一种面向多属性不确定数据流的模体发现算法

王 菊* 刘付显
(空军工程大学防空反导学院 西安 710051)

摘 要：该文针对多属性不确定数据流的频繁模式发现问题，借鉴生物信息学中的模体发现思想，提出了一种基于MEME(Multiple Expectation-maximization for Motif Elicitation)的多属性不确定数据流模体发现算法。该算法根据不确定数据流的特点，设计了基于混合型模型的不确定滑动窗口更新计算方法，改进了 SAX(Symbolic Aggregate approXimation)的符号化策略，提出了不同滑动窗口下多属性模体的相似性分析方法。在实验当中，用防空反导情报传感器网络中的一组不确定数据流验证了其功能，通过植入不同数目的模体测试了其发现准确率，并在元组有效概率设置为 1 的条件下与已有算法进行了比较，结果表明：该算法可以较准确地发现多属性不确定数据流中的频繁模式。

关键词：数据挖掘；模体发现；不确定数据流；MEME(Multiple Expectation-maximization for Motif Elicitation)算法
中图分类号：TP391 **文献标识码**：A **文章编号**：1009-5896(2017)01-0159-08
DOI: 10.11999/JEIT160247

Motif Discovery Algorithm for Multiple Attributes Uncertain Data Stream

WANG Ju LIU Fuxian
(*Air and Missile Defenses College*, *Air Force Engineering University*, *Xi'an* 710051, *China*)

Abstract: Algorithm of motif discovery for multiple attributes uncertain data stream is proposed on the basis of MEME (Multiple Expectation-maximization for Motif Elicitation), which consults the thought of sequential pattern discovery in bioinformatics to solve the problem of frequent pattern discovery for multiple attributes uncertain data stream. A new method for update calculation of uncertain sliding window is designed based on mixed type model, SAX (Symbolic Aggregate approXimation) symbolic strategy is improved, and similarity analysis method for multiple attributes motifs under different sliding windows is put forward. The proposed algorithm is verified to be correct functionally by a set of uncertain data stream in the wireless sensor network of air and missile defense. Its accuracy is measured through planting different number of motifs. Furthermore, comparison with previous algorithm with tuples' valid probability set to 1 shows that the proposed algorithm can discover frequent pattern for multiple attributes uncertain data stream precisely.

Key words: Data mining; Motif discovery; Uncertain data stream; Multiple Expectation-maximization for Motif Elicitation (MEME) algorithm

1 引言

多属性不确定数据流的频繁模式发现问题可以归结为数据挖掘领域中的关联规则挖掘问题，其实质是发现动态多属性数据中重复出现的相似模式。作为关联规则挖掘中的关键问题，多属性不确定数据流中频繁模式的发现准确率会受到其动态性、不确定性以及属性之间关联性的影响，已成为数据挖掘领域中的一个研究热点和难点[1–3]。

针对不确定数据流的频繁模式挖掘，Leung等人[4]提出了SUF-growth(mine Frequent itemsets from Streams on Uncertain data)算法，基于该算法的框架，出现了较多的改进算法，如基于滑动窗口的MFCIFUDS算法[5]、基于Hyper-structure的UHS-stream(Uncertain Hyper-Structure stream mining)算法[6]、基于衰减窗口的TUF-streaming(use the Time-fading model in an Uncertain data environment to mine Frequent patterns from streaming data)算法和基于界标窗口的XTUF-streaming(eXtended TUF-streaming)算法[7]等，但这些算法只适应于单属性的不确定数据流。在实际数据流的产生过程中，不仅数据是动态的，且可能

收稿日期：2016-03-17；改回日期：2016-08-16；网络出版：2016-09-30
*通信作者：王菊 yonglingjuke@126.com
基金项目：国家自然科学基金(61272011)
Foundation Item: The National Natural Science Foundation of China (61272011)

会有多个属性同时到达，而关于如何挖掘这类多属性不确定数据流的研究还相对较少。

模体发现属于数据挖掘中的序列模式发现[8]，来源于生物信息学当中，也称为转录因子结合位点识别，用于寻找在序列当中具有功能和排列相似的短核苷酸片段[9]，主要算法有随机投影(Random Projection)，EM(Expectation Maximization)，MEME 和 Voting 等[10]。借鉴生物信息学中模体发现的思想，Lin 等人[11]在 2002 年首次提出了时间序列模体发现的概念，其他专家学者又相继提出了一种线性时间下的随机投影方法[12]、致密球[13]、MK 算法[14]和 MOEN 算法[15]等。

本文把该思路做了进一步的拓展，将模体发现的概念应用于多属性不确定数据流的频繁模式发现中，提出了一种基于 MEME 的多属性不确定数据流模体发现算法，在传统时间序列模体发现的基础上，增加了处理不确定性、动态性和多属性模体相似性分析等功能。该算法设计了基于混合型模型的不确定滑动窗口更新计算方法，改进了 SAX 的符号化策略，提出了不同滑动窗口下多属性模体的相似性分析方法，能够较准确地发现多属性不确定数据流中的频繁模式。

