i)=p_{low}(x,X_i)+p_{up}(x,X_i)$

其中 $p_{low}(x,X_i)=w_{low}\sum\limits_{x_j\in\underline{B}(X_i)}\exp(-\frac{||x-x_j||^2}{2\sigma^2})$

$$p_{up}(x,X_i)=w_{up}\sum_{x_j\in\overline{B}(X_i)-\underline{B}(X_i)}\exp(-\frac{||x-x_j||^2}{2\sigma^2})$$

(2) 当 $\overline{B}(X_i)-\underline{B}(X_i)=\Phi$，

$$p(x,X_i)=\sum_{x_j\in\underline{B}(X_i)}\exp(-\frac{||x-x_j||^2}{2\sigma^2})$$

(3) 否则，当$\overline{B}(X_i)\neq\Phi\wedge\underline{B}(X_i)=\Phi$，

$$p(x,X_i)=\sum_{x_j\in\overline{B}(X_i)}\exp(-\frac{||x-x_j||^2}{2\sigma^2})$$

其中 w_{low}，w_{up} 分别表示下近似和上近似中数据对象对 x 对象势值的权重，并且满足 $w_{low}+w_{up}=1$。

性质 1 设 U 是一个论域，$X=\{X_1,X_2,\cdots,X_k\}$ 是论域 U 上的一个覆盖，对于任意给定的 $X_i\in X$，则 X_i 的上近似 $\overline{B}(X_i)$ 与下近似 $\underline{B}(X_i)$ 具有如下性质：

(1) $\Phi\subseteq\underline{B}(X_i)\subseteq\overline{B}(X_i)\subseteq U$；

(2) $\underline{B}(X_i)\cap\underline{B}(X_j)=\Phi, i\neq j$；

(3) $\underline{B}(X_i)\cap\overline{B}(X_j)=\Phi, i\neq j$；

(4) 如果 $x_i\in U$ 不属于任意集合的下近似，那它至少属于 2 个集合的上近似。

3.3 基于数据场的 RC 算法

根据上面介绍的数据场理论势函数概念和粗糙集理论上下近似所具有的特性，下面我们给出基于数据场的 RC 算法步骤。

输入：数据集 U，影响因子 σ，w_{low}，K；

输出：数据的聚类结果：$\{X_1,X_2,\cdots,X_K\}$；

算法步骤：

Step1 选取 K 个聚类簇的均值作为初始场源；

Step2 对 $x\in U$，计算 $p(x,X_i)$，$i\in 1,2,\cdots,K$，设 $p(x,X_l)=\max\{p(x,X_1),p(x,X_2),\cdots,p(x,X_K)\}$；

Step3 计算 $p(x,X_l)-p(x,X_j)$，$j=1,2,\cdots,K$，如果 $p(x,X_l)-p(x,X_j)\geqslant\alpha$，则 $x\in\underline{B}(X_l)$，否则 $x\in\overline{B}(X_l)\wedge\overline{B}(X_j)$；

Step4 对 $\forall i\in 1,2,\cdots,K$，如果 $\underline{B}(X_i)$ 基本保持不变，转 Step5；否则，如果 $\exists i,t\in 1,2,\cdots,K$，对 $\forall x\in(\overline{B}(X_i)-\underline{B}(X_i))\cap(\overline{B}(X_t)-\underline{B}(X_t))$，计算 $p(x,X_i)$ 与 $p(x,X_t)$，其中设 $p(x,X_l)=\max\{p(x,X_i),p(x,X_t)\}$，$p(x,X_j)=\min\{p(x,X_i),p(x,X_t)\}$，转 Step3；

Step5 算法结束，输出聚类结果。

4 实验分析

4.1 聚类有效性度量指标

基于划分的聚类算法，正如本文中所给出的基于数据场的 RC 算法、粗糙 K-means 算法和 K-medoids 算法，聚类个数 K 值预先是要设定的。在聚类簇个数 K 已知的情况下，目前已存在很多度量聚类效果的指标函数，如文献[13]中采用的 SICV(Sum of intra-cluster variation)度量指标。SICV 依赖于 K 值的选取。

Mitra 采用遗传算法来优化基于 Davies-Bouldin 聚类有效性度量指标函数[14]。由于 Davies-Bouldin 指标函数独立于聚类簇初始个数 K 的设定，因此不同的聚类划分算法都可以采用它来进行比较。

