时间序列的表示与分类算法综述

原继东 王志海
(北京交通大学计算机与信息技术学院 北京 100044)
(交通数据分析与挖掘北京市重点实验室 北京 100044)

摘 要 时间序列是按照时间排序的一组随机变量,它通常是在相等间隔的时间段内,依照给定的采样率,对某种潜在过程进行观测的结果。时间序列数据广泛地存在于商业、农业、气象、生物科学以及生态学等诸多领域,从时间序列中发现有用的知识已成为数据挖掘领域的研究热点之一。在时间序列表示方面,主要介绍了非数据适应性表示方法、数据适应性表示方法和基于模型的表示方法;针对时间序列的分类方法,着重介绍了基于时域相似性、形状相似性和变化相似性的分类算法,并对未来的研究方向进行了进一步的展望。

关键词 时间序列,时间序列分类,时间序列表示
中图法分类号 TP391.4 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.3.001

Review of Time Series Representation and Classification Techniques

YUAN Ji-dong WANG Zhi-hai
(School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)
(Beijing Key Lab of Traffic Data Analysis and Mining, Beijing 100044, China)

Abstract Time series is a set of random variables ordered in timestamp. It is often the observation of an underlying process, in which values are collected from uniformly spaced time instants, according to a given sampling rate. Since time series data exist widely in various application domains, such as finance, agriculture, meteorology, biological science, ecology and so on, discovering knowledge from time series has become one of the mainly research fields of data mining. In this paper, a comprehensive review on the existing time series representation and classification research was given. In the term of time series representation, three different categories named non-data adaptive, data adaptive and model based were summarized. A summary of several time series classification method, namely similarity in time, similarity in shape and similarity in change was also provided.

Keywords Time series, Time series classification, Time series representation

1 引言

一条时间序列是一组序列数据,它通常是在相等间隔的时间段内,依照给定的采样率,对某种潜在过程进行观测的结果。现实生活中,在一系列时间点上观测数据是司空见惯的活动[9],比如在商业上,我们会观测日股票闭盘价、周利率、月价格指数、年销售量等,图1展示了某股票近4年来日股票闭盘价的时间序列;在气象上,我们会观测太阳黑子的活动情况、每天的最高/最低温度温度、年降水量、每小时的风速等,图2展示了1770年到1869年间太阳黑子的活动情况;在生物科学上,我们会观测每毫秒心电或脑电活动的状况等,图3给出了某病人的心电图活动状况;另外,在农业上,我们会记录不同农作物每年的产量、土壤侵蚀情况、农产品出入口销量等方面的数字;在生态学上,我们会记录不同动物种群数量的变动情况等等。目前,时间序列数据正以不可预测的速度产生于现实生活中的几乎每一个应用领域。

如图1—图3所示,时间序列数据是实值型的序列数据,具有数据量大、数据维度高以及数据是不断更新的等特点。直接在初始时间序列数据上进行挖掘工作是非常耗时的,为高效地处理时间序列数据,需采用一种简洁的方式来表示时间序列数据,此表示方法需要从时间序列的形状出发,并在降低时间序列维度的同时保持其重要的特征[22]。

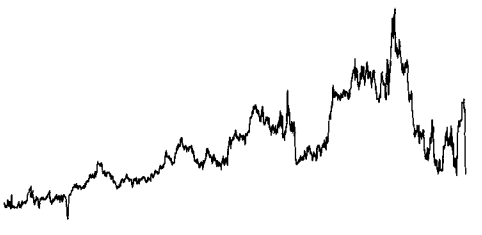

图1 股票交易数据

到稿日期:2014-04-01 返修日期:2014-07-16 本文受北京市自然科学基金(4142042),中央高校基本科研基金(2014YJS032)资助。
原继东(1989—),男,博士生,主要研究领域为数据挖掘和模式识别,E-mail:12112078@bjtu.edu.cn;王志海(1963—),男,博士,教授,博士生导师,主要研究领域为数据挖掘和机器学习。

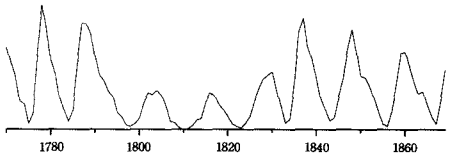

图 2　1770 年到 1869 年间太阳黑子的活动情况

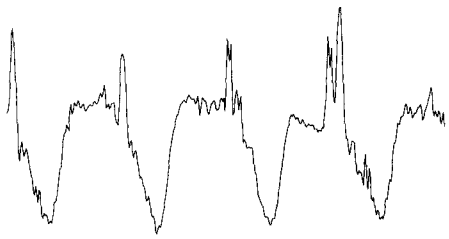

图 3　心电图示例

另外，在时间序列分类问题中，任意实值型有次序的数据被当作一条时间序列[42]。时间序列分类的目标是首先从标定类标的训练集中学习到能够区分不同序列的鉴别性特征。然后，当遇到一条未标定类标的时间序列时，它能够自动为该时间序列分配类标。时间序列分类问题与传统分类问题之间的差别在于，属性的次序在传统的分类问题中是不重要的，并且变量之间的相互关系独立于它们的相对位置；而对于时间序列数据而言，变量的次序在寻找最佳的辨别性特征时起着至关重要的作用[1]，因此，时间序列分类问题已成为机器学习领域的诸多挑战之一。