2 MEME 算法

MEME[16–18]是目前最流行的符号序列集合模体发现算法，可将模体从由背景成分和模体成分组成的混合序列中辨识出来。对于符号序列 $S_i = s_{i1}s_{i2}\cdots s_{iL}\ (i=1,2,\cdots,t)$ 组成的符号序列集合 $S=\{S_1,S_2,\cdots,S_t\}$，用 $l-\text{mer}$ 表示一个长度为 l 的模体，即 $l-\text{mer}=s_{i,j+1}s_{i,j+2}\cdots s_{i,j+l}$。MEME算法就是从符号序列集合中识别出重复出现的 $l-\text{mer}$。

文献[16]和文献[18]中详细介绍了MEME算法，其核心是定义了一个二元随机变量 $Z_{ij}(j=1,2,\cdots,L_i)$，通过计算每一个 $l-\text{mer}$ (表示一个长度为 l 的碱基片段)的似然值来寻找模体。其中，Z_{ij} 表示每一个 $l-\text{mer}$ 对应的背景成分或模体成分，即当字符 s_{ij} 表示为一个结合位点时，$Z_{ij}=1$，否则 $Z_{ij}=0$。该算法将整个序列集合的似然值表示为

$$\lg\left(p\left(X,Z\left|\theta_0,\Theta\right.\right)\right)=\sum_{i=1}^{t}\sum_{j=1}^{L_i-l+1}Z_{ij}\lg p\left(X_{ij}\left|Z_{ij},\theta_0,\Theta\right.\right) \quad (1)$$

式中，X_{ij} 表示第 i 行第 j 个 $l-\text{mer}$；θ_0 表示背景分布(文中采用零阶马尔科夫模型，即假设每一个字符各自独立的分布)；$\Theta=(\theta_{1h},\theta_{2h},\cdots,\theta_{zh},\cdots,\theta_{lh}),(h\in\Omega)$ 表示模体分布；θ_{zh} 表示字符 h 在模体第 z 个位置出现的概率。期望最大化算法正是通过更新潜在的隐变量 Z 使得似然值最大化，主要过程分为E步和M步。

E步的具体表达式为

$$Z_{ij}^{(T)}=\frac{p\left(X\left|Z_{ij},\Theta^{(T)}\right.\right)}{\sum_{L_i-l+1}p\left(X\left|Z_{ij},\Theta^{(T)}\right.\right)} \quad (2)$$

M步的具体表达式为

$$\Theta^{(T+1)}=\arg\max E\left[\log p\left(X,Z\left|\Theta^{(T)}\right.\right)\right] \quad (3)$$

E 步和 M 步重复执行，直至收敛。

3 基于 MEME 的多属性不确定数据流模体发现算法

3.1 算法设计思路

基于MEME的多属性不确定数据流模体发现的核心有两点：一是如何将具有时序性、海量性、无界性、实时性、高速性和连续性等特点的数据序列合理地转换为符号序列集合；二是如何从不同属性的模体中发现潜在的规律。

基于此，本文采用混合型模型对不确定数据流进行表示，利用滑动窗口对有效时间内的不确定数据流进行划分，对满足一定条件的数据进行符号化，执行MEME算法进行模体发现，最后再对不同滑动窗口下多属性模体的相似性进行分析，其思路如图1所示。

3.2 多属性不确定数据流表示

采用混合型模型可以对多属性不确定数据流进行表示，本文又进一步地定义了元组有效概率，可以简化用混合型模型所表示的多属性不确定数据流，达到数据校验和快速计算的目的。

3.2.1 混合型模型 文献[19]中所提出的混合型模型(mixed-type model)综合考虑了存在级不确定性和属性级不确定性，不仅能处理离散型属性，还可以处理连续型属性，是一种较为通用的不确定数据流模型，其定义如下。

定义1[19] 混合型模型 给定一个元组，将其 m 个连续不确定属性用 A^x 表示，n 个离散不确定属性

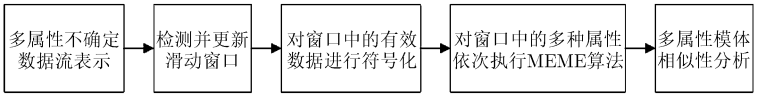


图1 算法设计思路

用 A^y 表示，其余确定属性用 A^d 表示，混合模型的分布 g 用 (p,f) 表示。其中，$p\in[0,1]$ 是元组的存在概率(Tuple Existence Probability, TEP)，f 是所有不确定属性的联合概率密度函数，定义为 $f(x,y)=f_{A^x|A^y}(x|y)\cdot P[A^y=y]$。因此，$g$ 可以用一个定义在 $R^m\times Q^n\times A^d\cup\{\perp\}$（$\perp$ 代表该元组不存在的情形）上的随机向量 (X,Y,Z) 表示，即

$$\left.\begin{aligned}&P[(X,Y,Z)=\perp]=(1-p)\\&P\left[X\subseteq I,Y=y,Z=A^d\right]=p\cdot\int_I f(x,y)\,\mathrm{d}x\end{aligned}\right\}\quad(4)$$

式中，$I\subseteq R^m$，$y\in Q^n$。

上述定义是基于元组之间的独立性假设，连续分布函数 f 采用高斯混合模型进行近似[20]。因此，假设各元组可以在离散域 B 中取多个规范化值，各元组在规范化连续区间 $[a,b]$ 中服从概率密度函数 f，当采用混合型模型时，一个具有离散属性和连续属性的不确定数据流 U 可以被描述为

$$\begin{aligned}&\left(\left\langle t_1,\left(i_1,p_{i_1}\right),\left([a_1,b_1],f\right),p_1\right\rangle,\cdots,\right.\\&\left.\left\langle t_v,\left(i_v,p_{i_v}\right),\left([a_v,b_v],f\right),p_v\right\rangle,\cdots\right)\end{aligned}\quad(5)$$

