doi:10.11835/j.issn.1000-582X.2020.03.010

储油罐液位时序数据模式发现

文必龙,马 强,李 菲
(东北石油大学 计算机与信息技术学院,黑龙江 大庆 163318)

摘要:储油罐液位时序数据模式发现对油田生产管理、灾害预警有重要意义,由于目前油气田领域的数据体系繁杂,并未对这些数据加以分类和标识。已有方法借助图形化工具进行人工筛选与检查,这样的方法不适用于长时间不间断生产的石油工业。面对上述问题及已有方法的不足,针对储油罐液位时序数据的特点,提出基于层叠分段与层次聚类模式发现的处理方法。将观测序列转换为离散的线性分段序列,并对各线性分段进行基于 DTW(距离的无监督层次聚类,可自动发现时序模式并分配标识符标注时序序列。以储油罐液位时序数据进行实验,发现了隐含的变化模式和变化规律。方法对液位时序变化模式有很好的识别及分类能力,无需人工筛选与检查,并可根据需要,查看不同粒度的变化模式,可为时序数据模式识别,异常检测提供参考和途径。

关键词:层叠线性分段;形态相似度量;时间序列分析;模式发现
中图分类号:TP391 **文献标志码**:A **文章编号**:1000-582X(2020)03-088-12

Pattern discovery of liquid level time series data in oil tank

WEN Bilong, *MA Qiang*, *LI Fei*
(School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, Heilongjiang, P. R. China)

Abstract: The liquid level of oil tank time series data model is of great significance for oilfield production, management and disaster warning. Due to the miscellaneous data system in the field of oil and gas fields, these data are not classified and marked. There are some methods for manual screening and checking with graphical tools, which are not suitable for long time uninterrupted production of petroleum industry. In the face of the above problems and the shortcomings of the existing methods, a processing method which based on cascade piecewise linear representation and hierarchical clustering for the characteristics of reservoir tank level data is proposed. The observation sequence is transformed into discrete linear piecewise sequence, and each linear segment is clustered by unsupervised hierarchical clustering based on DTW distance, which can automatically discover the temporal pattern and assign identifiers to annotate the sequence. Based on the data of oil tank level sequence data, and the implied models and the changing rules were found. The method has a good ability to recognize and classify the time series change patterns of liquid level, without manual

收稿日期:2019-02-13
基金项目:国家自然科学基金面上资助项目(41574117);国家重大专项资助项目(2016ZX05033-005-004)。
Supported by General Program of National Natural Science Foundation of China(41574117), National Science and Technology Major Projects of China(2016ZX05033-005-004).
作者简介:文必龙(1967—),男,教授,博士,主要从事大数据、软件工程方向研究,(E-mail)bilong_wen@126.com。
通信作者:马强,男,硕士研究生,主要从事大数据。知识工程方向研究(E-mail)maqiang52@163.com;(Tel)15776543719。

screening and inspection, and can view the changing patterns of different granularity according to the need, which can provide a reference and avenues for time series data pattern recognition and abnormality detection.

Keywords: cascade piecewise linear representation; similarity measure; time series analysis; pattern discovery

油气集输过程中产生了大量的储油罐液位时序数据，使用自动记录方法，按油田生产周期中的时间，以储罐中油液位为观测对象采集得到，是油气生产中一类重要数据形式和数据资源，发现这些大量时序数据中的知识成为一个巨大的问题。液位时序数据中常包含独特模式，这些模式可以作为偷油、漏油、跑料、抽空等事件预警、检测的主要指标，也是进行知识发现(KDD，knowledge discovery in database)的基本元素。它反映了生产状况、调度安排管理、生产设备是否安全稳定运行等情况，是生产灾害确定的直接预警因素和构建灾害防护系统的重要依据。对液位数据的模式进行分析挖掘，揭示生产、管理等内在规律和存在的问题，对生产指挥以及做出科学决策具有重要意义[1-4]。

许多研究者对时序数据模式发现处理进行研究，液位时序数据属于非平稳序列，其变化复杂多样造成液位模式地发现和预测以及潜在时间趋势地理解更具挑战性。要想实现数据有效挖掘而获取知识，必须对数据进行处理与描述，识别液位序列的变化模式。模式的发现有 2 个作用，检测生成数据序列的系统何时发生变化，创建观测序列的高级数据表示。[3-8]。生产单位现有的数据处理是借助图形化工具进行人工筛选与检查，这样的方法不适用于长时间不间断生产的石油工业，并且面对大量数据，人工分辨成本太高。文献[4]、[9]利用线性化分段与神经网络模糊聚类结合的方法对时间序列进行符号化处理，为液位模式的发现提供了参考，但需要对序列预先构造存储样本模式。文献[10]对时间序列进行微分再求其双谱的方法判断序列间的相似性，方法对高相似度的序列具有很好效果，但对非平稳不连续的序列难以处理。文献[11]、[12]利用统计特征矢量符号化方法对时序数据进行处理，方法侧重于时序数据的降维与压缩，会损失部分信息，不利于后续数据异常检测等处理。文献[13]利用 K-Means 聚类对时序数据进行聚类并预测，而实际观测对象的数据变化模式多样，难以预先指定类簇个数。文献[14-15]基于动态时间弯曲聚类算法对时序数据进行聚类分析，寻找相似序列，方法给出了分段后序列相似判断的方法，但无法确定合适的类簇个数。根据相关研究的经验和不足，提出基于层叠线性分段与基于动态弯曲相似的层次聚类处理方法，可以自适应对液位进行分段处理，能对相似序列进行发现，并确定合适的模式个数，用于解决储油罐液位时序数据模式发现的问题。

方法基本思想是，首先利用变化阈值估算方法确定液位稳态变化的阈值，然后利用基于阈值的层叠线性化分段方法(CPLR，cascade piecewise linear representation)将时间序列转换为离散的拟线性分段序列。最后根据各子序列变化形态，利用基于 DTW 距离的层次聚类算法对各线性分段进行聚类，利用轮廓系数确定最佳类簇个数，发现液位变化模式并分配标识符。该方法可根据设定的变化阈值自适应进行分段，获得不同的表示精度，对不同变化趋势的序列有很好的识别与发现能力，帮助实现对液位数据进行多粒度、多层次的挖掘分析。

