一种基于频繁模式的时间序列分类框架

万　里[①③]　廖建新[①②]　朱晓民[①②]　倪　萍[①③]
[①](北京邮电大学网络与交换技术国家重点实验室　北京　100876)
[②](东信北邮信息技术有限公司　北京　100191)
[③](卡耐基梅隆大学　匹兹堡　15213)

摘　要：如何提取和选择时间序列的特征是时间序列分类领域两个重要的问题。该文提出 MNOE(Mining Non-Overlap Episode)算法计算时间序列中的非重叠频繁模式，并将其作为时间序列特征。基于这些非重叠频繁模式，该文提出 EGMAMC(Episode Generated Mixed memory Aggregation Markov Chain)模型描述时间序列。根据似然比检验原理，从理论上推导出频繁模式在时间序列中出现的次数和 EGMAMC 模型是否能显著描述时间序列之间的关系；根据信息增益定义，选择能显著描述时间序列的频繁模式作为时间序列特征输入分类模型。在 UCI (University of California Irvine)公共数据集和实际智能楼宇数据集上的实验表明，选择频繁模式作为特征进行分类的准确率、召回率和 F-Measure 均优于不选择频繁模式作为特征的分类结果。高效的计算和有效的选择非重叠频繁模式作为时间序列特征有助于提高时间序列分类模型的各项评价指标。

关键词：时间序列分类；频繁模式挖掘；智能楼宇
中图分类号：TP393　　**文献标识码**：A　　**文章编号**：1009-5896(2010)02-0261-06
DOI: 10.3724/SP.J.1146.2009.00135

A Frequent Pattern Based Time Series Classification Framework

Wan Li[①③]　Liao Jian-xin[①②]　Zhu Xiao-min[①②]　Ni Ping[①③]
[①](*State Key Laboratory of Networking and Switching Technology Beijing University of Posts and Telecommunications*, *Beijing* 100876, *China*)
[②](*EBUPT Information Technology Co.*, *Ltd*, *Beijing* 100191, *China*)
[③](*Carnegie Mellon University*, *Pittsburgh*, *US* 15213, *USA*)

Abstract: How to extract and select features from time series are two important topics in time series classification. In this paper, a MNOE (Mining Non-Overlap Episode) algorithm is presented to find non-overlap frequent patterns in time series and these non-overlap frequent patterns are considered as features of the time series. Based on these non-overlap episodes, an EGMAMC (Episode Generated Mixed memory Aggregation Markov Chain) model is presented to describe time series. According to the principle of likelihood ratio test, the connection between the support of episode and whether EGMAMC could describe the time series significantly is induced. Based on the definition of information gain, significant frequent patterns are selected as the features of time series for classification. The experiments on UCI (University of California Irvine) datasets and smart building datasets demonstrate that the classification model trained with selecting significant frequent patterns as features outperforms the one trained without selecting them on precision, recall and F-Measure. The time series classification models can be improved by efficiently extracting and effectively selecting non-overlap frequent patterns as features of time series.

Key words: Time series classification; Frequent pattern mining; Smart building

1　引言

给定一个数据样本集合，每个数据样本包括：一个输入时间序列 $X_i = \{\boldsymbol{x}(t) \mid t \in \{1,2,\cdots,T\}\}$ 及其离散的分类标签 C_s ，其中，$\boldsymbol{x}(t) \in R^n$ 是一个n维向量，称作 t 时刻发生的事件，$C_s \in \{1,2,\cdots,S\}$ 。时间序列分类的目标是预测新给出的时间序列 X_j 的类标签。时间序列分类技术在通信[1]、生物信息[2]、自动控制[3]等领域已有广泛应用，但通常情况下时间序列的长度不相等，即使所有待分类时间序列长度相

2009-02-02 收到，2009-09-03 改回
国家杰出青年科学基金(60525110)，国家 973 计划项目(2007CB307100，2007CB307103)和电子信息产业发展基金项目(基于 3G 的移动业务应用系统)资助课题
通信作者：万里　wanly@ebupt.com