式中，对于任意 $v\geq1$ 的整数，t_v 是元组到达的时刻，p_v 表示元组的存在概率，p_{i_v} 表示属性 i_v 的出现概率，且满足 $i_v\in B$ 和 $[a_v,b_v]\subseteq[a,b]$。

3.2.2 元组有效概率　基于混合型模型，本文定义了元组的有效概率，该概率不仅能够综合属性出现概率和元组存在概率的影响，起到数据校验的效果，还能简化多属性不确定数据流的表示形式，提高计算效率。

定义 2　元组有效概率　由定义1可知，对于任意元组 u_v，它的属性出现概率为 $q(u_v)=\int_I f\left(x_{u_v},y_{u_v}\right)\mathrm{d}x_{u_v}$，元组存在概率为 $p(u_v)$，因此该元组的有效概率被定义为 $Pr(u_v)=p(u_v)\cdot q(u_v)$。

元组有效概率的含义可以理解为：如果 $Pr(u_v)=0$，即使 $p(u_v)\neq0$，但考虑到该元组的属性值已超出了其定义域，使得 $q(u_v)=0$，该元组仍被认为是无效元组。此时，不确定数据流的描述可以简化为

$$\left(\left\langle t_1,i_1,[a_1,b_1],Pr(u_1)\right\rangle,\cdots,\left\langle t_v,i_v,[a_v,b_v],Pr(u_v)\right\rangle,\cdots\right)\quad(6)$$

3.3 基于混合型模型的不确定滑动窗口

多属性不确定数据流具有无限性，因此要处理从数据出现直至当前时刻的所有数据是不可能的。针对这个问题，本文设计了一种基于混合型模型的不确定滑动窗口更新计算方法，保证滑动窗口中有效元组的数目依置信概率 α 为 w 个。通常情况下，有效元组数目 w 由所处理不确定数据流的稀疏程度以及计算机的处理能力决定，置信概率 α 由所处理不确定数据流的用途和用户的偏好决定。

3.3.1 不确定滑动窗口的定义及更新　在一个不确定数据流 U 中，w' 表示滑动窗口 $W(U,w')$ 的大小，$|\widehat{W}(U,w')|$ 表示该窗口中有效元组的数目，不确定滑动窗口的定义和更新过程如下。

定义 3[21]　不确定滑动窗口　由于滑动窗口的实际长度 w' 具有不确定性，因此该窗口被定义为不确定滑动窗口，其满足

$$\left.\begin{aligned}&\min\quad w'\\&\text{s.t.}\quad P\left(\left|\widehat{W}(U,w')\right|\geq w\right)\geq\alpha\end{aligned}\right\}\quad(7)$$

当新元组到来时，更新滑动窗口的计算流程如表1所示。

表 1　不确定滑动窗口更新过程

输入：U,w,α
输出：$W(U,w',\alpha)$
$W(U,w',\alpha)$ 被赋予空集
Loop
If (U 中的新元组 u 到达) then
$W(U,w',\alpha)\leftarrow W(U,w',\alpha)\cup\{u\}$
While $P\left(\left|\widehat{W}(U,w'-1)\right|\geq w\right)\geq\alpha$ do
$W(U,w',\alpha)\leftarrow W(U,w',\alpha)\setminus\{u'\}$
(u' 是窗口 $W(U,w',\alpha)$ 中最早的元素)
End while
End if
End loop

3.3.2 基于混合型模型的不确定滑动窗口更新计算

根据定义 2 和定义 3，本文结合滑动窗口的更新过程和于混合型模型，对约束条件 $P\left(\left|\widehat{W}(U,w')\right|\geq w\right)\geq\alpha$ 进行了变形和简化。

对于任意元组 u_i，若随机变量 $X_i=0$，则表示 u_i 无效，无效概率为 $1-Pr(u_i)$；若 $X_i=1$，则表示 u_i 有效，有效概率为 $Pr(u_i)$，即 X_i 服从以 $Pr(u_i)$ 为参数的 $(0-1)$ 分布，因而有

$$P\left(\left|\widehat{W}(U,w')\right|\geq w\right)\geq\alpha\Leftrightarrow P\left(\sum_{i=1}^{w'}X_i\geq w\right)\geq\alpha\quad(8)$$

设 $Y=\sum_{i=1}^{w'}X_i$，则 $P\left(\sum_{i=1}^{w'}X_i\geq w\right)\geq\alpha=P(Y\geq w)\geq\alpha$，由于 $P(Y=w)=\sum_{A\in F_w}\prod_{u_i\in A}Pr(u_i)\cdot\prod_{u_i\notin A}\left(1-Pr(u_i)\right)$，可得

$$P(Y\geq w)=1-\sum_{i=0}^{w-1}P(Y=i)\quad(9)$$

式中，F_w 代表在以 w' 为长度的滑动窗口中出现 w 个

有效元组的所有可能组合；A 是 F_w 中的一种组合方式。文献[22]分别利用泊松近似、高斯近似以及修正高斯近似对式(9)的运算进行了简化，本文采用均值为 $\mu = E[Y] = \sum_{i=1}^{w'} Pr(u_i)$，标准差为 $\sigma = \sqrt{\mathrm{Var}(Y)} = \sum_{i=1}^{w'} Pr(u_i)\cdot\left(1 - Pr(u_i)\right)$ 的高斯近似来计算该概率，即

$$P(Y \ge w) \approx 1 - \Phi\left(\frac{w + 0.5 - \mu}{\sigma}\right) \tag{10}$$