1　液位时序特征分析及描述

为达到有效利用液位数据特征进行模式发现的目的，下面对液位数据基本特征进行分析，并在此基础上对数据进行形式化描述。

1.1　数据特征

储油罐液位时序数据是依时间间隔取值得到的离散观测记录，数据包含 3 个主要属性，观测时间(单位：s)、液位高度(单位：cm)、储油体积(单位：m^3)，以时间和液位高度构建时序序列，如图 1 所示。液位整体上呈现连续性，按时间次序可拟合成连续的曲线。曲线的形态变化蕴含着

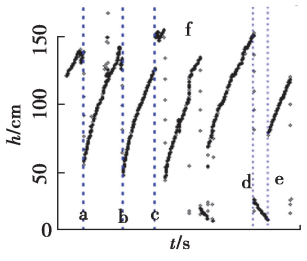


图 1　某井液位数据（部分）

Fig. 1　Liquid level data of a well

围绕生产的工作变化信息。

液位时序数据具有时间性和周期性，从一次拉油结束为起始点，到另一次拉油结束为终结点，该时间单位为一个完整的工作周期，如图1中ab段，液位数据以工作周期为单位，成周期变化。数据从不同周期对比上看，数据变化形态呈现相似性。观测数据周期变化与变化形态的相似性是模式提取的基础。

数据在工作周期内的变化具有阶段性。如液位变化中对应生产过程的液位上升阶段(如图1中bc段)和对应拉油调度阶段的液位下降阶段(如图1中de段)。在阶段变化当中，存在观测值稳定变化的状态与拐点。拐点是数据进行分析的关键点，它反映了外部因素对液位变化的影响，如液位上升阶段与下降阶段的分隔点就是重要的拐点，反映了工作变化的调整。工作周期内还存在毫无规律的若干波动点形成的突变，突变在观测序列中表现十分明显，如液位幅值急剧变化，引起液位的不连续(如图1中(f))。

1.2 形式化描述

1.2.1 时序序列

设全体液位时序序列 T，$T=\{t_1,t_2,t_3,\cdots,t_i,\cdots,t_n\}$。$T$ 是液位总体变化，在数据 T 中，如果 $0<i<j<=n$，则称 t_i 先于 t_j 发生。称 t_i 与 t_{i+1} 为相邻的。在实际的模式发现中，更关心 T 在某个时间段内的变化，即其某个子序列 S 的变化。

1.2.2 观测对象子序列

给定一个长度为 n 的序列 T，设 T 的长度为 m $(m<n)$ 的子序列 S 为：$S=\{t_p,t_{p+1},\cdots,t_{p+m-1}\}$。$p$ 是子序列 S 的起始点，且 $1\leqslant p<n-m+1$。定义子序列的目的是为了方便将全体序列数据 T 进行分段描述与处理，求取 T 在其子序列上的变化特征。研究将具有一致变化趋势的有序数据定义为一个模式。不同的模式根据子序列的定义描述如下：设某个子序列为 $S=\{t_p,t_{p+1},\cdots,t_{p+m-1}\}$，如果 $t_p\leqslant t_{p+1}\leqslant\cdots\leqslant t_{p+m-1}$ 则称 S 为 T 的上升 m 长子序列。如果 $t_p\geq t_{p+1}\geq\cdots\geq t_{p+m-1}$ 则称 S 为 T 的下降 m 长子序列。

数据 T 是其全体子序列 S 的有序集合，在数据的处理中，可通过线性化分段方法获得满足特定条件的子序列。原始序列 T 的线性分段后的序列为 $T=\{S_1,S_2,S_3,\cdots,S_i,\cdots,S_q\}$。

1.2.3 观测对象变化阈值

液位的观测值在工作周期中呈波动状态，设观测值变化阈值为 ε $(\varepsilon>=0)$，ε 是液位稳态变化所能达到的最大波动范围。ε 越小，对液位稳态波动的限制越严格，则分段时越敏感。反之，ε 越大，对液位稳态变化的波动幅度限制越低，切分时段就分得越粗。液位变化阈值与拐点密切相关。

2 层叠线性分段方法

2.1 趋势估值及变化阈值

为了从液位的时间序列中分割和提取基本的相对独立的变化模式，采用线性化分段方法对序列 T 进行分段，使分段后序列满足以下条件：1)每个子序列 S 都近似线性，即单个子序列 S 中序列值的变化趋势都较为接近；2)各子序列 S 之间序列值的变化趋势区分明显。特殊情况为单个序列值作为1个分段，表示与相邻序列值变化趋势差距较大[16]。特殊值是液位异常检测的依据。

液位在 i 时刻的变化趋势为 $\dfrac{x_i-x_{i-1}}{t_i-t_{i-1}}$，由于观测数据为等时间间隔采样，因此将问题简化，对时间变化 t_i-t_{i-1} 取1，构造趋势估值函数 F，定义如下

$$F=\begin{cases}x_{t+1}-x_t, & \text{如果 } t=i;\\ x_t-x_{t-1}, & \text{如果 } t=n,\end{cases}$$

利用估值函数 F 对观测序列进行近似求导，输入数据 T，输出为序列 T 上各点的变化趋势估计值。设DIFF为数据 T 的一阶趋势估值序列，则 $\mathrm{DIFF}=F(T)=\{d_1,d_2,\cdots,d_i,\cdots,d_{n-1},d_n\}$ 是时间序列 $T=\{t_1,t_2,t_3,\cdots,t_i,\cdots,t_n\}$ 各个时刻变化趋势的合理估计。

DIFF为 T 的趋势序列，为取得趋势相近的分段，根据液位稳态变化所能达到的最大限制，即变化阈值 ε 进行判断。变化阈值 ε 取不同值，分段方法可以对同一时间序列产生不同颗粒度下的观察。一般情况下，波动幅值为一离散的随机变量，为了能在粗分阶段取得最小颗粒度分段，需根据观测值波动统计幅值波动规律