本文拟介绍时间序列的表示和分类方法，第 2 节阐述时间序列数据挖掘的基本定义；第 3 节描述时间序列表示的 3 种方法，包括非数据适应性表示方法、数据适应性表示方法和基于模型的表示方法；第 4 节主要介绍时间序列的不同分类方法，即基于时域相似性的分类算法、基于形状相似性的分类算法和基于变化相似性的分类算法；第 5 节介绍时间序列表示和分类算法的进一步研究方向；最后对全文进行总结。

2　相关定义

本节介绍文中涉及到的基本定义。

定义 1(时间序列数据集)　D 为一个时间序列数据集，其中包含有 n 条时间序列，即 $D=\{T_1,T_2,\cdots,T_n\}$。

定义 2(时间序列)　时间序列 T 是一条长度为 m 实值的序列，可表示为 $T=t_1,t_2,\cdots,t_m$。一般情况下，数据点 t_1，$t_2,\cdots,t_m$ 是按照时间顺序排列的，两两之间具有相同的时间间隔，此时的时间序列为离散型时间序列。连续型时间序列的观测值是在连续的时间点上取得的。另外，按照某一时间点上观测变量的多少，时间序列可划分为单变量时间序列和多变量时间序列，而本文只讨论单变量离散型时间序列的表示和分类算法。

定义 3(时间序列子序列)　给定一条时间序列 $T=t_1$，$t_2,\cdots,t_m$，其子序列 $S_{i,l}=t_i,t_{i+1},\cdots,t_{i+l-1}$ 是一条 T 中从位置 i 开始长度为 $l(l\leqslant m)$ 的连续子序列，此处 $1\leqslant i\leqslant m-l+1$。

定义 4(时间序列的表示)　给定一条长度为 m 的时间序列 $T=t_1,t_2,\cdots,t_m$，若维度为 $d(d\ll m)$ 的模型 T' 与 T 非常接近，则称 T' 为时间序列 T 的表示。

定义 5(时间序列分类)　假设一个时间序列数据集中有 n 条时间序列，$D=\{T_1,T_2,\cdots,T_n\}$，每一条时间序列有 m 个观测值和一个类值 c_i，即 $T_i=\langle t_{i1},t_{i2},\cdots,t_{im},c_i\rangle$。时间序列分类问题可以表述为一个将时间序列的观测值映射到类值的函数。

$$dist(x,y)=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_i-y_i)^2} \tag{1}$$

定义 6(时间序列间的相似性度量)　为获得两条时间序列间的有效比对并保证缩放和偏移的不变性，在处理时间序列之前，须对每条时间序列进行规范化处理[51]。另外，为允许不同长度时间序列间的比对，需要通过除以时间序列的长度来规范化距离，并称之为长度规范化(如式(1)所示)。

为得到时间序列子序列与整个时间序列间的距离，短的子序列必须在长的序列上滑动以得到它们间的最短距离，我们称此距离度量为子序列距离，并定义为：

$$subdist(x,y)=\min(dist(x,y_{|x|})) \tag{2}$$

其中，$y_{|x|}$ 表示时间序列 Y 中长度为 $|x|$ 的子序列。需要注意的是，$subdist()$ 得到的是两个时间序列间的最小距离。注意，为方便起见，我们将式(1)中的规范化距离表示为欧氏距离(Euclidean Distance，ED)，而其他距离度量方式如曼哈顿距离、动态时间规整(Dynamic Time Wrapping，DTW)等皆可应用于时间序列间的相似性度量。

3　时间序列表示

如前文所述，为有效存储和加快时间序列的处理过程，我们需要采用一种简洁的方法来表示高维的时间序列数据。怎样表示一条时间序列是时间序列挖掘的基础问题[25]，一种有效的时间序列表示方法不仅能够允许序列间进行相似性比对，也可以较好地应用于不同的数据挖掘任务中。时间序列表示方法的基本特征包括：有效地降低数据维度；强调局部或全局的形状特征；较低的计算消耗；能够根据约减后的表示较好地重构原数据；对噪音不敏感或者能够隐式地处理噪音等[22]。