3.4 改进的 SAX 符号化策略

SAX[23,24]是一种针对一元时间序列符号化的方法，可以用于聚类、分类、索引和异常检测等。为了将该方法应用于不确定数据流，确保每个滑动窗口中存在的元组数目差别不大，使每一种属性(包括连续属性和离散属性)都能用 SAX 进行符号化，本文将 SAX 的符号化策略改进为以下 4 个步骤：

步骤 1　在长度为 w' 的滑动窗口中，以置信概率 α 抽取该窗口中的 w 个有效元组；

步骤 2　若为离散属性，用向量 $\boldsymbol{D} = d_1, d_2, \cdots, d_w$ 对有效元组中的该属性值进行表示；若为连续属性，用向量 $\boldsymbol{D} = d_1, d_2, \cdots, d_w$ 对有效元组中该属性的期望值进行表示；

步骤 3　PAA(Piecewise Aggregate Approximation)过程，即用一小段序列的平均值代表该序列；

步骤 4　符号化，即用不同的符号表示每小段的平均值。

步骤1的目的是确保每个滑动窗口中存在元组的数目近似为 w，步骤2是将连续属性用期望值进行表示，为后续计算做准备，文献[24]中给出了PAA的计算过程、常用的分界点值以及所需字符集的数目。

3.5 不同滑动窗口下多属性模体的相似性分析

为了发现具有相似性的多属性模体，本文提出了一种不同滑动窗口下多属性模体(如图2所示)的相似性分析方法，该方法可以为具有多属性的不确定数据流预测、分类、异常检测和规则挖掘等提供依据，其具体处理流程如下。

步骤 1　列出窗口中不同属性间所有模体的组合，如图 2 中的窗口 1，其所有模体的组合为 $(M_{11}, M_{21}, \cdots, M_{k1})^{\mathrm{T}}$ 和 $(M_{12}, M_{21}, \cdots, M_{k1})^{\mathrm{T}}$；

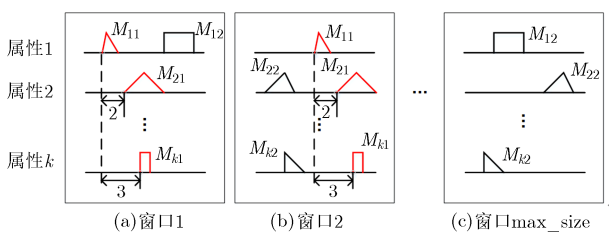


图2 不同滑动窗口下多属性模体示意图

步骤 2　采用 Hamming 距离对所有的模体组合进行相似性匹配，记录满足阈值条件的模体组合(阈值根据实际需要进行设定，本文中设定为 2)，图 2 中得到的模体组合为 $(M_{11}, M_{21}, \cdots, M_{k1})^{\mathrm{T}}$；

步骤 3　计算每一个满足阈值条件模体组合所对应的若干个多属性模体间的相对时间向量。将第 1 个模体的起始点设置为 0，向量中的值表示模体起始点相对于 0 点的时间，图 2 中模体组合 $(M_{11}, M_{21}, \cdots, M_{k1})^{\mathrm{T}}$ 所对应的两个多属性模体的相对时间向量都为 $(0,2,3)^{\mathrm{T}}$；

步骤 4　计算每一个满足阈值条件模体组合所对应的若干个多属性模体中任意两个相对时间向量之间的欧氏距离。当该距离小于给定 ε (根据实际的精度要求进行设定，本文中设定 ε 为0.01)时，其相应的多属性模体具有相似性，输出并记录该多属性模体。

3.6 算法设计

根据以上的分析，文中提出了一种基于MEME的多属性不确定数据流模体发现算法，该算法可以对任意一组多属性不确定数据流进行模体发现，其总体描述如表2所示。表中，步骤4和步骤5至步骤8可以同时运算，步骤3-步骤8的具体实现方法已在前文中进行过介绍。

4 实验与分析

4.1 案例分析

防空反导情报传感器网络是产生不确定数据流的典型应用，本文根据该网络对飞机速度和发动机温度的实时测量数据进行模体发现。其实验平台是装有 64 位 Windows7 操作系统的个人电脑，具有 Core i5-4590 的双核处理器和 4GB 内存，可运行 Matlab 和 MEME 程序包。

(1)针对本文中所处理的不确定数据流，可设置窗口中有效元组数目 $w = 200$，置信概率 $\alpha = 0.98$，$\max\_size = 3$，$n = 0$，$k = 2$。

(2)判断不确定数据流是否实时到来，若数据流中断，则算法终止，否则转下一步继续执行计算。

(3)通过混合型模型，对不确定数据流进行建模，利用 $(x - \mu)/\sigma$ 对其规范化，其中温度是连续属性，规范化后概率密度函数 $\varphi \sim N(0,1)$，计算元组有效概率，其部分数据如表3所示，灰色表示无效元组。