及分布情况来确定用于不同颗粒度下的变化阈值 ε，初分时选择能最大区分类别的范围确定限制阈值 ε。

2.2 层叠线性分段

在层叠线性分段方法(CPLR)中，引入变化阈值 ε、估值函数 F，对给定的观测数据序列 T，将相邻的并且变化趋势接近的序列值划分在同一个分段中，使分段之间序列值的变化趋势区分明显。层叠线性分段方法描述如算法 1。

算法 1：基于变化阈值及估值函数的层叠线性分段

输入：时序序列 $T=\{t_1,t_2,t_3,\cdots t_i\cdots,t_n\}$；

变化阈值 ε，

估值函数 F。

过程：1)初始类标签 flag=1，

T 的初分类标签集合 Label= $\{l_1,l_2,\cdots l_i,\cdots l_n\}$；

2)对所有观测到的序列值计算波动趋势：$d_i=F(t_i)$；

a)如果 $|d_i|\leqslant\varepsilon$，$l_i=$flag；否则转 b)；

b)如果 $|d_i|>\varepsilon$，flag=flag+1；$l_i=$flag；

3)对 T 中所有观测值 t_i 依据其对应的粗分类标签 l_i 划分类别，得到初分子序列集合 $T'=\{S_1,S_2,S_3,\cdots,S_i,\cdots S_q\}$。

4)对初分子序列集合每个子序列计算序列长度：$\text{len}_i=\text{length}(S_i)$，

如果 $\text{len}_i=1$，添加 S_i 到 O 集合中；将 S_i 从初分集合中删除：delet Si from T'；

5)对数据集合 O，统计并确定变化阈值 ε′，并进行步骤 1)、2)、3)；

6)合并初分序列集、再分序列集为新序列集合：$T''=T'+O'$；

输出：T 的模式集合 T''。

序列 T 经过层叠线性处理，根据其变化趋势与变化形态被分成不同的子序列，每个子序列都表达了液位在该段时间内的变化特征，并且不同子序列在液位的阶段变化当中相对独立。可以认为一个子序列是某个相对独立的变化模式在该时间段内的具体体现。

时序数据的分段线性表示是为了方便模式发现的后续处理，初分阶段得到趋势相近的子序列是拟线性的，其对应的分段标签 l_i 是 T 中可能的变化模式在某时间段的标记，要发现的模式与时间段无关，是观测对象自身的变化规律。各子序列变化形态的相似性及周期性，为观测数据通过相似性判断发现模式提供了可能。

3 基于聚类的模式发现

聚类是数据挖掘中通过相似性、距离等度量方法对数据进行汇聚而发现信息的有效方法。观测数据可以通过时间序列间的相似性进行聚类分析，发现观测对象状态变化的相同模式。聚类目标是得到变化趋势与变化形态相似的子序列的汇聚的簇的集合，使得每个类簇内的子序列是相似的，不同类簇的子序列是不相似的，以不同的簇代表不同的模式。液位数据聚类需要解决 2 个问题，子序列相似性度量和聚类方法。相似性度量应能对时序数据的变化、波动等特征有很好的识别与区分能力；聚类方法要适合观测数据的表示形式并能捕获数据的自然结构[17]。

3.1 形态相似性度量

序列相似性度量的依据是数据在时间尺度上的变化相似，常用的时序数据相似度量方法有欧式距离和动态时间扭曲距离(DTW)等[18]。

给定 2 个多维向量，$\boldsymbol{Q}=\{q_1,q_2,\cdots,q_n\}$ 和 $\boldsymbol{S}=\{s_1,s_2,\cdots,s_n\}$，则 $\boldsymbol{Q}$ 和 $\boldsymbol{S}$ 之间的欧式距离为

$$\text{dis}(\boldsymbol{Q},\boldsymbol{S})=\sqrt{\sum_{i=1}^{n}(Q_i-S_i)^2},$$

欧式距离存在无法处理不同维度数据的局限性[19-21]，因此选择对时序数据中相位偏移、形状扭曲有识别能力的(DTW)方法计算 2 个子序列的相似度[18]。

对给定的2个子序列：$\boldsymbol{S}=(S_1,S_2,\cdots,S_i,\cdots S_m)$、$\boldsymbol{Q}=(Q_1,Q_2,\cdots,Q_j,\cdots Q_n)$，其相似性计算如下，构造 $m*n$ 的矩阵 $\boldsymbol{A}$。矩阵 $\boldsymbol{A}$ 中的每个元素为 $a_{ij}=\mathrm{dis}(S_i,Q_j)$。在矩阵 $\boldsymbol{A}$ 中搜寻扭曲路径，其中起点 $P_1=A_{11}$，终点 $P_k=A_{mn}$。对 $P_h=A_{ij}$，$P_{h-1}=A_{xy}$，必须满足连续性和单调性约束：$i-x\geq 0$；$0\leqslant j-y\leqslant 1$。序列中某2个时间点之间的距离 $\mathrm{dis}(i,j)=\min\{\mathrm{dis}(i-1,j-1),\mathrm{dis}(i,j-1),\mathrm{dis}(i-1,j)\}$，将满足条件的 i、j 作为路径值添加到规整路径 P 中。根据规整路径 $P=(P_1,P_2,\cdots,P_h,\cdots P_k)$，将每个时间点上的距离求和作为2个时间序列之间的相似值 $\mathrm{sim}(\boldsymbol{S},\boldsymbol{Q})$。

3.2 子序列层次聚类及模式发现

聚类算法的选择取决于数据的类型、聚类的目的和应用[18]。液位数据子序列聚类的目的是为了将不同工作周期内的相同变化模式挖掘出来，并将模式应用到数据库中，分析生产事件或生产任务状态及变化。期望各子序列根据自身结构无监督的进行聚类，同时聚类结果可以从不同角度和层次观察液位变化的关系。依据上述目的和要求对聚类算法进行考察，层次聚类算法使用数据的连接规则，通过一种层次架构方式，无监督且无需指定参数，反复将数据进行分裂或聚合，以形成一个层次序列的聚类问题解，最终将数据按层次结构组织[22]，适合于液位数据序列模式的分析与挖掘。