等，不同序列相同时刻的事件不一定可比，直接套用一般的分类算法，如SVM，k-近邻搜索[4]等，效果不一定好。因此，特征提取和选择是研究时间序列分类的重要课题。

本文主要研究如何从离散时间序列中提取并选择频繁模式(frequent pattern)作为分类特征(连续序列可用文献[5]提出的方法转换为离散序列)。现有的基于频繁模式的分类算法[6–8]大多利用频繁模式生成基于关联规则的分类模型，文献[6]利用信息增益建立了根据频繁模式支持度选择分类属性的框架。这些方法存在两个问题：(1)没有考虑频繁模式在时间序列中的分布。(2)没有系统的讨论如何根据频繁模式在时间序列中出现的次数(支持度)选择其作为分类属性。

本文主要贡献如下：提出一种基于非重叠频繁模式(Non-overlap Episode)的时间序列分类框架：(1)提出非重叠频繁模式挖掘算法，基于此种模式提出 EGMAMC 模型(Episode Generated Mixed memory Aggregation Markov Chain)；(2)根据似然比检验和信息增益原理，提出利用非重叠频繁模式支持度进行特征选择的理论框架。(3)在公共数据集和私有数据集上的实验表明，基于非重叠频繁模式的时间序列分类方法的分类结果优于传统分类算法。

2 基于非重叠频繁模式的 EGMAMC

2.1 非重叠频繁模式挖掘算法

文献[9]首次提出非重叠频繁模式的概念：某个频繁模式在时间序列中出现两次，一个实例中的事件不在另一个实例的两个事件之间出现。非重叠频繁模式的支持度是非重叠实例在时间序列中出现的最大次数。然而文献[9]并没给出直接计算非重叠频繁模式的方法，因此本文提出 MNOE(Mining Non-Overlap Episode)算法直接计算非重叠频繁模式。

MNOE 算法如下：

输入：(1)带时间戳的时间序列投影 $P<\text{head, body}>$ 的集合 $S=\{P_1,P_2,\cdots,P_n\}$

(2) 当前迭代频繁模式长度：l

(3) 频繁模式中两个连续事件间允许的最大时间间隔：maxGap

(4) 最小支持度：$\text{Spt}_{\min}$

输出：非重叠频繁模式集合

MNOE 是递归算法，具体步骤如图 1 所示

```
1  将 S 中的 P_i 按 P_i.head 中最后一个元素的时间戳进行升序排列。
2  for (S 中每个投影 P_i ){
3      For  P_i.body 中每个元素 e
4          If ( e.time_stamp 减去 P_i.head 中最后一个元素的时间戳大于 maxGap)
5              结束循环
6          If ( e 是 P_i.body 中属于|e|事件类型且时间戳最小的元素且 P_i.head 中第一个元素的时间戳大于 HashtableFE(|e|)对应集合中最后一个元素的时间戳)
7              |e|的支持度增加 1。
8      P_i' =projection( e , P_i )
9      S'.add(P_i')
10     For HashtableFE 中每个键值|e|
11         if  HashtableFE(|e|).size 大于 Spt_min
12             调用函数 MNOE( S' , l+1 , max Gap , Spt_min )
```

图 1 MNOE 算法步骤

$|\text{e}|$ 表示一个事件类型，$>e=<|\text{e}|$，time_stamp 表示 $|\text{e}|$ 类型事件在时间序列中 time_stamp 时刻出现的一个实例。P_i.head 表示当前迭代步骤所计算频繁模式的前缀，P_i.head 表示该前缀的一个实例，P_i.body 表示 P_i.head 中最后一个时刻事件以后到时间序列结束时刻的子时间序列。算法每次迭代的第 1 步对 $S=\{P_1,P_2,\cdots,P_n\}$ 的排序保证所得非重叠频繁模式在时间序列中出现实例达到最大值。HashtableF 是以|e|为键值的哈希表，HashtableFE (|e|)为存放现有前缀(P_i.head)后紧跟事件|e|的实例的集合。projection(e)中的投影规则：新的投影为 P_i'，P_i'.head $=P_i$.head $+e$，P_i'.body 为 P_i.body 中所有时间戳大于 e 的时间戳的事件实例。S 的初始值即给定的整个时间序列，l 初始值为 0。