Davies-Bouldin 指标函数就是聚类簇内部的距离与聚类簇之间的距离之比的平均值。好的聚类效果就是使得类内距达到最小，类间距达到最大。Davies-Bouldin 指标函数定义如下：

$$DB=\frac{1}{K}\sum_{k=1}^{K}\max_{k\neq l}\{\frac{S(X_k)+S(X_l)}{d(X_k,X_l)}\}$$

然而，在 RC 算法中，处于聚类 X_i 上近似中的对象和处于下近似中的对象对类内距的贡献不一样，因此 Mitra 等人在文献[15]中对 Davies-Bouldin 指标函数进行了改进。改进之后的 Davies-Bouldin 指标函数定义为 RDB：

$$RDB=\frac{1}{K}\sum_{k=1}^{K}\max_{k\neq l}\{\frac{RS(X_k)+RS(X_l)}{d(X_k,X_l)}\}$$

在下面的实验分析中，我们采用基于改进的 Davies-Bouldin 指标函数(RDB)来对文中涉及的 3 种 RC 算法进行对比分析。实验分析数据包括 Synthetic 数据、UCI 标准数据库中的 IRIS 数据集、ZOO 数据和 Colon Cancer 数据。

4.2 Synthetic 数据

Synthetic 数据是一个二维分布数据，由 10 个取值在[0，1]上的数值型数据对象构成。在文献[16]中，Georg Peters 等人分别采用粗糙 K-means 算法和粗糙 K-medoids 算法对 Synthetic 数据进行聚类分析，初始聚类个数 $K=2$，聚类结果分别如图 1 和图 2 所示。

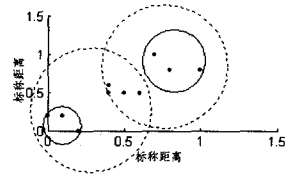

图 1 粗糙 K-means 算法

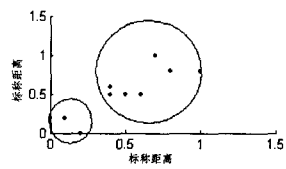

图 2 粗糙 K-medoids 算法

从图 1 中我们可以看出，采用粗糙 K-means 算法聚类分析之后，两类对象之间的交界处是不容易区分的，采用粗糙 K-medoids 算法对之聚类分析之后得到的结果却和粗糙 K-means 算法有很大的差别，所有对象被精确地划分到两个类的下近似中，处于聚类簇边界的集合为空集。

同时我们利用本文提出的算法也对 Synthetic 数据进行了聚类分析，初始聚类个数 $K=2$，势值函数的影响因子 $\sigma=0.20$，$w_{low}=0.9$，聚类结果如图 3 所列。

从图 3 中我们可以看出，对象划分结果采用粗糙 K-medoids 算法相同，即图 3 所示的左边的 3 个对象属于一个类，右边的 7 个对象属于另外一个类。下面我们采用 RDB 指标来度量这 3 种 RC 算法的效果，结果如表 1 所列。

表 1 Synthetic 数据的 RDB 指标

Algorithm	Rough Davies-Bouldin index
Rough K-means	0.638
Rough K-medoids	0.654
RCDF	0.654

4.3 IRIS 数据

UCI 数据库中的 IRIS 是模式识别中的一个经典数据集，记录的是一些植物特征数据。共有 5 个属性，其中包括 4 个数值型条件属性、1 个分类型决策属性。共有 150 条数据对象和 3 种决策值，我们把这些数据样本简记为 1，2，…，150，其中 1～50，51～100，101～150 分别对应一种决策。我们去掉决策属性后，分别采用粗糙 K-means 算法、粗糙 K-medoids 算法和本文提出的 RCDF 算法对数据进行聚类分析。对粗糙 K-means 算法和粗糙 K-medoids 算法我们设初始聚类个数 $K=3$，距离阈值 $\xi\in[0.05,0.1]$，$w_{low}=0.9$；我们用本文提出的 RCDF 算法也对 IRIS 数据集做了聚类分析，势值阈 $\alpha\in[0.15,0.18]$，$\sigma=1.4$，$w_{low}=0.9$，$K=3$，采用 RDB 指标来度