层次聚类一般可采用“自底向上”的聚合策略[23]，和“自顶向下”的分拆策略[19]。经过实验对比，选择自底向上的聚合策略，将层叠线性分段输出的子序列集中的每个序列看作一个初始聚类簇，然后按照簇间距离找出最近的2个类簇进行合并，不断重复该过程，直至达到预设的想查看的聚类簇个数。算法的关键是如何计算聚类簇之间的距离，定义2个类簇间的距离为不同簇的所有子序列间的DTW扭曲距离的平均值为簇间距离[17]，即平均距离：$d_{avg}(C_i,C_j)=\dfrac{1}{|C_i||C_j|}\sum_{x\in C_i}\sum_{z\in C_j}dtw(x,z)$。

对时间序列 T 分类集合 T'' 中的子序列的聚类过程如算法2所述。

算法2：基于DTW形态相似距离的聚类算法

输入：T 的模式集合 $T''=\{S_1,S_2,\cdots,S_m\}$，

目标查看簇数目 k，

过程：1）初始化原始簇 C，对时间序列 T 分类集合 T'' 中的每个子序列 S_j，都作为原始簇：$C_j=\{S_j\}$

2）对原始簇 C 中的每个簇两两计算相似度，得到子序列相似度矩阵 $\boldsymbol{M}$：$\boldsymbol{M}(i,j)=\mathrm{dtw}(C_i,C_j)$；$\boldsymbol{M}(j,i)=\boldsymbol{M}(i,j)$

3）设置当前聚类簇个数 q：$q=m$，

4）在当前簇个数大于要聚类的个数时：$q>k$，

a）在簇相似矩阵 $\boldsymbol{M}$ 中找出距离最近的2个聚类簇 C_{i*} 和 C_{J*}；

b）将簇 C_{i*} 和 C_{J*} 合并成新的簇：$C_{i*}=C_{i*}\cup C_{j*}$；

c）对相似度矩阵中的簇更新编号：

for $j=j^*+1,j^*+2,\cdots,q$ do

将聚类簇C_j 重新编号为 C_{j-1}；

d）删除相似度矩阵 $\boldsymbol{M}$ 中的第 j^* 行与第 j^* 列；

e）计算更新后的簇 C_{i*} 与其它簇的相似度，更新相似度矩阵 $\boldsymbol{M}$：$\boldsymbol{M}(i^*,\mathrm{j})=d_{avg}(C_{i^*},C_j)$；$M(j,i^*)=M(i^*,j)$

6）$q=q-1$；

输出：模式的划分 C。

层次聚类后得到各子序列间的簇间关系，对应于不同粒度和层次的待查看模式类别数由用户根据经验和实际应用需求设定，即可得到不同的模式划分集合。根据划分集合为每个类分配一个类标识符，将类中对应的子序列以标识符标注或表示，得到时序数据 T 的模式表示。

4　实验结果及分析

4.1　实验结果

取某井储油罐中 2015 年 10 月～2016 年 1 月的 62 万条液位数据，进行实验。实验中以时间和液位高度构建时序序列。液位的波动值 X 为离散的随机变量，利用趋势估值函数 F 计算液位变化趋势，统计 DIFF 中观测数据点间的波动情况，波动幅值频率(前 10)分布情况如表 1 所示。

表 1　液位数据波动幅值分布概率

Table 1　Probability of fluctuation amplitude distribution of liquid level data

序号	波动值	概率
1	0	0.903 559 228
2	1	0.054 731 525
3	−1	0.022 628 398
4	−2	0.001 722 678
5	3	0.001 489 095
6	4	0.001 357 704
7	2	0.001 124 12
8	5	0.000 744 547
9	−3	0.000 642 355
10	−14	0.000 613 157

由表 1 可见，波动值在[−1，1] 之间取值的概率约为 98%，占据了绝大多数。液位在稳态变化区间[−1，1]变化为主要趋势。对液位数据应用层叠线性分段与聚类方法处理，聚类的结果是层次化树结构，为确定最佳簇个数，先选取可能的类簇范围，应用不同簇个数下的轮廓系数来确定最佳聚类数[25]，轮廓系数其值在−1 到+1 之间，值越大表示聚类效果越好。实验中设置簇的取值范围为[2，30]，计算的轮廓系数变化如图 2 所示，系数最大值在 $k=2$ 时取得，但聚为 2 簇应用意义不大，因此选择次最大值 9 或 12，同时结合工作调度的先验知识，最终确定簇个数为 9。所识别的 9 类模式基本特征和时间序列数据模式后标注表示如表 2、图 3 所示。利用模式替换对应子序列，形成液位序列的符号表示，以液位下降阶段对应的模式 0 为分割，得到液位序列的模式表示如表 3。

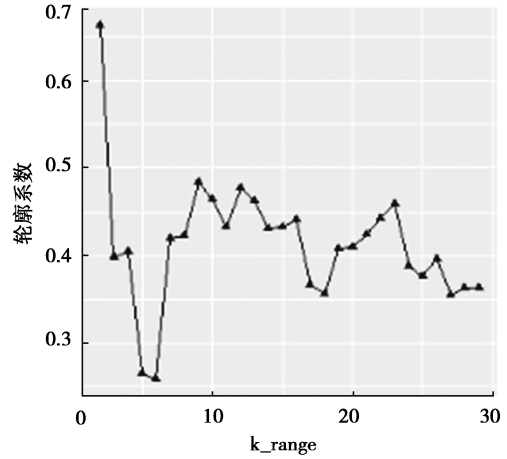


图 2　轮廓系数随簇个数变化

Fig. 2　Silhouette Coefficient

表 2　模式基本特征

Table 2　Basic characteristics of patterns

模式序号	子序列个数	平均长度/s	平均变化率
1	57	286.6	0.059 697 033 8
2	9	1 803.0	0.043 382 826 1
3	42	370.7	0.059 774 960 9
4	12	1 261.2	0.041 312 614 5
5	104	9.1	0.236 385 975 4
6	1	2 519.0	0.040 095 276
7	10	338.4	0.036 691 904 8
8	16	503.2	−0.021 955 382 8
9	50	184.0	0.029 061 752 6