每一种时间序列表示方法都从不同侧面强调了上述的多个基本特征。根据不同的转换方式，Ratanamahatana 和 Keogh 等人[53]将不同的时间序列表示方法分类为非数据适应性的、数据适应性的和基于模型的 3 种，下面将一一阐述。

3.1　非数据适应性表示方法

在非数据适应性表示方法中，每一条时间序列的转换参数是一致的。频谱分析是一种比较常见的非数据适应性表示方法。Agrawal 等人[1]首次采用离散傅里叶变换(Discrete Fourier Transform，DFT)将时间序列映射到频域，并使用 R*-tree 对序列进行索引和相似性查询。该方法有效解决了时间序列挖掘中“特征抽取的完备性”和“维度灾难”这两个问题。自从 DFT 被应用于时间序列上之后，多数研究者采用欧氏距离来度量转换后的时间序列，而 Bagnall 等人[3]提出了一种基于似然函数的距离度量方式来比较经过 DFT 处理后的时间序列间的相似性。Chan 等人[11]首次提出用离散小波变换(Discrete Wavelet Transform，DWT)来处理时间序列。DWT 能同时表示时间序列中的时域和频域信息，而 DFT 只能表示频域信息。Popivanov 等人[49]对比了不同小波变换在时间序列数据上相似性搜索的效率，通过实验验证了多贝西(Daubechies)小波相较于哈尔(Harr)小波的高效性(前者更光滑，拟合性较好)。Liabotis 等人[37]采用小波变换来降低时间序

列的维度，并使用 X-Trees 进行索引，由于所用的小波变换更接近初始的时间序列，实验结果也表明了此方法相比较于 DFT 的高效性。

除频谱分析之外，研究者还提出了其它专门用于时间序列表示的方法。比如，Keogh 等人[31]提出了一种基于逐段线性分割(Piecewise Linear Segments)的方法来表示时间序列的形状，此方法允许快速的时间序列分类、聚类以及相关反馈等工作。之后，Keogh 等人[32]又提出了一种新的维度约减技术 PAA(Piecewise Aggregate Approximation)，通过在索引速度和灵活性等方面与传统的奇异值分解(Singular Value Decomposition，SVD)、DFT 和 DWT 做比较，阐述了 PAA 在时间序列相似性度量和索引上的优势。

3.2 数据适应性表示方法

数据适应性表示方法在数据转换时，转换参数随着时间序列数据的变化而变化。非数据适应性表示方法可以转化为数据适应性表示方法，比如 Keogh 等人[33]在 PAA 的基础上提出了一种自适应的维度约减技术 APCA(Adaptive Piecewise Constant Approximation)用于时间序列表示。

众所周知，数值型高维时间序列数据是难以掌控的，但当把它们表示为符号序列而不是实值序列时，从中挖掘和发现有趣的模式或规则将变得更加容易。Lin 等人[39]提出了一种符号聚集近似(Symbolic Aggregate approXimation，SAX)表示方法，该方法可以将初始的实值型高维数据转换成离散的低维数据。如图 4 所示，一条维度为 150 的时间序列经过 SAX 表示，转换成了维度为 10 的符号序列“aaacddcbba”。在低维离散型的数据上进行搜索是更有效的。更重要的是，研究者可以更加容易地处理 SAX 转换后的数据，比如可以从文本处理或者是生物信息处理中借鉴相应的处理方法(如随机映射等)，也可以将哈希表、后缀树等数据结构应用到时间序列数据中。

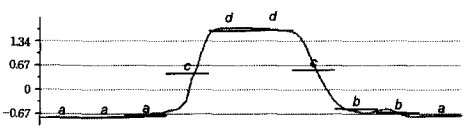

图 4 SAX 表示

时间序列 shapelets 是时间序列中能够最大限度地表示一个类别的子序列，而时间序列表示的主要动机是通过一种简洁的方式突出数据的主要特征，正好与时间序列表示的动机相符。时间序列 shapelets 的具体概念是由 Ye 等人[64,65]提出的。如 Esling 等所述，此方法大大缩小了时间序列和形状分析间的差距[22]。

3.3 基于模型的表示方法

基于模型的表示方法假设一条时间序列是对某潜在模型的观察结果。Azzouzil 等人[2]首次提出采用隐马尔科夫模型(Hidden Markov Model，HMM)来定义时间序列变量间的关系，并将此模型成功应用于石油勘探数据中。Kalpakis 等人[30]采用求和自回归移动平均模型(Auto Regression Integrated Moving Average，ARIMA)来简洁地表示时间序列，并为此表示方法定义了高效的相似性度量方式。Nanopoulos 等人[48]提出了一种基于统计模型(如均值、方差等)的特征抽取方法来表示整个时间序列。一般情况下，基于模式的表示方法具有较强的可解释性，若两条时间序列可以由同一潜在模型的相同参数集表示，则认为它们是相似的。