(4)继续采集诸如表3中的数据，根据元组的有

表 2 基于 MEME 的多属性不确定数据流模体发现算法

输入：$U, w, \alpha, \max\_size, k$ // max_size 为最多可同时处理的滑动窗口数目，k 为属性个数。

输出：具有相似性的多属性模体。

步骤 1　初始化参数 $w, \alpha, \max\_size$，设置变量 $n=0$；

步骤 2　判断不确定数据流是否到来，是转步骤 3，否则算法结束；

步骤 3　对 U 建模、规范化及有效概率计算，得到

$U'=\{u_1', u_2', \cdots\}$

$=\{\langle t_1, i_1, [a_1, b_1], Pr(u_1)\rangle, \langle t_2, i_2, [a_2, b_2], Pr(u_2)\rangle, \cdots\}$

步骤 4　Loop

If (U' 中的新元组 u' 到达) then

$W(U', w', \alpha) \leftarrow W(U', w', \alpha) \cup \{u'\}$；

If $P\left(\left|\widehat{W}(U', w'-1)\right| \geq w\right) \geq \alpha$ then

$S_n = \{u_1', u_2', \cdots, u_{w'}'\}$；

$W(U', w', \alpha)$ 被赋予空集；

$n = n+1$；

If $n = \max\_size$ then

$n = 0$；

另存 $S_1, S_2, \cdots, S_{\max\_size}$ 为 $S_1', S_2', \cdots, S_{\max\_size}'$，

转步骤 5；

Else

Continue;

End if

Else

Continue;

End if

End if

步骤 5　对序列集 $S_1', S_2', \cdots, S_{\max\_size}'$ 进行符号化，得到 max_size 个 $k \times w$ 的字符矩阵；

步骤 6　对字符矩阵同一行中的字符串执行 MEME 算法，存储并输出所发现的模体

$M_{11}, M_{12}, \cdots, M_{21}, M_{22}, \cdots, \cdots, M_{k1}, M_{k2}, \cdots$

步骤 7　释放 $S_1', S_2', \cdots, S_{\max\_size}'$ 所占用的存储空间；

步骤 8　对不同滑动窗口下所发现的多属性模体进行相似性分析，存储具有相似性的多属性模体。

End loop

表 3 部分规范化不确定数据流示例

时间	飞机速度	发动机温度距离	存在概率	有效概率
t_1	(0.84,0.8)	$([-0.3,1.2], \varphi)$	0.9	0.362
t_2	(2.43,0.7)	$([-1.2,0.3], \varphi)$	0.8	0.282
t_3	(−0.35,0.8)	$([-2.2,0.9], \varphi)$	0.9	0.577
t_4	(1.38,0.9)	$([-2.1,0.3], \varphi)$	0.6	0.324
t_5	(−0.86,0.8)	$([0.7,1.8], \varphi)$	0.7	0.115
t_6	(0.21,0.9)	$([-0.4,0.6], \varphi)$	0.9	0.309
t_7	(−21,0)	$([0.8,1.9], \varphi)$	0.9	0
t_8	(0.34,0.8)	$([0.1,1.4], \varphi)$	0.8	0.243
t_9	(−0.49,0.8)	$([-0.2,0.9], \varphi)$	0	0
…	…	…	…	…

效概率来判断是否满足 $P\left(\left|\widehat{W}(U', w'-1)\right| \geq w\right) \geq \alpha$，当 $n = \max\_size$，得到 $S_1', S_2', \cdots, S_{\max\_size}'$ 后，设置 $n=0$，不断重复步骤 3。

(5)当连续属性取期望值后，对 $S_1', S_2', \cdots, S_{\max\_size}'$ 进行符号化，字符集采用 $\{A, G, C, T\}$，具体见 3.4 节的分析。在符号化过程中，需要根据文献[24]中分界点值的划设规范，确立 3 个分界点：−0.67, 0 和 0.67，设置维度 $c=4$，进而得到符号序列集合如表 4 所示。

表 4　S_1', S_2', S_3' 的符号化结果

速度属性	seq1	TGCATTGCTCACGCCCGCGACGCGCA
		GGCCGCGTCGGAACACGCCCTGCG
	seq2	CGAGGGCGACTAGGGGGAACGCCGAC
		GGGGGGCGCCGCGGTCCGAGGGGC
	seq3	GGCTGGTACCACGGGCTGGCGGCGGT
		ACCGTTGGGGCCCGTGGAGAAGGG
温度属性	seq1	GGGTGGCCTGAGGTTCGGGCCCCCCC
		GGCAGACCCGCGGCAGCCACGACC
	seq2	GACGCATTCGCGCCTCCGGCGCCTGG
		ACAAGACCCGGCTGCAGGCGCTAC
	seq3	CCGATCGGCCCGGCCGCCGGACGGGC
		CGCGCTCGATCCTTCACCGGACAA