量这 3 种 RC 算法的效果，如表 2 所列。

表 2　IRIS 数据的 RDB 指标

Algorithm	Rough Davies-Bouldin index
Rough K-means	0.630
Rough K-medoids	0.627
RCDF	0.622

由于 IRIS 数据集的分布特性，即第二类与第三类数据对象具有较强的难分性，因此实验得到的 RDB 指标相差不大。但是，我们对 3 种 RC 算法的收敛速度进行了比较。从图 4 中我们可以看出，本文提出的 RCDF 算法具有更快的收敛速度。

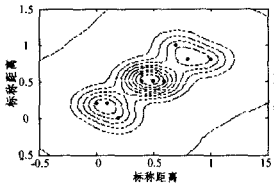

图 3　基于数据场的粗糙聚类算法

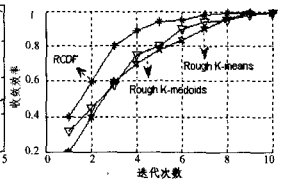

图 4　IRIS 的收敛速度

4.4　ZOO 数据

UCI 数据库中的 ZOO 也是模式识别中的一个经典数据集，记录的是一些动物的特征数据。共有 101 条记录，每个对象有 18 个特征属性。我们分别采用基于粗糙 K-means 算法和粗糙 K-medoids 算法对其进行 RC 聚类分析，设 $K=7$，距离阈值 $\xi\in[0.08,0.15]$，$w_{low}=0.9$；我们采用本文提出的算法也对其进行了聚类分析，势值阈值 $\alpha\in[0.20,0.25]$，$\sigma=0.7$，$w_{low}=0.9$，$K=7$，采用 RDB 来度量这 3 种 RC 算法的效果，如表 3 所列。

表 3　ZOO 数据的 RDB 指标

Algorithm	Rough Davies-Bouldin index
Rough K-means	0.805
Rough K-medoids	0.819
RCDF	0.782

从表 3 中我们可以容易看出，本文提出的算法具有较小的 RDB 值，具有比粗糙 K-means 算法和粗糙 K-medoids 算法更好的聚类效果。

下面我们从实验算法的收敛速度来对 3 种 RC 算法加以比较，比较结果如图 5 所示。

4.5　Colon Cancer 数据

Colon Cancer 数据是基因表达数据，其中包括 62 条基因表达数据。每个基因表达数据共有 2000 个特征属性，有 2 个类别信息，即正常基因和癌症基因。由于基因各个特征属性之间具有极强的相关性，在文献[16]中，Peters 等人对 Colon Cancer 数据首先进行了约简，然后采用 Davies-Bouldin 指标度量了粗糙 K-means 算法和粗糙 K-medoids 算法的效果。而本文我们采用原始数据进行实验分析，改用 RDB 指标重新度量 3 种 RC 算法的效果，设 $K=2$，势值域值 $\alpha\in[0.05,0.10]$，$\sigma=1.2$，$w_{low}=0.9$，结果如表 4 所列。

表 4　Colon Cancer 数据的 RDB 指标

Algorithm	Rough Davies-Bouldin index
Rough K-means	0.561
Rough K-medoids	0.528
RCDF	0.503

从表 4 中我们可以看出 RCDF 算法具有更好的聚类效果，同时我们也对算法的收敛速度进行了比较，结果如图 6 所示。

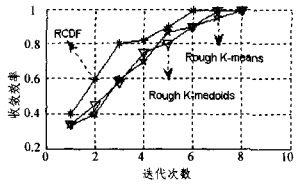

图 5　ZOO 的收敛速度

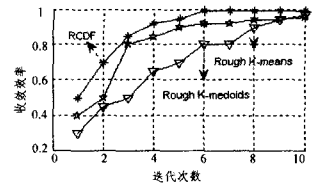

图 6　Colon Cancer Data 的收敛速度

结束语　本文将经典数据场理论引入到 RC 算法中来，提出了一种新的 RC 算法。传统的 RC 算法是根据对象到聚类中心或质心距离来划分对象的归属。本文提出的 RC 算法不是根据计算点与点的欧氏距离，而是采用数据场势值的思想将对象划分到不同的聚类簇中；对象划分通过粗分再到细分，直至形成稳定的聚类效果。实验分析阶段通过引入改进的 Davies-Bouldin 指标来度量聚类算法的效果，分别对 Synthetic 数据、IRIS 数据、ZOO 数据集和 Colon Cancer 数据集进行了实验分析比较，结果表明，在具有交叉的数据集上本文提出的算法较粗糙 K-means 算法和粗糙 K-medoids 算法具有更好的聚类效果；同时本文提出的算法还具有迭代次数较少、收敛速度较快的优点。算法的缺点就是处理的数据必须是数值型的，对非数值型数据对象不能很好的处理，且初始聚类个数 K 需要预先给出，这些都是有待进一步研究的工作。