4 时间序列分类

时间序列分类问题作为序列分类问题的一个分支[59]，已经在时间序列挖掘领域引起了广泛关注。该问题广泛地存在于现实生活中的诸多领域，如健康信息处理中的心电图或脑电图分类[51]、气象中天气状况预测[44]、根据传感数据来区分不同的行为动作[26,47,64]、根据用电量来区分不同的家用电器[41]等。Keogh 等人还专门收集了用于时间序列分类/聚类的 UCR 数据集[1]。

在时间序列分类问题中，任意实值型有次序的数据被当作一条时间序列[42]。也就是说，数据是不需要在时间上有序的，任何逻辑上有序的实值型数据都可用于时间序列分类问题中。因此，研究者们也将时间序列分类方法用于植物叶片的识别、古文物中的箭头识别等[64]。

由于时间序列数据的特殊性，时间序列分类问题面临着 3 个主要方面的挑战。首先，对于多数分类器如决策树或神经网络来说，输入数据为特征向量，然而时间序列数据并没有明确的特征；其次，尽管可以在时间序列上使用特征选择的方法，但由于时间序列特征空间的维度非常大，特征选择的过程是非常繁琐的，此举的计算量很大；最后，在某些应用中，除了精确的分类结果之外，我们还希望得到具有可解释性的分类器。但由于时间序列数据没有明确的特征，建立一个可解释性的分类器是非常困难的。

所有的分类问题都依赖于数据间的相似性度量，时间序列分类问题也不例外。对于时间序列来说，同类时间序列间的相似性有以下 3 种形式[4]。

(1)时域相似性(Similarity in Time)：同一类别的时间序列都是在时间维度上对某一潜在相同曲线观察的结果，它们之间的不同可能是由噪音和相位漂移所引起的。1-NN 分类器最适合处理此类问题，而 DTW 度量可缓解噪音等带来的影响。

(2)形状相似性(Similarity in Shape)：同一类别的时间序列是通过一些相同的子序列或形状来区分的，而且这些子序列可能出现在时间序列的任意位置，这是它与时域上相似性的主要不同。子序列与时间的相关性越小，基于时域的 1-NN 分类器就越难处理此类问题，此时可通过使用基于时间序列特征的方法来区分不同的类别。

(3)变化相似性(Similarity in Change)：最不容易被观察到的相似性，此类相似性出现在自相关性较强的序列中。此问题可以用产生式模式如隐马尔科夫模型(HMM)、自回归移动平均模型(AutoRegressive Moving Average，ARMA)等来处理。

接下来，本文将详细阐述这 3 种相似性以及相对应的分类算法。

4.1 基于时域相似性的分类算法

在过去的十多年中，针对时间序列分类问题的研究主要

1) http://www.cs.ucr.edu/eamonn/time_series_data/

集中在基于不同距离度量方式的最近邻(1-NN)算法上,如基于ED或者DTW的1-NN等[20,34]。Faloutsos等人[23]首次将欧氏距离用于时间序列子序列的匹配算法中。Batista等人[6]提出了一种复杂性不变的距离度量方式CID(Complexity-Invariant Distance),用于缓解1-NN在处理复杂和简单时间序列时遇到的困境,并指出,"尽管目前有大量的分类算法应用于时间序列分类问题中,但实验表明,简单的1-NN分类器是很难被击败的"。另外,Buza等人[10]还介绍了一种融合不同距离度量方式的1-NN分类器。

欧氏距离对噪音数据和相位漂移比较敏感,而DTW能够较好地处理时间轴上的变形。Ding和Wang等人[20,57]对已有的时间序列维度约减和相似性度量方法做了统一的对比实验,并指出基于DTW的距离度量方法可能是当前最好的度量时间序列相似性的方法。Berndt等人[7]首次将之前应用于语音识别领域的DTW度量方法应用于时间序列的模式发现中。随后,Keogh等人[35]将基于DTW的精确索引应用于时间序列挖掘中。Fu等人[24]将DTW和US(Uniform Scaling)结合起来用于时间序列查询,并采用多维索引技术来提升查询效率。传统的DTW在计算两条时间序列间的距离时,对各对观测值赋予了相同的权重,而忽略了参考值和测试值之间的相位差,此缺陷可能引起形状相似性对比时的误分类问题。Jeong等人[29]为解决此问题,提出了一种加权的DTW来分类时间序列。虽然基于DTW的1-NN分类器是难以击败的,但该算法需要消耗大量的运算时间,所以它并不能很好地处理需要实时反馈的应用。为缓解此问题,Xi等人[58]提出将数据块消减(numerosity reduction)技术应用于基于DTW的1-NN分类器中,使之在拥有更快分类速度的同时保持较高的分类准确率。另外,Rakthanmanon等人[51]提出了一种基于DTW的快速的时间序列查询方法,并将其扩展到模式发现、聚类、分类以及时间序列数据流的挖掘中。