2.2 EGMAMC 模型

基于一个频繁模式 α 可定义一个 EGMAMC 模型。EGMAMC 模型结合 Aggregate Markov(AM)和 Mixed Memory Markov(MMM)模型，将状态空间分为 K 类，并假设当前时刻状态在一定概率下受 l 个时刻前状态影响：

设时间序列 $X=\{x_1,x_2,\cdots,x_t,\cdots,x_T\}$，$t\in\{1,\cdots,T\}$，$x_t\in E=\{e_1,e_2,\cdots,e_n\}$ 为当前时刻状态，E 为状态空间。本文将 EGMAMC 模型状态空间划分为两类：频繁模式状态(K_e)和噪声状态(K_n)。频繁模式状态代表时间序列中在频繁模式 $\alpha=\{S_e^1,S_e^2,\cdots,S_e^{N-1},S_e^N\}$ (S_e^i 表示在 α 中第 i 个位置出现的事件)内出现的事件；噪声状态则代表不在 α 内出现的事件。若规定状态间相互影响的最大时间间隔为 L，

则基于 α 的 EGMAMC 模型 Λ_α 为

$$\begin{aligned}&P(x_t \mid x_{t-1},\cdots,x_{t-L})\\&=\sum_{l=1}^{L}P(l \mid x_{t-1},\cdots,x_{t-L})P^l(x_t \mid x_{t-l})\\&=\sum_{l=1}^{L}P(l \mid x_{t-1},\cdots,x_{t-L})\sum_{k=1}^{K}P(x_t \mid k,x_{t-l})P^l(k \mid x_{t-l})\end{aligned} \tag{1}$$

$P(l \mid x_{t-1},\cdots,x_{t-L})$，$P(x_t \mid k,x_{t-l})$，$P^l(k \mid x_{t-l})$ 定义如下：$\exists x_{t-l} \in K_e$ 且 $\forall l' < l, x_{t-l'} \in K_n$ 时，$P(l \mid x_{t-1}, \cdots, x_{t-L}) = 1$ 否则 $P(l=1 \mid x_{t-1} \in K_n, \cdots, x_{t-L} \in K_n) = 1$。$P^l(k \mid x_{t-l})$ 表示当前时刻状态类型受 l 个时刻前状态影响，令 $P^l(k = K_n \mid x_{t-l}) = \eta$。当 $k = K_e$ 时，如果 $x_{t-l} = S_e^{n-1}$，则 $P(x_t = S_e^n \mid K_e, x_{t-l})=1$，$P(x_t \neq S_e^n \mid K_e, x_{t-l}) = 0$。当 $k = K_n$ 时，$P(x_t \mid K_n, x_{t-l}) = 1/M$（$M$ 为时间序列中所有事件类型个数）。

令当前时刻前 L 个时刻状态均为噪声状态的概率，即 $P(x_{t-1} \in K_n, \cdots, x_{t-L} \in K_n) = \lambda$。