(6)根据MEME算法，采用MEME程序包，对表4中的符号序列集合进行模体发现，其结果如图3所示。

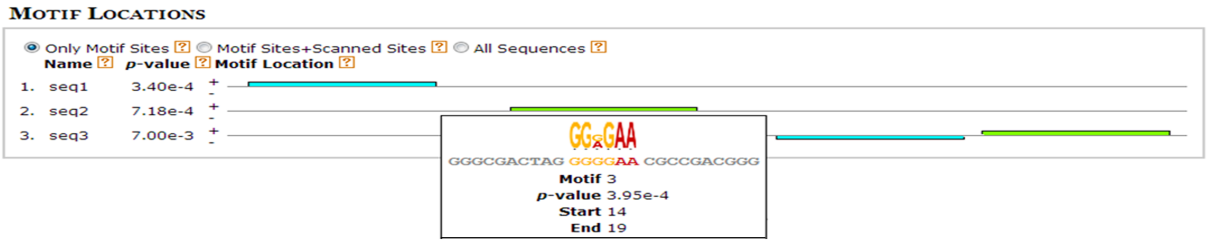
由图 3 可知，在表 4 的两种属性中，共发现了 4 种模体。进而根据所发现模体的排序及位置(如图 3 中方框内所示)，可以反推出原数据中模体的位置及形状，如图 4 所示。

从图 4 可发现，所发现模体在形状上具有一定的相似性，满足“模体”在时间序列中的定义，证明文中所提出的算法具有发现多属性不确定数据流中频繁模式的功能。

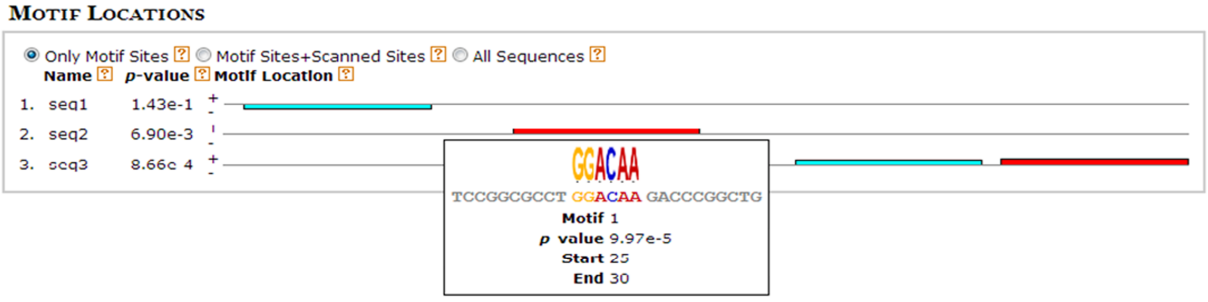
(7)MEME 算法执行完毕后，释放出 S_1', S_2', S_3' 所占用的存储空间，并存储所发现的模体。

(8)根据图 3 和图 4，可得到如图 5 所示的示意图。

根据 3.5 节的分析，结合图 5 可知，$(M_{11}, M_{21})^{\mathrm{T}}$ 和 $(M_{12}, M_{22})^{\mathrm{T}}$ 是满足阈值条件的模体组合，其对应的多属性模体具有相似性。从而可以得出：飞机的高速机动会引起其发动机温度的变化，且根据模体组合 $(M_{11}, M_{21})^{\mathrm{T}}$ 在窗口 1 中的多属性模体，可以预测窗口 3 中其相应多属性模体温度属性的变化。