尽管1-NN分类器具有分类准确率高、易于实现等优点,但1-NN分类器只能说明被分类对象与所分到的类别间具有较大的相似性,并没有指出它与其它类别之间具体的不同点,即分类结果的可解释性较差。另外,1-NN分类器是一种懒惰式的分类器,它需要为每一个测试实例建立相应的分类器,并且需要对比整个数据集,所以采用1-NN进行分类还需要消耗大量的时间和空间。

4.2 基于形状相似性的分类算法

基于时域相似性的时间序列分类器在分类时间序列时,倾向于从整个时间序列入手。然而,在许多时间序列数据集中,同一类别的不同时间序列间往往存在着较大的差异性,基于时域相似性的分类器并不能很好地处理这种差异性,而基于形状相似性的分类算法能够很好地发现区分不同类别的最佳特征。

时间序列shapelets是序列中最具辨别性的子序列[64]。最初的基于shapelets的分类算法是由Ye等人[64,65]提出的,其采用信息增益度量数据的分裂点,并通过递归搜索最具有辨别性的shapelets来构建决策树。如图5所示,对于经典的Gun/NoGun(手中有枪/没枪)问题[64],最佳shapelet如加粗标注部分所示,它能够对Gun/NoGun数据集进行准确分类。Mueen等人[47]提出了使用逻辑shapelets来构建决策树的思想,并采用加速技术和剪枝策略来提升文献[64]中算法的可解释性和运行效率。Rakthanmanon等人[52]针对原有的shapelets发现算法在时间效率上的不足,提出了一种基于SAX[39]表示的快速shapelets发现算法,此算法在提升发现shapelets速率的同时保证了一定的分类准确性。

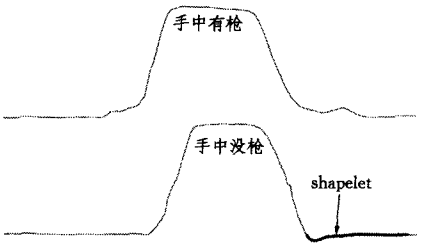

图5 shapelet示例

采用决策树解决时间序列分类问题的思想可追溯到Yamada等人[62]提出的二叉分类树。在文献[62]中,作者提出了两种分裂策略,第一种策略通过穷举搜索得到信息增益最大的一条时间序列并将其作为分裂节点;第二种策略通过穷举搜索得到信息增益最大的一对时间序列并将其作为分裂节点。两种策略均采用DTW距离来度量时间序列的相似性。Balakrishnan等人[5]在获取分裂节点时,采用k-means聚类($k=2$)来发现不同类别的代表性序列,这些代表性序列并不是原有的时间序列,而是位于类簇中心的一条抽象序列。Douzal-Chouakria等人[21]提出了一种综合考虑趋势相似与数值相似的时间序列相似性度量方法,并采用此度量方法建立能够处理时间序列的决策树。Deng等人[18]提出了一种随机森林方法[8],用于解决时间序列分类问题,同时提出了一种联合信息增益和距离度量的评价方式来选取分裂节点。

上述几种方法都是在发现shapelets的同时构建分类器。而Bagnall等人[4]强调了数据转换在时间序列分类中的重要性,并指出解决时间序列分类问题的最简方式是将时间序列映射到其他辨别性特征容易被检测到的空间。作者通过在建造分类器之前将时间序列分类问题转换到其他空间,显著提升了分类的准确性。Lines等人[42]提出了一种基于shapelets转换的时间序列分类方法,即将shapelets的发现与分类器的构建过程相分离,其主要优点是优化了shapelets的选择过程并能够灵活应用不同的分类策略。Hills等人[27]评估了其他3种不同的相似性度量方法(除信息增益外):Kruskal-Wallis、F-statistic、Mood's median对shapelets选择的影响。另外,时间序列shapelets还被广泛地应用于聚类[67]、姿势识别[26]、天气预测[44]、早期分类[60,61]等多个方面。