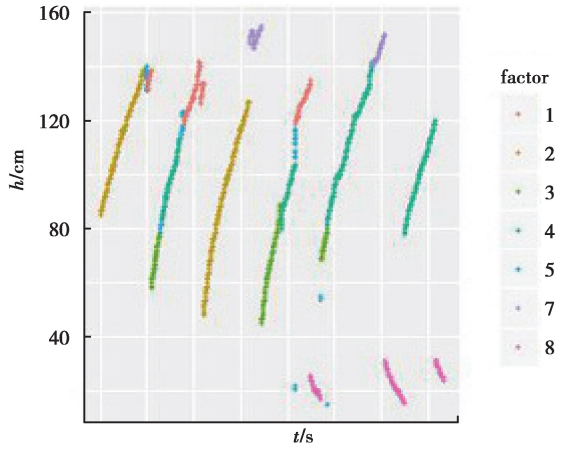


图3 某井液位数据及其模式标注(部分)

Fig. 3 Liquid level data and pattern annotation of a well

表3 液位序列的模式表示

Table 3 Symbolic representation of liquid level sequence

序号	模式串	序号	模式串	序号	模式串
1	<2,1>	16	<2,7>	31	<4>
2	<3,4,5>	17	<3,4,5,1>	32	<8,5>
3	<3,4,1>	18	<2,7>	33	<5,3,1>
4	<3,4,1>	19	<3,4,5,1>	34	<0,8,5>
5	<2,5,1>	20	<0,8>	35	<5,4>
6	<3,6>	21	<4>	36	<0,8,5>
7	<1>	22	<0,8,5>	37	<5,4>
8	<2,1,4,1>	23	<4>	38	<5,4,1>
9	<3,1,4>	24	<0,8>	39	<0,8,5>
10	<4,5,7>	25	<4>	40	<5,4>
11	<2>	26	<5,8,5>	41	<0,8,5>
12	<2,5,1>	27	<5,4>	42	<5,4>
13	<2,1>	28	<0,8>	43	<4>
14	<2,5,1>	29	<5,4>		
15	<3,5,4,5,1>	30	<0,8>		

由表2、3可知,在储油罐的液位变化中。模式5出现频次最高,该模式的特点是,出现时间短,变化率大,且经常出现在变化模式的边界处。模式3、4有可能是某种固定的变化模式,从液位数据来看,它有如下特点,液位上升一段时间,然后出现拐点,且拐点造成整体液位下移,之后继续上升。模式1通常出现在液位达到最高点附近,且在模式1之前通常存在液位短暂下降。模式8表示在液位降到最低点后缓慢下降。液位按工作调度变化,原则上只有上升和下降部分,且上升时均匀平缓,下降时速度快,2个阶段对应的模式应该是单一的。因此任何上升或下降阶段内发生的模式改变的衔接点都有可能是液位异常。如模式3、4间的衔接点,模式2、3间的衔接点等。由于导致液位模式异常变化的外部因素很多,如抽油机的检修、温度、液位计故障等。因此对异常的判定还需要参考其它信息,发现液位模式及模式标注为油气生产故障分析与异常监测、检测提供预警与参考。

4.2　方法评价

实验中根据表 1 统计值设定限制阈值为 1.1。液位数据第一次分段后的平稳段识别如图 4 所示。分段的合理与否取决于分段内是否趋势相同与接近，分段后的液位子序列波动方差与标准差反映了波动是否平稳。对每个子序列计算波动状态，求其均值，见表 4。

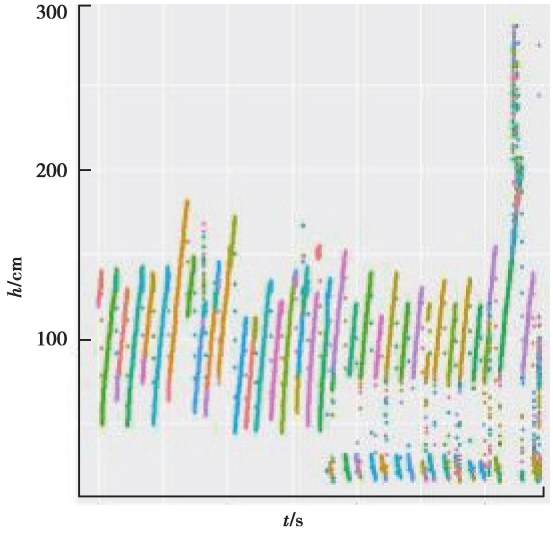


图 4　阈值 ε 取 1.1 时液位分段

Fig. 4　Sectional representation of liquid level($\varepsilon=1.1$)

表 4　不同阈值下线性分段结果及 PAA 对比

Table 4　Comparison of linear segmentation results and PAA under different thresholds

序号	CPLR				PAA	
	阈值 ε	平稳段数目	各段平均方差	各段平均标准差	各段平均方差	各段平均标准差
1	1.1	301	34.377 48	2.705 2	210.651 4	7.986 194
2	2.1	278	39.892 77	3.130 304	243.325 4	8.693 907
3	3.1	300	38.764 17	3.341 303	223.28	8.069 101
4	4.1	297	43.414 12	3.793 738	204.856 4	7.729 249
5	5.1	288	47.537 99	4.101 278	212.290 1	80.985 87
6	6.1	276	51.114 23	4.353 104	224.735 6	8.620 702

表 4 给出了利用 CPLR 方法变化阈值取不同值时，识别出的平稳分段平均方差与平均标准差的变化情况，由表可知，当变化阈值越大，即对稳态波动限制越小，所识别的平稳分段数目越小，平稳分段的平均方差与平均标准差越大，这表示分段粒度越大，各分段内的变化波动趋势越大。按照 CPLR 选择的阈值情况下的识别出的平稳段数据为依据，采用 PAA 分段方法[24]，对液位数据进行分段对比实验，CPLR 方法能保证线性分段后的子序列变化趋势相近，波动平稳。