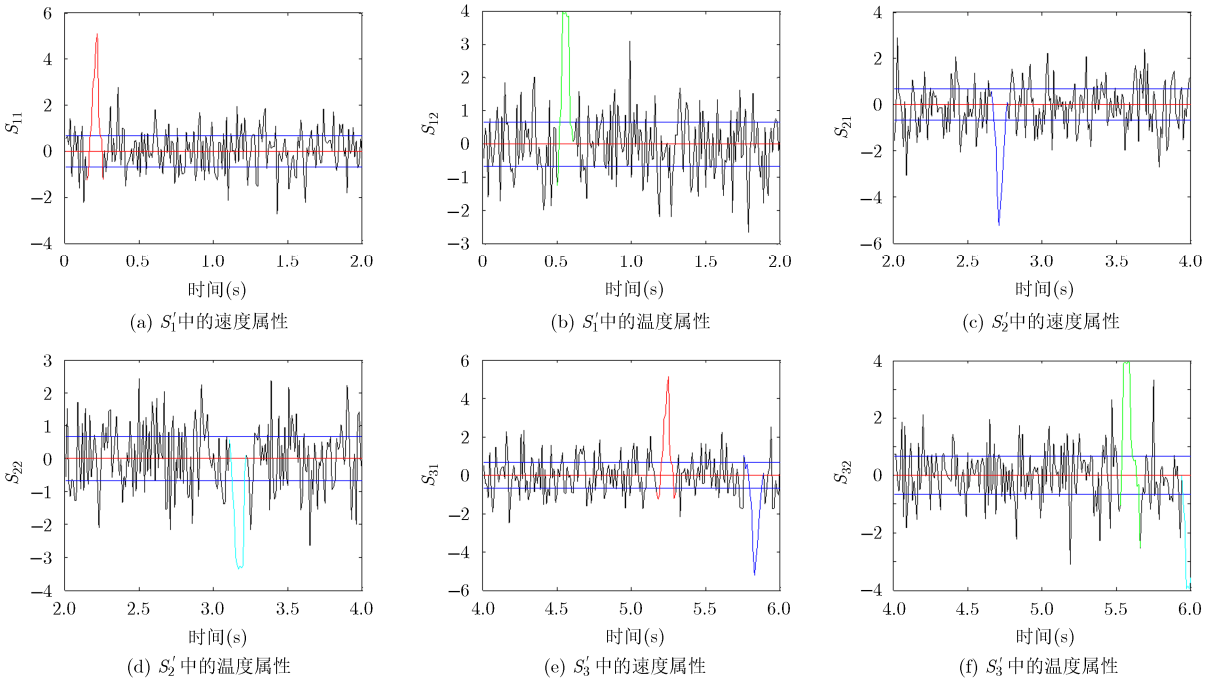


(a)速度属性中所发现的模体



(b)温度属性中所发现的模体

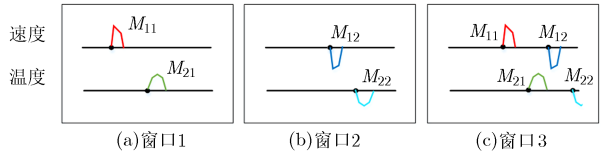
图 3 模体发现结果



(a) S_1'中的速度属性 (b) S_1'中的温度属性 (c) S_2'中的速度属性

(d) S_2'中的温度属性 (e) S_3'中的速度属性 (f) S_3'中的温度属性

图 4 模体在原数据中的位置及形状



(a)窗口1 (b)窗口2 (c)窗口3

图 5 不同滑动窗口下所发现模体的示意图

4.2 实验分析

通过 4.1 节的案例，验证了文中算法的功能。为进一步分析该算法的性能，本节设计了两个实验，分别用于测试其模体发现准确率和将其与已有算法进行比较。

实验 1 对不确定数据流进行模体植入，当植入模体的数目依次为 $x=1,10,20,\cdots,200$ 时，本文算法所发现模体的准确率如图 6 所示。

从图 6 可知，在植入不同数目模体的情况下，本文算法的模体发现准确率在 0.8~0.9 之间，稳定性强。

实验 2 考虑到目前还没有关于多属性不确定

数据流模体发现的相关算法，因此文中需要设定不确定数据流中所有属性的出现概率和元组存在概率都为 1。进而通过设置滑动窗口长度 $w=200$，属性个数 $k=1$，滑动窗口不更新，将文中算法与 Mueen 提出的 MK[14]和 MOEN[15]算法在随机植入模体的 3 组白噪声数据集上进行比较，其模体发现准确率结果如图 7 所示。

从图 7 可知，文中算法的模体发现准确率高于 MK 和 MOEN 算法。

5　结束语

本文在传统时间序列模体发现的基础上引入了不确定性和动态性，建立了序列数据挖掘和不确定数据流挖掘之间的桥梁，并采用生物信息学算法完成了不确定数据流的模体发现。主要工作有：(1)提出了基于 MEME 的多属性不确定数据流模体发现算法，根据防空反导传感器网络对飞机速度和发动机温度的实时测量数据进行了模体发现，验证了其功能；(2)通过多次模体植入实验和算法性能对比实验，在同等仿真条件下，相比于 MK 和 MOEN 算法，验证了其优越性。文中部分内容属于探索性的研究，不确定性理论、智能搜索算法与不确定数据流模体发现的结合将是本文下一步的研究内容。

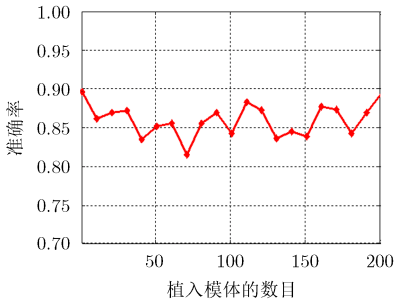


图 6　本文算法所发现模体的准确率

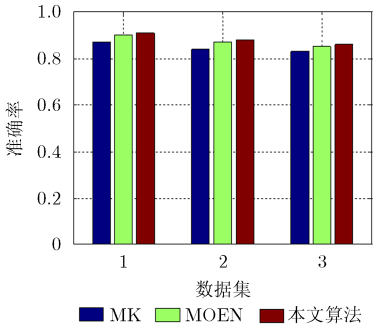


图 7　算法准确率比较

参 考 文 献

[1] LEUNG C K S, JIANG F, and HAYDUK Y. A landmark-model based system for mining frequent patterns from uncertain data streams[C]. 2011 International Database Engineering and Applications Symposium, Lisbon, Portugal, 2011: 249–250. doi: 10.1145/2076623.2076659.

[2] CHUI C K and KAO B. A decremental approach for mining frequent itemsets from uncertain data[C]. 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Osaka, Japan, 2008: 64–75. doi: 10.1007/978-3-540-68125.

[3] LEUNG C K S, HAO B, and BRAJCZUK D A. Mining uncertain data for frequent itemsets that satisfy aggregate constraints[C]. 25th Annual ACM Symposium on Applied Computing, Sierre, Switzerland, 2010: 1034–1038. doi: 10.1145/1774088.1774305.

[4] LEUNG C K S and HAO B. Mining of frequent items from streams of uncertain data[C]. 25th IEEE International Conference on Data Engineering, Piscataway, NJ, USA, 2009: 1663–1670. doi: 10.1109/ICDE.2009.157.