如前文所述,由于时间序列数据没有明确的特征,建立一个可解释性的分类器是非常困难的,而关联规则能够对知识进行简洁、直观的描述。关联规则挖掘已引起了研究者广泛的关注,其在分类方面的成功应用包括生物数据[15]、关系数据[12,13,36,43,55,56,66]、文本[28]以及图[19,63]等等。高维的时间序列数据天然地需要强有力的分析工具来抽取出最显著的和令人信服的规则,用以揭示序列模式和类标之间的重要关系,并将实值型的序列数据转换成相对容易理解的知识。但传统的关联规则挖掘算法(如CBA[43]、CPAR[36]、CMAR[66]等)主要致力于基于符号表示的事务数据挖掘。为将关联规则发现应用于时间序列数据中,我们需要将高维实值型序列离散化为

低维的符号序列才能进行关联规则的发现。Das 等人[16]通过聚类和离散化时间序列的子序列得到符号序列，并应用简单的规则发现方法来获取序列中的规则，但是此算法缺乏可解释性并且需要设置过多的参数。另外，尽管可以设置较高的支持度和置信度，仍可能从高维时间序列数据中发现大量的规则。Ting[54]等针对股票数据中的模式发现问题，提出了一种基于关联分析的方法对股票数据进行预测，然而，它容易产生过量的规则，并且许多规则在分类过程中是无用的。

4.3 基于变化相似性的分类算法

基于变化相似性的分类算法本质上是基于模型的分类算法，比如 Zhong 等人[68]成功将广泛应用于语音识别领域的 HMM 模型应用于时间序列中的脑电图(EEG)分类问题。在临床学习方面，为对齐和分类时间序列基因表达式，Lin 等人[40]提出了一种辨别性的 HMM 模型来提升分类准确率。Povinelli 等人[50]提出了一种基于相空间重构的高斯混合模型(Gaussian Mixture Models，GMM)，用于处理时间序列分类中的信号分类问题。Deng 等人[17]提出将自回归移动平均模型(AutoRegression Moving Average，ARMA)用于时间序列的辨别中。Nanopoulos 等人[48]提出了一种用于时间序列分类的多层传感神经网络方法(Multi-layer Perceptron，MLP)，此方法对数据的长度和噪音不敏感。尽管 HMM、GMM 等模型能突出时间序列间变化的相似性，但在分类准确性方面，它们中的多数被证明不如简单的基于 DTW 的 1-NN 算法[58]。

5 进一步的研究方向

时间序列的表示和分类方法在过去的十多年中得到了长足的发展，但现有的方法仍存在着不足之处，这也为我们未来的研究提供了一定的方向。

(1)将时间序列表示技术与分类方法相结合。时间序列表示技术多用于时间序列的索引和相似性查询，并未能很好地与时间序列分类技术相结合，因此有必要针对不同表示方法在分类方面的适用性，分析设计出适用于时间序列分类的表示技术，并将其应用于时间序列分类中。另外，除了 SAX 方法之外，现阶段研究者并没有提出其他能有效地将实值型时间序列转换为符号序列的方法，所以将时间序列数据表示为符号序列仍具有巨大的研究空间。研究者可以从研究数值型属性的离散化方法出发，将普遍的离散化方法转变为适合于时间序列挖掘的离散化方法，或者从时间序列数据的本身特性出发，设计出适合不同时间序列数据的离散化方法。

(2)研究时间序列中关联规则的发现方法，并将关联式分类应用于时间序列中。关联规则能够对知识进行简洁、直观的描述，依据关联规则进行分类明显具有很强的可解释性。但传统的关联规则挖掘算法主要集中于基于符号表示的事务数据上，所以针对时间序列数据的关联规则挖掘主要存在两方面的挑战：一是必须将时间序列数据离散化为片段，然后将每一个片段转换成一个符号；二是尽管可以设置较高的支持度和置信度，仍可能从高维时间序列数据中发现大量的未能应用到最终分类过程中的规则。上述两方面的困难严重制约了关联规则在时间序列数据中的应用。另外，时间序列 motif 是指反复出现在长时间序列中的子序列[14,38,45,46]，它的概念与频繁模式相近，亦可考虑将时间序列 motif 的发现与关联规则的生成相结合，用于生成具有可解释性的分类器。

(3)改进基于 shapelets 的时间序列分类算法。其包括：研究时间序列中 shapelets 的发现方法，提高 shapelets 的发现效率(如采用启发式方法)和辨别性，降低 shapelets 之间的相似性；将 shapelets 的发现方法与不同种类的数据特征相结合，即根据不同的应用问题设计出不同评价策略的 shapelets 发现算法；探索挖掘出的各个时间序列 shapelets 之间的依赖关系等。另外，候选 shapelets 最大、最小长度的设置也是一个难题。由于它们定义了候选 shapelets 的长度范围，参数设置不对时可能发现不了最具有辨别性的 shapelet，从而对分类器的准确率造成影响。