综上，模型 Λ_α（式(1)）生成给定时间序列 o 的概率为

$$\begin{aligned}P(o \mid \Lambda_\alpha) &= \left(\frac{\lambda\eta}{M}\right)^{q_n^n(o)}(\lambda(1-\eta))^{q_n^e(o)}\left(\frac{(1-\lambda)\eta}{M}\right)^{q_e^n(o)}\\&\cdot((1-\lambda)(1-\eta))^{q_e^{e\prime}(o)} = \left(\frac{\lambda}{1-\lambda}\right)^{q_n^n(o)+q_n^e(o)}\\&\cdot\left(\frac{(1-\lambda)\eta}{M}\right)^{q_n^n(o)+q_n^e(o)+q_e^n(o)+q_e^{e\prime}(o)}\left(\frac{M(1-\eta)}{\eta}\right)^{q_n^e(o)+q_e^{e\prime}(o)}\\&=\left(\frac{\lambda}{1-\lambda}\right)^{\lambda T}\left(\frac{(1-\lambda)\eta}{M}\right)^{\mathrm{T}}\left(\frac{M(1-\eta)}{\eta}\right)^{q_e^e(o)}\\&=\left(\frac{\lambda^\lambda(1-\lambda)^{(1-\lambda)}\eta}{M}\right)^{\mathrm{T}}\left(\frac{M(1-\eta)}{\eta}\right)^{q_e^e(o)}\end{aligned} \tag{2}$$

其中 $T = |o|$ 表示时间序列长度，$q_n^n(o)$ 为噪声状态转移到噪声状态的次数，$q_n^e(o)$ 为噪声状态转移到 α 第 1 个事件状态（S_e^1）的次数，$q_e^n(o)$ 为频繁模式状态转移到噪声状态的次数，$q_e^{e\prime}(o)$ 为频繁模式状态转移到“部分”频繁模式状态的次数（S_e^1 除外），$q_e^e(o)$ 为频繁模式状态在 o 出现的次数，且 $\lambda = [q_n^n(o) + q_n^e(o)]/T$。若 α 在 o 中出现次数为 f_α，则有 $Nf_\alpha \le q_e^e(o) < N(f_\alpha + 1)$。因为如果 $q_e^e(o)$ 出现次数超过 $N(f_\alpha + 1)$，则得到 α 出现次数为 $f_\alpha + 1$，和前提条件矛盾。通常情况下 $N \ll T$，所以 $q_e^e(o) = Nf_\alpha$。

$$P(o \mid \Lambda_\alpha) = \left(\frac{\lambda^\lambda(1-\lambda)^{(1-\lambda)}\eta}{M}\right)^{\mathrm{T}}\left(\frac{M(1-\eta)}{\eta}\right)^{Nf_\alpha} \tag{3}$$

在式(3)中，若 λ, η 一定，$[M(1-\eta)/\eta] > 1$，其中 $\eta < M/(M+1)$，则 f_α 越大，$P(o \mid \Lambda_\alpha)$ 越大。当 $\lambda = 1/2$，$\lambda^\lambda(1-\lambda)^{(1-\lambda)}$ 取得最小值 1/2，其它参数一定时，在区间 $(0,1/2)$ 上 $P(o \mid \Lambda_\alpha)$ 与 λ 负相关。η 越大，α 出现的概率越小，λ 越大 α 的分布越稀疏。

2.3 显著频繁模式

本文用似然比检验衡量 EGMAMC 模型 Λ_α 是否比独立同分布模型更适合作为给定时间序列生成模型。如果 Λ_α 适合生成时间序列 o 的假设成立，称 α 为 o 中的显著频繁模式。

假设 H_1：时间序列 o 由 Λ_α 生成；H_0：时间序列 o 由独立同分布模型 Λ_{iid} 生成。似然比检验将拒绝假设 H_0（接受 H_1），如果

$$L(o) = \frac{P(o \mid H_1)}{P(o \mid H_0)} > \gamma > 0 \text{（I类误差决定 } \gamma \text{ 大小）} \tag{4}$$

由式(3)，当 $M/(M+1) < \eta < 1$ 或 $1/2 < \lambda < 1$ 时，o 中噪声占主要比例，H_1 和 H_0 等价。以下讨论均假设 $\eta < M/(M+1)$ 且 $0 < \lambda < 1/2$。