在聚类实验中，分别采用分裂策略层次聚类方法和 L2 范数作为对比实验进行层次聚类，统计不同方法下聚类数据及实验结果如表 5、图 5 所示。基于 DTW 的凝聚层次聚类的聚类效果最好，对不同形状的子序列的识别与区分最清楚，其它方法中都存在一定的缺陷，如图 5(a)中第 6 类、图 5(b)中第 3 类、图 5(c)中第 5 类，将不同趋势的变化序列识别为一簇。而在时间复杂度上，由于 DTW 需要比较两个子序列中所有点间的

情况,方法存在较大的时间消耗。

表 5 层次聚类不同策略对比实验统计

Table 5 Comparison of different hierarchical clustering strategy experiments

序号	层次分类方法	距离度量	聚类所需时间/s
1	分裂策略	L2	5.35
2	凝聚策略	L2	5.41
3	分裂策略	DTW	387.22
4	凝聚策略	DTW	377.21

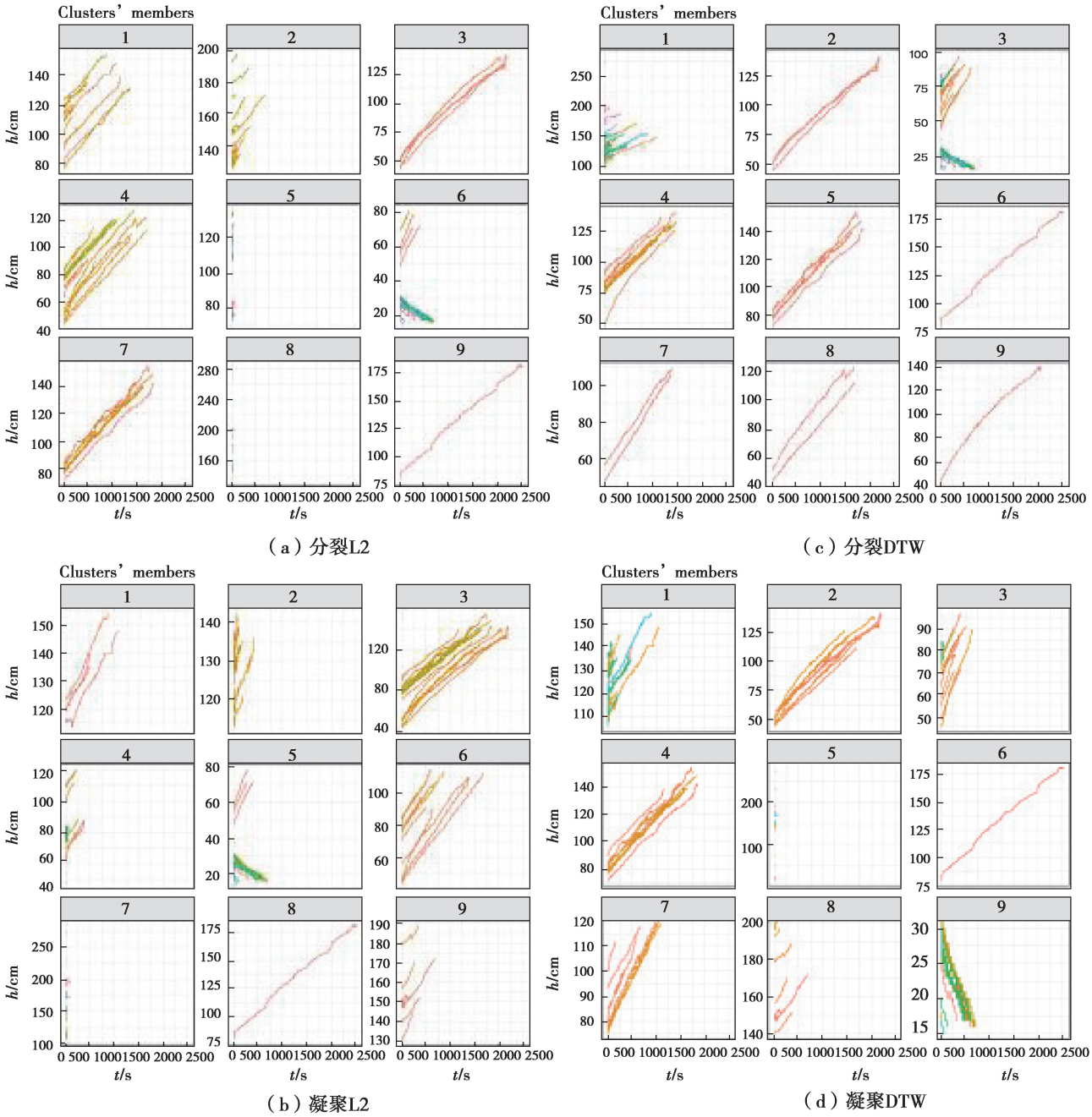
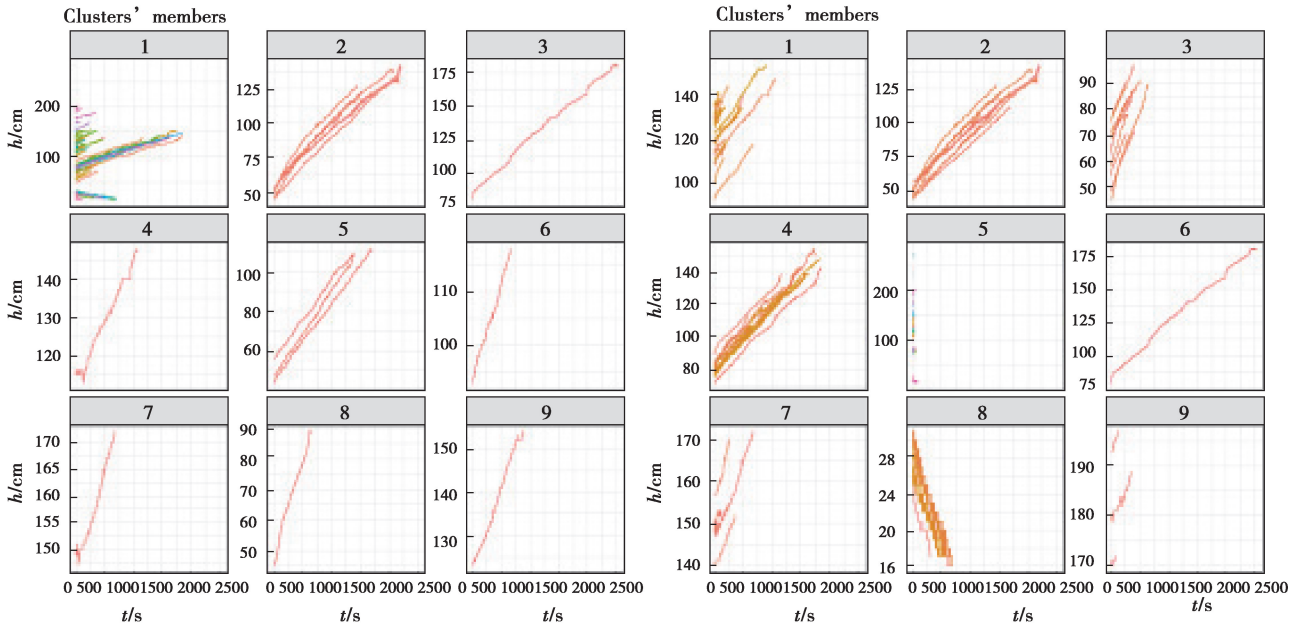