参 考 文 献

[1] Joshi A，Krishnapuram R. Robust Fuzzy Clustering Methods to Support Web Mining // Proceedings of the Workshop on Data Mining and Knowledge Discovery. SIGMOD'98. Seattle, 1998: 15/1-15/8

[2] Lingras P, Yao Y Y. Time Complexity of Rough Clustering: GAs versus K-means // Proceedings of the Third International Conference of Rough Sets and Current Trends in Computing 2002. Berlin, Germany: Springer Verlag, 2002: 263-270

[3] Ozyer T, Alhajj R, Barker K. Utilizing Rough Sets and Multi-objective Genetic Algorithms for Automated Clustering // the 4th International Conference, TSCTC 2004. Heidelberg Germany: Springer Verlag, 2004: 567-572

[4] Lingras P, West C. Interval Set Clustering of Web Users with Rough K-means. Journal of Intelligent Informmation Systems, 2004, 23(1): 5-16

[5] Fahim A M, Salem A M, Torkey F A. An efficient enhanced k-means clustering algorithm. Journal of Zhengjiang University-Science A, 2006, 7(10): 1626-1633

[6] Peters G. Outliers in rough k-means clustering // Proc. First International Conference on Pattern Recognition and Machine Intelligence. Volume 377 of LNCS. Kilkata: Springer Verlag, 2005: 702-707

[7] Peters G. Some refinements of the rough k-means. Pattern Recognition, 2006, 20: 1481-1491

[8] Mitra S. An Evolutionary Rough partitive clustring Pattern Recognition letters, 2004, 25(12): 1439-1449

（下转第 244 页）

离，短暂停留后继续执行航班 t_6，到达机场 p_1。在到达 p_1 之后由同一飞机机组执行航班 t_7，到达机场 p_7，至此结束当天任务。

图 1 所示的 t_m，t_n 为两类特殊变迁，表示标识的资源结束当天任务，并将于以后执行飞行任务，目的地暂时未知，用 p_m，p_n 保存托肯。航班的执行顺序和使用资源如图所示，每架航班的执行都需要使用前一架航班的某一资源，前一架航班的延误，必然使后续航班的执行受到影响。图中展现了初始航班 t_1 的机组和飞机两种资源对后续航班的影响范围，而 $(p_1)t_1(p_2)t_3(p_4)t_5(p_6)$ 则反映了飞机 a_1 在单机执行多航班时产生的链式波及延误情况。

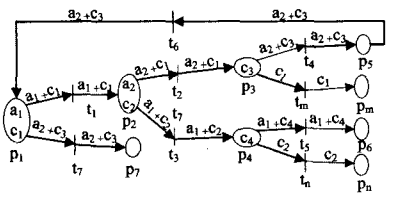

图 1　基于 CPN 的航班延误波及链模型

4.2　有色出现网的应用模型

为观察航班执行的先后顺序和并发情况，以及某架航班的延误对后续航班的影响，我们对上述模型进行动态行为的描述，以出现网的形式反映由一架航班的延误带来的影响。

以 t_1 作为初始变迁为例，得到图 2 所示的一个有色出现网，该出现网反映了一天内 t_1 所使用的资源对后续航班的全部影响。其中 $(p_1)t_1(p_2)t_3(p_4)t_5(p_6)$ 则反映了初始飞机 a_1 在单机执行多航班的情况下其延误将对下游航班产生的影响。

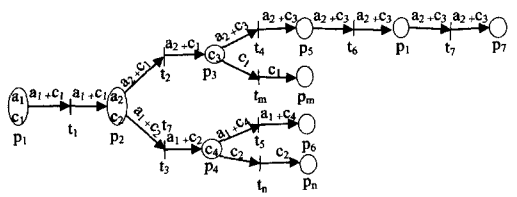

图 2　航班延误波及链模型的一个有色出现网

我们把有向弧的方向看作时间流动方向，该出现网给出的是偏序时间。网中没有顺序关系的两架航班是并发的。这种有色出现网直观地反映了机组和飞机资源的分配和流动情况，航班执行的顺序和并发关系，以及一架航班的延误对其他航班的影响，客观地反映了航班执行过程中的相互关系和延误波及情况。