因为 o 在假设 H_0 下的似然值为 $P(o \mid H_0) = (1/M)^{\mathrm{T}}$，所以由式(4)，$P(o \mid H_1) > \gamma' = \gamma P(o \mid H_0)$。由式(2)，$P(o \mid H_1) > \gamma'$ 等价于 $q_e^e(o) > \Gamma$，则式(4)等价于

$$L'(o) = q_e^e(o) > \Gamma \text{（I类误差决定 } \Gamma \text{ 大小）} \tag{5}$$

I类误差 P_{fa} 是似然比 $L'(o) > \Gamma$ 时接受 H_0 的概率 $P_{\mathrm{fa}} = P(L'(o) > \Gamma; H_0) = (1/M)^{\mathrm{T}} Q(\Gamma)$，其中 $Q(\Gamma) = |\{o; q_e^e(o) > \Gamma\}|$ 表示使得似然比 $L'(o) > \Gamma$ 且长度为 T 的时间序列个数，且 $Q(\Gamma) \le \sum_{k>\Gamma} C_T^k M^{\lambda T} \cdot (M-1)^{(1-\lambda)T-k}$，因为长度为 T 的时间序列中，选择 k 个时刻放置频繁模式状态组合数为 C_T^k，λT 个状态由其前一时刻噪声状态决定，共有 $M^{\lambda T}$ 种组合，剩余状态由离其最近的频繁模式状态决定，共有 $(M-1)^{(1-\lambda)T-k}$ 种组合。因此

$$\begin{aligned}P_{\mathrm{fa}} &\le (1/M)^{\mathrm{T}}\sum_{k>\Gamma}C_T^k M^{\lambda T}(M-1)^{(1-\lambda)T-k}\\&< 1-\left(\frac{M}{M-1}\right)^{T\lambda_{\min}(k\le\Gamma)}\\&\cdot\sum_{k\le\Gamma}C_T^k\left(1-\frac{1}{M}\right)^{T-k}\left(\frac{1}{M}\right)^k\\&\approx 1-\left(\frac{M}{M-1}\right)^{T\lambda_{\min}(k\le\Gamma)}\Phi\left(\frac{\Gamma-\frac{T}{M}}{\sqrt{T\left(\frac{1}{M}\right)\left(1-\frac{1}{M}\right)}}\right)\\&=1-c\Phi\left(\frac{\Gamma-\frac{T}{M}}{\sqrt{T\left(\frac{1}{M}\right)\left(1-\frac{1}{M}\right)}}\right)\end{aligned}$$

其中 $\Phi(\cdot)$ 为正态随机变量累计分布函数，由中心极限定理，当 T 足够大时上式第 3 步近似推导成立。$c=\left(\frac{M}{M-1}\right)^{T\lambda_{\min}(k\le\Gamma)}$ 为常数，$\lambda_{\min}(k\le\Gamma)$ 表示在所讨论的 $k\le\Gamma$ 时间序列集合中 λ 的最小值。给定 I 类误差上限 ε（$\varepsilon=0.5$），则 $\Gamma=\frac{T}{M}+\sqrt{\left(\frac{T}{M}\right)\left(1-\frac{1}{M}\right)}\cdot\Phi^{-1}\left(\frac{1-\varepsilon}{c}\right)$，当 T 足够大时 Γ 近似等于 T/M，又因为 $q_e^e(o)=Nf_\alpha$，根据式(5)可得 $q_e^e(o)=Nf_\alpha>T/M$ 时拒绝假设 H_0。综上，当 $f_\alpha>T/(MN)$，$0<\lambda<1/2$ 时，α 为 o 中显著频繁模式。