[5] 汤克明. 不确定数据流中频繁数据挖掘[D]. [博士论文], 南京航空航天大学, 2012.
TANG Keming. Study on frequent data mining from uncertain data streams[D]. [Ph.D. dissertation], Nanjing University of Aeronautics and Astronautics, 2012.

[6] HEWANADUNGODAGE C, YUNI X, and LEE J J. Hyper-structure mining of frequent patterns in uncertain data streams[J]. *Knowledge and Information Systems*, 2013, 37: 219–244. doi: 10.1007/s10115-012-0581-y.

[7] LEUNG C K S, CUZZOCREA A, FAN J, *et al*. Discovering frequent patterns from uncertain data streams with time-fading and landmark models[J]. *Transactions on Large-Scale Data and Knowledge-Centered Systems VIII*, 2013: 174–196. doi: 10.1007/978-3-642-37574-3_8.

[8] 朱跃龙, 彭力, 李士进, 等. 水文时间序列模体挖掘[J]. 水利学报, 2012, 43(12): 1422–1430.
ZHU Yuelong, PENG Li, LI Shijin, *et al*. Research on hydrological time series mining [J]. *Journal of Hydraulic Engineering*, 2012, 43(12): 1422–1430.

[9] 张懿璞. 转录因子结合位点识别问题的算法研究[D]. [博士论文], 西安电子科技大学, 2014.
ZHANG Yipu. Algorithm research on the problem of transcription factor binging sites identification[D]. [Ph.D. dissertation], Xidian University, 2014.

[10] 杨矫云. 大规模生物序列分析的高性能算法和模型[D]. [博士论文], 中国科学技术大学, 2014.
YANG Jiaoyun. High performance algorithms and models for large-scale biological sequence analysis[D]. [Ph.D. dissertation], University of Science and Technology of China, 2014.

[11] LIN J, KEOGH E, PATEL P, *et al*. Finding motifs in time series[C]. Proceedings of the 2nd Workshop on Temporal Data Mining at KDD, District of Colombia, USA, 2002: 53–68.

[12] CHIU B, KEOGH E, and LONARDI S. Probabilistic discovery of time series motifs[C]. 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, District of Colombia, USA, 2003: 493–498. doi: 10.1145/956750.956808.

[13] FERREIRA P G, AZEVEDO P J, SILVA C G, *et al*. Mining approximate motifs in time series[C]. 9th international conference on Discovery Science, Berlin, Germany, 2006: 89–101 .

[14] MUEEN A, KEOGH E, ZHU Q, *et al*. Exact discovery of time series motif[C]. 9th SIAM International Conference on Data Mining 2009, Nevada, USA, 2009: 469–480.

[15] ABDULLAH M and NIKAN C. Enumeration of time series motifs of all lengths[J]. *Knowledge and Information Systems*, 2015, 45: 105–132. doi: 10.1007/s10115-014-0793-4.

[16] 张懿璞, 霍红卫, 于强, 等. 用于转录因子结合位点识别的定位投影求精算法[J]. 计算机学报, 2013, 36(12): 2545–2559. doi: 10.3724/SP.J.1016.2013.02545.
ZHANG Yipu, HUO Hongwei, YU Q, *et al*. A novel fixed-position projection refinement algorithm for TFBS Identification[J]. *Chinese Journal of Computers*, 2013, 36(12): 2545–2559. doi: 10.3724/SP.J.1016.2013.02545.

[17] TIMOTHY L B. DREME: motif discovery in transcription factor ChIP-seq data[J]. *Original Paper*, 2011, 17(12): 1653–1659. doi: 10.1093/bioinformatics/btr261.

[18] DANIEL Q and XIE Xiaohui. EXTREME: an online EM algorithm for motif discovery[J]. *Original Paper*, 2014, 30(12): 1667–1673. doi: 10.1093/bioinformatics/btu093.

[19] THANH T L T, PENG Liping, DIAO Yanlei, *et al*. CLARO: modeling and processing uncertain data streams[J]. *The VLDB Journal*, 2012, 21: 651–676. doi: 10.1007/s00778-011-0261-7.

[20] ARCHAMBEAU C and VERLEYSEN M. Manifold constrained finite Gaussian mixtures [C]. 8th International Work Conference on Artificial Neural Networks, Berlin, Germany, 2005: 820–828.

[21] MICHELE D. Modeling and querying data series and data streams with uncertainty[D]. [Ph.D. dissertation], University of Trento, 2014.

[22] HONG Y. On computing the distribution function for the sum of independent and non-identical random indicators [R]. Technical Report, Department of Statistics, Virginia Tech, 2011.

[23] 曲文龙, 张克君, 杨炳儒, 等. 基于奇异事件特征聚类的时间序列符号化方法[J]. 系统工程与电子技术, 2006, 28(8): 1131–1134.
QU Wenlong, ZHANG Kejun, YANG Bingru, *et al*. Time series symbolization based on singular event feature clustering[J]. *Systems Engineering and Electronics*, 2006, 28(8): 1131–1134.

[24] JESSICA L, EAMONN K, LI W, *et al*. Experiencing SAX: a novel symbolic representation of time series[J]. *Data Minning and Knowledge Discovery*, 2007, 15: 107–144. doi: 10.1007/s10618-007- 0064-z.

王 菊： 女，1991 年生，博士生，研究方向为数据挖掘.

刘付显： 男，1962 年生，教授，研究方向为作战建模与仿真、数据挖掘.