(4)设计更好的分类器融合策略。分类器融合策略的主要思想是将多样性融入所集成的分类器中。这里所说的多样性包括：使用不同的分类算法训练每一个基分类器来构成一个异构的集成分类器；通过采样机制或直接复制实例为每一个基分类器提供不同的训练数据(Bagging 思想)；随机选择不同的属性来训练每一个分类器；通过重新评估训练数据的权重来修正每一个分类器(AdaBoost 思想)等。

融合学习策略已被应用于时间序列数据挖掘中。Deng 等人[18]提出了一种集成多个时间序列分类树的时间序列森林(Time Series Forest，TSF)。Buza[10]提出了一种集成灵活性与非灵活性距离度量的分类器。Bagnall 等人[4]为了解决数据转换后可能引起的准确率大幅度差异的问题，提出了一种分类器集成策略来提升分类性能，此举降低了只在时间域上建立分类器时的可变性。此外，我们可以进一步设计出综合考虑时域、形状和变化相似性的集成分类器，避免只在时间域上或形状相似性上建立分类器时所带来的偏差，并根据不同的数据集和不同的应用场景，为基分类器设置不同的权重。

(5)多变量时间序列的表示和挖掘算法研究。研究者对时间序列的研究多数集中在单变量的时间序列表示和分类算法上，尽管目前已经有一些针对多变量复杂时间序列的研究工作[21,44]，但此部分的研究仍具有很大的发展空间。

结束语 本文首先从 3 个不同方面介绍了时间序列的表示方法，然后根据时间序列间不同的相似性度量方式，总结了 3 种不同的基于相似性的分类算法。由于时间序列数据的特殊性，现有的研究成果仍有许多不完善的地方，比如怎样将时间序列表示和分类方法相结合，怎样建造具有可解释性的分类器等，这些需要研究者进一步地努力。

参 考 文 献

[1] Agrawal R, Faloutsos C, Swami A. Efficient similarity search in sequence databases[C]// Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms (FODO 1993). 1993: 69-84

[2] Azzouzi M, Nabney I T. Analysing time series structure with Hidden Markov Models[C]// Proceedings of the IEEE Conference on Neural Networks and Signal Processing. 1998: 402-408

[3] Bagnall A, Janacek G J, Powell M. A likelihood ratio distance measure for the similarity between the fourier transform of time series[C]// Proceedings of the Advances in Knowledge Discovery and Data Mining, 9th Pacific-Asia Conference (PAKDD

2005). 2005:737-743

[4] Bagnall A, Davis L, Hills J, et al. Transformation based ensembles for time series classification[C]// Proceedings of the 2012 SIAM International Conference on Data Mining (SDM 2012). 2012:307-318

[5] Balakrishnan S, Madigan D. Decision trees for functional variables[C]// Proceedings of the 2006 International Conference on Data Mining (ICDM 2006). 2006:798-802

[6] Batista G, Wang X, Keogh E. A complexity-invariant distance measure for time series[C]// Proceedings of the eleventh SIAM conference on data mining (SDM 2011). 2011:699-710

[7] Berndt D J, Clifford J. Using dynamic time warping to find patterns in time series[C]// KDD Workshop. 1994:359-370

[8] Breiman L. Random forests [J]. Machine Learning, 2001, 45(1): 5-32

[9] Cryer J D, Chan K. 时间序列分析及应用 [M]. 潘红宇, 等译. 北京: 机械工业出版社, 2011

[10] Buza K. Fusion methods for time-series classification[D]. University of Hildesheim, Germany, 2011

[11] Chan K, Fu A W. Efficient time series matching by wavelets [C]// Proceedings of the 15th International Conference on Data Engineering (ICDE 1999). 1999:126-133

[12] Cheng H, Yan X, Han J, et al. Discriminative frequent pattern analysis for effective classification[C]// Proceedings of the 23rd International Conference on Data Engineering (ICDE 2007). 2007:716-725

[13] Cheng H, Yan X, Han J, et al. Direct discriminative pattern mining for effective classification[C]// Proceedings of the 24th International Conference on Data Engineering (ICDE 2008). 2008:169-178

[14] Chiu B, Keogh E, Lonardi S. Probabilistic discovery of time series motifs[C]// Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 2003). 2003:493-498

[15] Cong G, Tan K, Tung A, et al. Mining top-k covering rule groups for gene expression data[C]// Proceedings of the ACM SIGMOD International Conference on Management of Data. 2005:670-681

[16] Das G, Lin K, Mannila H, et al. Rule discovery from time series [C]// Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD98). 1998:16-22