3 基于支持度的特征选择

给定数据集，属性集合 X，类别集合 C，信息增益 $\mathrm{IG}(C\mid X)$ 越大，属性集合 X 对类别集合 C 的区分能力越大[6]：

$$\mathrm{IG}(C\mid X)=H(C)-H(C\mid X) \tag{6}$$

其中 $H(C)$ 是数据集中样本所属类别概率的信息熵，$H(C\mid X)$ 是条件信息熵，数据集给定则 $H(C)$ 一定。由式(6)可知，当 $H(C\mid X)$ 达到最小下界 $H(C\mid X)_{lb}$ 时，$\mathrm{IG}(C\mid X)$ 达到最大上界 $\mathrm{IG}(C\mid X)_{ub}$，即 $\mathrm{IG}_{ub}(C\mid X)=H(C)-H_{lb}(C\mid X)$。不妨设 $X\in\{0,1\}$，$C=\{0,1\}$。属性 x 在数据集中出现的概率 $P(x=1)=h(f)$，类别为 1 的样本出现的概率 $P(c=1)=p$，属于类别 1 具有属性 x 的样本的概率 $P(c=1\mid x=1)=q$。当 x 为频繁模式时，$h(f)$ 表示 x 在时间序列样本集合中为显著模式的概率，f 表示 x 在整个数据集中出现次数[1)]。则

$$\begin{aligned}H(C\mid X)&=-\sum_{x\in\{0,1\}}P(x)\sum_{c\in\{0,1\}}P(c\mid x)\log_2P(c\mid x)\\&=-h(f)q\log_2q-h(f)(1-q)\log_2(1-q)\\&\quad+(h(f)q-p)\log_2\frac{p-h(f)q}{1-h(f)}+(h(f)\\&\quad\cdot(1-q)-(1-p))\log_2\frac{(1-p)-h(f)(1-q)}{1-h(f)}\end{aligned}$$

若给定数据集，上式中 p 为常数，若 $h(f)\le p$，当 $q=0$ 或 1 时 $H(C\mid X)$ 达到下界；如果 $h(f)\ge p$，当 $q=p/h(f)$ 或 $1-(1-p)/h(f)$ 时 $H(C\mid X)$ 达到下界。$h(f)\le p$ 和 $h(f)\ge p$ 对称，当 $h(f)\le p$ 时 $q=0$ 和 $q=1$ 也对称，因此只列出 $h(f)\le p$，$q=1$ 时 $H(C\mid X)$ 下界取值的讨论：

$$\begin{aligned}H_{lb}(C\mid X)_{q=1}&=(h(f)-1)\left(\frac{p-h(f)}{1-h(f)}\log_2\frac{p-h(f)}{1-h(f)}\right)\\&\quad+\frac{1-p}{1-h(f)}\log_2\frac{1-p}{1-h(f)},\\\frac{\partial H_{lb}(C\mid X)_{q=1}}{\partial f}&=h'(f)\left(\log_2\frac{p-h(f)}{1-h(f)}-\frac{p-1}{1-h(f)}-\frac{1-p}{1-h(f)}\right)\\&=h'(f)\log_2\frac{p-h(f)}{1-h(f)}\end{aligned} \tag{7}$$

因为 $\log_2\frac{p-h(f)}{1-h(f)}\le\log_21=0$，所以 $h'(f)\ge0$，$\frac{\partial H_{lb}(C\mid X)_{q=1}}{\partial f}\le0$，反之亦然。通常情况下 λ 一定，$h(f)$ 一定单调递增。所以，当 $h(f)\le p$ 时 f 增加 $H_{lb}(C\mid X)$ 单调递减，$\mathrm{IG}_{ub}(C\mid X)$ 增加。基于频繁模式的时间序列分类框架如下：