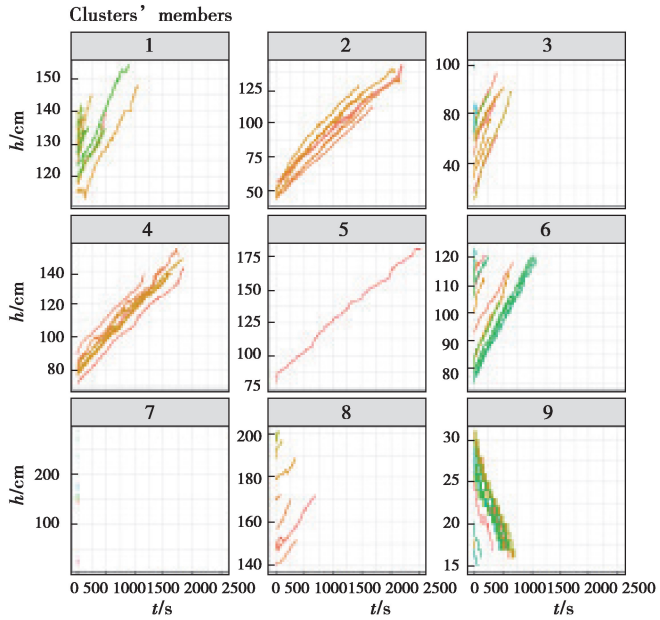
图 5 不同聚类策略和距离测度下聚类结果

Fig. 5 Clustering results under different clustering strategies and distance measures

在确定凝聚策略与 DTW 距离的前提下，分别对不同的类簇间邻近性度量方法进行实验。查看聚为 9 簇的结果，实验结果及数据见图 6、表 6。单链的方式倾向于将大多数序列合为同一簇，能发现极为独立的簇，但对大多数的变化识别不明显。中位数的方式容易将序列长度较为接近的部分序列识别成一簇，如图 6(c)中类簇 6。对比全链与组平均的聚类结果(图 6(b)与图 5(d))，对相似序列的聚类结果基本相同，但组平均的方式将更多相似序列识别为 1 簇且其类簇内距离更为紧凑。见表 6 全链簇 1 与组平均簇 1、全链簇 3 与组平均。



(a) 单链　　(b) 全链



(c) 中位数

图 6　DTW 距离与凝聚策略下不同类簇间邻近性度量方法结果

Fig. 6　Different proximity measure results

表6 DTW距离与凝聚策略下不同类簇间邻近性度量方法簇内个数及平均距离

Table 6 The number and average distance of clusters in the proximity measure between different clusters

簇编号	单链		全链		中位数		组平均	
1	286	5 034.633 4	22	2 226.727 3	29	656.862 1	57	812.350 9
2	6	866.666 7	9	2 014.000 0	9	2 014.000 0	9	2 014.000 0
3	1	0.000 0	10	1 193.500 0	49	647.755 1	42	717.571 4
4	1	0.000 0	20	2 233.650 0	12	779.916 7	12	779.916 7
5	3	367.666 7	213	479.347 4	1	0.000 0	104	215.740 4
6	1	0.000 0	1	0.000 0	46	1 883.413 0	1	0.000 0
7	1	0.000 0	5	1 655.600 0	91	261.956 0	10	909.900 0
8	1	0.000 0	18	529.833 3	22	2 921.636 4	16	4 693.375 0
9	1	0.000 0	3	2 815.333 3	42	829.500 0	50	801.560 0

5 结 语

时序数据模式发现是数据挖掘和应用的基本任务。模式发现方法的研究对生产、安全管理、作业调度以及知识发现有重要的意义和应用价值。但由于油气田领域的数据体系繁杂，并未对这些数据加以分类和标识。研究在对液位时序数据分析的基础上进行形式化描述，给出了一种层叠线性分段方法，并基于序列变化的形态相似性进行聚类，识别变化模式。实验证明，该方法对模式有很好的识别及分类能力，发现了液位数据库中存在的液位变化模式、拐点及可能的异常模式，为液位异常检测及生产调度提供参考，同时也为油田知识发现提供可能。为进一步运用该方法准确地分析生产管理及液位间关系提供了一种可行的途径。同时，该方法的时间消耗较大，在线性分段时，稳态波动阈值的确定方法单一。接下来的工作中，将对方法进行深入研究，提高方法的执行效率与适用性。

参考文献：

[1] Gong X Y, Si Y W, Fong S, et al. Financial time series pattern matching with extended UCR suite and support vector machine[J]. Expert Systems With Applications, 2016, 55: 284-296.