这种有色出现网不仅保留了有色 Petri 网的优点，清楚地反映了飞机和机组资源的出处、流动情况和航班的执行细况，也遵循了网论在时间系统上的观点，使航班执行的先后和并发顺序一目了然，避免了因大量库所和变迁给 Petri 网模型的描述和理解带来的复杂性。

结束语　有色 Petri 网作为描述异步并发系统的一种高级网系统，图形简洁但函数关系复杂。用出现网对其动态行为进行刻画，可以清楚地展现各种资源的分布和流动情况，以及各个变迁的顺序和并发关系。鉴于基本出现网在描述高级网系统时带来的图形复杂性，本文在基本出现网的基础上提出了一种有色出现网，对库所容量和托肯、变迁的颜色进行扩充，使之更适合直观简洁地记录有色 Petri 网这种高级网系统的客观行为。本文利用有色 Petri 网对航班延误的链式波及情况进行建模，并构造了它的一个有色出现网，简洁而清晰地展现了各航班执行时所需的资源在机场的分布和流动情况，以及一架航班的执行情况对后续航班的链式影响，同时，出现网提供的偏序时间关系也使得各航班执行的顺序和并发关系一目了然。可见，这种有色出现网能够对基于有色 Petri 网的模型进行更加简洁和直观的行为记录。

参 考 文 献

[1] 蒋昌俊. Petri 网的行为理论及其应用. 第一版. 北京：高等教育出版社，2003：19-21

[2] 袁崇义. Petri 网原理与应用. 第一版. 北京：电子工业出版社，2005：97-103

[3] Lomazova I. On Occurrence Net Semantics for Petri Nets with Contacts. Fundamentals of Computation Theory, 1997, 1279：317-328

[4] Kurt J. Colored Petri Nets and The Invariant-method. Theoretical Computer Science, 1981, 14：317-336

[5] 唐培和. 着色网到基本网的等价变换. 广西工学院学报，1995，6(3)：53-57

[6] 郝克刚，葛玮. 论高级 Petri 网系统的等价谱系. 计算机学报，1993，16(7)：553-558

[7] Schaefer L, Wojcik L. Flight Connections and Their Impacts on Delay Propagation// Digital Avionics Systems Conference. 2003, 1：5. B. 4-5. 1-9

[8] Ahmad B S, Cphn A, Guan Yihan. Analysis of the Potential for Delay Propagation in Passenger Aviation Flight Networks. Sloan Industry Studies Working Papers, WP-2007-11. 2007

(上接第 206 页)

[9] Vogts K, Pope N. Generating compact rough cluster descriptions using an evolutionary algorithm// Proceedings of the 6th Annual Genetic and Evolutionary Computation Conference. Heidelberg, Germany：Springer Verlag, 2004：1332-1333

[10] Voges K E, Pope N K, Brown M R. Heuristics and Optimization for Knowledge Discovery. chapter Cluster Analysis of Marketing Data Examining On-Line Shipping Orientation：A Comparison of k-Means and Rough Clustering Approaches. Hershey PA：Idea Group Publishing, 2002：207-224

[11] 淦文燕，李德毅. 一种基于数据场的层次聚类方法. 电子学报，2006(2)：68-72

[12] Xie Feng, Bai Shou. Dectecting Novel Network Attacks with a Data Field. Springer Berlin, Heidelberg, Volume 3917. 2006：66-72

[13] Sheng Weiguo, Liu Xiaohui. A genetic k-medoids clustering algorithm. Journal of Heuristics, 2006, 12：447-466

[14] Davies D L, Bouldin D W. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979, 1：224-227

[15] Mitra S, Banka H, Pedrycz W. Collaborative Rough Clustering. Rough Sets, Case-Based Reasoning and Knowledge Discoery, 2005, 3776：768-773

[16] Peters G, Lampart M. A Partitive Rough Clustering Algorithm. Rough Sets and Current Trends in Computing, 2006, 4259：657-666