输入：时间序列数据集，信息增益阈值 IG_0

输出：分类模型

(1)由式(7)计算使 $\mathrm{IG}_{ub}(h(f))\le\mathrm{IG}_0$ 成立的 $h(f)$ 的最大值 $h^*(f)$

(2)循环：

(a)数据集中每个时间序列样本

(b)计算每个样本中的非重叠显著频繁模式(支持度大于 $S_{\min}=T/(MN)$，且 $0<\lambda<1/2$)。

(c)记录每个频繁模式在时间序列样本集合中为显著模式的次数 h_α。

(3)结束循环

(4)选择 $h_\alpha>h^*(f)\times n$ 的频繁模式作为时间序列属性(n 为时间序列样本个数)。

(5)将时间序列中各时刻事件作为样本的属性和频繁模式属性一起用于训练分类模型。

(6)输出训练后的分类模型

采用文献[10]的方法选择信息增益阈值 IG_0。分类模型采用 C45，SVM 等传统模型。

4 实验

实验采用准确率(precision)、召回率(recall)和 F-Measure 评价分类模型，采用 SAX[2)]方法[5]离散化连续时间序列，C45[3)]模型作为分类模型，选择时间序列样本中事件和频繁模式作为属性。

4.1 UCI 时间序列数据集实验

“synthetic control signal”(简称“scs”)和

1) f 和 f_α 的含义不同，f 表示频繁模式 α 在数据集中“所有”时间序列样本中出现的次数；f_α 表示 α 在“一个”时间序列样本 o 中出现的次数。

2) http://www.cs.ucr.edu/~eamonn/SAX.htm

3) http://www.cs.waikato.ac.nz/ml/weka/

“Japanese vowel”(简称“jv”)是 UCI 数据库中公共数据集。“scs”数据集包括 6 种控制信号，每种信号有 100 个长度为 60 的样本，随机抽取 10%的样本作为测试集，其余作为训练集。“jv”数据集包括 9 名男性发/ae/音的记录，每条记录长度在 7–29 帧之间，数据集已划分训练集和测试集，共有 640 个样本。

从表 1 看出，在各数据集上的实验均表明，选择频繁模式作为时间序列属性对 3 种指标均有明显提升。只选择显著频繁模式作为属性分类性能优于选择所有频繁模式作为属性的分类性能，可见研究频繁模式的显著性是有必要的。

4.2 智能楼宇传感器采集数据实验

本文将 MNOE 算法和分类框架用于实现某智能楼宇 WSAN(Wireless Sensor Ad-hoc Network)数据分析系统 EventMiner。WSAN 采集 6 种类型事件，每类事件有 3 种取值，采集周期为 1 min 。本文在 2008 年 6 月到 11 月的数据上进行实验，数据集中每条记录包括时间戳和在该时刻 6 个传感器节点所采集的事件。本文用长度为 61 的滑动窗口取滑动步长为 1，分别将 6–8 月和 9–11 月的数据分为两组时间序列集合(Dataset1 和 Dataset2)，其中每条记录的前 60 个时刻的事件作为时间序列样本，第 61 个事件作为分类标志。分别将 Dataset1 和 Dataset2 中所占比例小于 3%的分类标志和其对应的样本作为噪声删除，然后取 80%的样本作为训练集，20%的样本作为测试集(如表 2 所示)。

宏平均准确率、召回率和 F-Measure 示于表 3。

由表 3 看出，在实际数据集中，选择显著频繁模式作为时间序列属性进行分类在准确率、召回率和 F-Measure 上任具有较大优势。综上，本文所提出的分类框架在公共数据集和实际应用数据集上均有很好表现，显著频繁模式的研究是必要的，本文提出的分类框架具有良好的鲁棒性，简单的选择所有频繁模式作为属性进行分类并不一定能提升分类性能。

5　结束语

本文提出一种基于非重叠频繁模式的时间序列分类框架。框架在特征提取阶段由 MNOE 算法提取非重叠频繁模式作为待选特征。在特征选择阶段，利用 EGMAMC 模型选择出显著频繁模式，并根据信息增益定义选择信息增益大于给定阈值的显著频繁模式作为时间序列特征进行分类。本文从理论上推导出频繁模式在时间序列中出现次数与分布和其是否显著以及是否被选为时间序列特征之间的关系，因此，在实际分类过程中不用进行似然比检验，简化了分类过程。根据 EGMAMC 的定义可知模型是稳定的，从实验中也得到证明。