[2] Lu C, Shi Y T, Chen Y Y, et al. Data mining applied to oil well using K-means and DBSCAN[C/OL]. 2016 7th International Conference on Cloud Computing and Big Data (CCBD). New York, USA: IEEE, 2016[2019-09-25]. https://doi.org/10.1109/ccbd.2016.018.

[3] Shokoohiyekta M. Meaningful rule discovery and adaptive classification of multi-dimensional time series data[J]. Dissertations & Theses - Gradworks, 2015, 13(10):323-327.

[4] 李斌，谭立湘，章劲松，等. 面向数据挖掘的时间序列符号化方法研究[J]. 电路与系统学报，2000，5(2)：9-14.
LI Bin, TAN Lixiang, ZHANG Jinsong, et al. The study of the data mining oriented method for the symbolization of time series[J]. Journal of Circuits and Systems, 2000, 5(2): 9-14. (in Chinese)

[5] 李爱国，覃征. 在线分割时间序列数据[J]. 软件学报，2004，15(11)：1671-1679.
LI Aiguo, QIN Zheng. On-line segmentation of time-series data[J]. Journal of Software, 2004, 15(11): 1671-1679.(in Chinese)

[6] Fayyad U, Uthurusamy R. Data mining and knowledge discovery in databases[J]. Communications of the ACM, 1996, 39(11): 24-26.

[7] 蔡智，岳丽华，王熙法. 时序模式发现算法研究[J]. 计算机研究与发展，2000，37(9)：1107-1113.
CAI Zhi, YUE Lihua, WANG Xifa. Research on an algorithm for time series patterns discovery[J]. Journal of Computer Research and Development, 2000, 37(9): 1107-1113. (in Chinese)

[8] Berndt D J, Clifford J. Using dynamic time warping to find patterns in time series[J]. Intelligent Information Management, 1994, 359-370.

[9] 杨雨浓，修春波. 联想神经网络的风速序列预测分析[J]. 重庆大学学报，2016，39(4)：139-146.
YANG Yunong, XIU Chunbo. Wind speed time series prediction based on associative network[J]. Journal of Chongqing University, 2016, 39(4): 139-146. (in Chinese)

[10] 吴文兵，黄荣华，刘日华，等. 耦合信号微分后对双谱的影响[J]. 重庆大学学报，2017，40(7)：82-90.
WU Wenbing, HUANG Ronghua, LIU Rihua, et al. The affection of differentiated coupled signals on bispectrum [J]. Journal of Chongqing University, 2017, 40(7): 82-90. (in Chinese)

[11] Baydogan M G, Runger G. Learning a symbolic representation for multivariate time series classification[J]. Data Mining and Knowledge Discovery, 2015, 29(2): 400-422.

[12] Bondu A, Boullé M, Cornuéjols A. Symbolic representation of time series: A hierarchical coclustering formalization [M]. Cham: Springer International Publishing, 2016: 3-16.

[13] Dai D, Mu D. A fast approach to K-means clustering for time series based on symbolic representation[J]. International Journal of Advancements in Computing Technology, 2012, 4(5): 233-239.

[14] Kate R J. Using dynamic time warping distances as features for improved time series classification[J]. Data Mining and Knowledge Discovery, 2016, 30(2): 283-312.

[15] Lee J, Yoo S, Kim H, et al. The spatial and temporal variation in passenger service rate and its impact on train dwell time: A time-series clustering approach using dynamic time warping [J]. International Journal of Sustainable Transportation, 2018, 12(10): 725-736.

[16] 尹锐，李雄飞，李军，等. 基于线性分段与 HMM 的时间序列分类算法[J]. 模式识别与人工智能，2011，24(4)：574-581.
YIN Rui, LI Xiongfei, LI Jun, et al. Time series classification algorithm based on linear segmentation and HMM[J]. Pattern Recognition and Artificial Intelligence, 2011, 24(4): 574-581. (in Chinese)

[17] 陈封能，斯坦巴赫，库玛尔.数据挖掘导论(完整版)[M].北京：人民邮电出版社，2011：305，32-327.
Fengneng C, Michael Steinbach, Vipin Kumar. Introduction to data mining[M]. BeiJing: the People's Posts and Telecommunications Press, 2011: 305, 32-327. (in Chinese)

[18] Aljawarneh S, Radhakrishna V, Kumar P V, et al. A similarity measure for temporal pattern discovery in time series data generated by IoT [C/OL]. 2016 International Conference on Engineering & MIS (ICEMIS). New York, USA: IEEE, 2016(2016-11-17)[2019-09-25]. https://ieeexplore.ieee.org/document/7745355/.

[19] 周东华，叶银忠. 现代故障诊断与容错控制[M]. 北京：清华大学出版社，2000：209-214.
ZHOU Donghua, YE Yinzhong. Modern Fault Diagnosis and Fault-Tolerant Control [M]. Beijing: Tsinghua University Press, 2000: 209-214. (in Chinese)

[20] Keogh E J, Pazzani M J. Scaling up dynamic time warping for datamining applications[C]// Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '00. New York, USA: ACM Press, 2000: 285-289.

[21] Ratanamahatana C A, Keogh E. Multimedia retrieval using time series representation and relevance feedback[M]. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, 3815: 400-405.

[22] 孙吉贵，刘杰，赵连宇. 聚类算法研究[J]. 软件学报，2008，19(1)：48-61.
SUN Jigui, LIU Jie, ZHAO Lianyu. Clustering algorithms research[J]. Journal of Software, 2008, 19(1): 48-61. (in Chinese)

[23] Fox W R, Kaufman L, Rousseeuw P J. Finding groups in data: an introduction to cluster analysis[J]. Applied Statistics, 1991, 40(3): 486.

[24] Guo C H, Li H L, Pan D H. An improved piecewise aggregate approximation based on statistical features for time series mining[M].Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, 6291: 234-244.

[25] Rousseeuw P J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis[J]. Journal of Computational and Applied Mathematics, 1987, 20: 53-65.

（编辑　侯　湘）