实验证明，研究显著频繁模式是必要的，简单的选择所有频繁模式作为时间序列特征并不一定能提高分类模型性能。下一步工作将集中在研究时间序列中事件间空间-时序(spatial-temporal)模式对时间序列分类模型性能的影响。

表 1 宏平均准确率、召回率和 F-Measure

数据集	未包含频繁模式			所有频繁模式			显著频繁模式		
	准确率	召回率	F-Measure	准确率	召回率	F-Measure	准确率	召回率	F-Measure
Scs	0.714	0.5	0.588	0.889	0.8	0.842	**0.909**	**0.852**	**0.88**
Jv	0.556	0.5	0.526	**0.667**	0.6	0.632	0.636	**0.7**	**0.667**

表 2 数据集

数据集	记录数	月份	训练集大小	测试集大小	分类数
Dataset1	131298	6–8 月	101887	25471	5
Dataset2	131072	9–11 月	101711	25427	5

表 3 宏平均准确率、召回率和 F-Measure

数据集	未包含频繁模式			所有频繁模式			显著频繁模式		
	准确率	召回率	F-Measure	准确率	召回率	F-Measure	准确率	召回率	F-Measure
Dataset1	**0.667**	0.6	0.632	0.6	0.6	0.6	0.571	**0.8**	**0.667**
Dataset2	0.444	**0.4**	0.421	0.5	0.3	0.375	**0.571**	**0.4**	**0.471**

参 考 文 献

[1] Boukerche. Handbook of Algorithms for Qireless Networking and Mobile Computing. Chapman & Hall/CRC, 2005.

[2] Aach J and Church G. Aligning gene expression time series with time warping algorithms. *Bioinformatics*, 2001, 17(6), 495–508.

[3] Laxman S. Stream prediction using a generative model based on frequent episodes in event sequences. Proceeding of Knowledge Discovery and Data Mining Conference 2008, Las Vegas, Nevada, USA,30 Jul. 2008: 453–461.

[4] Vladimir Vapnik. The Nature of Statistical Learning Theory. New York: Springer Verlag, 1999, Chapter 4.

[5] Lin J, Keogh E, Lonardi S, and Chiu B. A symbolic representation of time series with implications for streaming algorithms. Proceedings of the 8th ACM SIGMOD workshop on research issues in data mining and knowledge discovery, San Diego, California, 9 Jun. 2003: 2–11.

[6] Cheng H, Yan X, Han J, and Hsu C W. Discriminative frequent pattern analysis for effective classification. Proceeding of International Conference on Data Engineering 2007, Istanbul, 17 April, 2007: 716–725.

[7] Liu B, Hsu W, and Ma Y. Integrating classification and association rule mining. Proceedings of the 7th International Workshop on New Directions in Rough Sets, Data Mining, and Granular-Soft Computing, London, UK, Springer-Verlag 1999: 443–447.

[8] Patel D, Hsu W, and Lee M L. Mining relationships among interval-based events for classification, Proceeding of International Conference on Management of Data / Principles of Database Systems, Vancouver, Canada, 10 Jun. 2008: 393–404.

[9] Laxman S, Sastry P S, and Unnikrishnan K P. Discovering frequent episodes and learning Hidden Markov Models: A formal connection. *IEEE Transactions on Knowledge and Data Engineering*, 2005, 17(11): 1505–1517.

[10] Yang Y and Pedersen J O. A comparative study on feature selection in text categorization. Proceeding of International Conference on Machine Learning, San Francisco, USA, 8 Jul. 1997: 412–420.

万 里： 男，1981年生，博士生，卡耐基梅隆大学访问学者，研究方向为网络智能化、人工智能、信号处理.

廖建新： 男，1965 年生，教授，博士生导师，主要研究方向为网络智能化.

朱晓民： 男，1974 年生，副教授、硕士生导师，研究方向为智能网、下一代业务网络、3G 核心网、协议工程等.

倪 萍： 男，1978 年生，博士生，卡耐基梅隆大学访问学者，研究方向为网络智能化、人工智能、信号处理.