[17] Deng K, Moore A W, Nechyba M C. Learning to recognize time series: combining ARMA models with memory-based learning [C]// Proceedings of the IEEE International Symposium on Computational Intelligence in Robotics and Automation. 1997: 246-251

[18] Deng H, Runger G, Tuv E, et al. A time series forest for classification and feature extraction[J]. Information Sciences, 2013, 239:142-153

[19] Deshpande M, Kuramochi M, Karypis G. Frequent sub-structure-based approach for classification chemical compounds[C]// Proceedings of the IEEE International Conference on Data Mining (ICDM 2003). 2003:35-42

[20] Ding H, Trajcevski G, Scheuermann P, et al. Querying and mining of time series data: experimental comparison of representations and distance measures[C]// Proceedings of the 34th International Conference on Very Large Data Bases (VLDB 2008). 2008:1542-1552

[21] Douzal-Chouakria A, Amblard C. Classification trees for time series[J]. Pattern Recognition, 2012, 45:1076-1091

[22] Esling P, Agon C. Time-series data mining[J]. ACM Computing Surveys, 2012, 45(1):12

[23] Faloutsos C, Ranganathan M, Manolopoulos Y. Fast subsequence matching in time-series databases[C]// Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data. 1994:419-429

[24] Fu A, Keogh E, Lau L, et al. Scaling and time warping in time series querying[J]. The VLDB Journal, 2008, 17(4):899-921

[25] Fu T. A review on time series data mining[J]. Engineering Applications of Artificial Intelligence, 2011, 24:164-181

[26] Hartmann B, Link N. Gesture recognition with inertial sensors and optimized dtw prototypes[C]// Proceedings of IEEE International Conference on Systems Man and Cybernetics (SMC). 2010:2102-2109

[27] Hills J, Lines J, Baranauskas E, et al. Time series classification with shapelets [J]. Data Mining and Knowledge Discovery, 2013, 27(1)

[28] Hu M, Liu B. Opinion feature extraction using class sequential rules[C]// AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. 2006:61-66

[29] Jeong Y, Jeong M, Omitaomu O. Weighted dynamic time warping for time series classification[J]. Pattern Recognition, 2011, 44 (9):2231-2240

[30] Kalpakis K, Gada D, Andputtagunta V. Distance measures for effective clustering of ARIMA time series[C]// Proceedings of the IEEE International Conference on Data Mining (ICDM 2001). 2001:273-280

[31] Keogh E, Pazzani M. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback[C]// Proceedings of the 4th International Conference of Knowledge Discovery and Data Mining. 1998:239-241

[32] Keogh E, Chakrabarti K, Pazzani M, et al. Dimensionality reduction for fast similarity search in large time series databases[J]. Knowledge for Information System, 2001, 3(3):263-286

[33] Keogh E, Chu S, Hart D, et al. An online algorithm for segmenting time series[C]// Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM 2001). 2001:289-296

[34] Keogh E, Kasetty S. On the need for time series data mining benchmarks: a survey and empirical demonstration [J]. Data Mining and Knowledge Discovery, 2003, 7(4):349-371

[35] Keogh E, Ratanamahatana C A. Exact indexing of dynamic time warping[J]. Knowledge and Information Systems, 2004, 7(3): 358-386

[36] Li W, Han J, Pei J. CMAR: Accurate and efficient classification based on multiple-class association rules[C]// Proceedings of the 2001 IEEE International Conference on Data Mining. 2001:369-376

[37] Liabotis I, Theodoulidis B, Saraee M. Improving similarity search in time series using wavelets[J]. International Journal of Data Warehousing and Mining, 2006, 2 (2):1116-1137

[38] Lin J, Keogh E, Lonardi S, et al. Finding motifs in time series [C]// Proceedings of 2nd Workshop on Temporal Data Mining at KDD. 2002:53-68

[39] Lin J, Keogh E J, Wei L, et al. Experiencing SAX: a novel symbolic representation of time series[J]. Data Mining Knowledge Discovery, 2007, 15(2): 107-144

[40] Lin T, Kaminski N, Bar-Joseph Z. Alignment and classification of time series gene expression in clinical studies[J]. Bioinformatics, 2008, 24(13): 147-155

[41] Lines J, Bagnall A, Caiger-Smith P, et al. Classification of household devices by electricity usage profiles[C]// Proceedings of the 12th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2011). 2011: 403-412

[42] Lines J, Davis L M, Hills J, et al. A shapelet transform for time series classification[C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2012). 2012: 289-297

[43] Liu B, Hsu W, Ma Y. Integrating classification and association rule mining[C]// Proceedings of the Fourth International Conference on Knowledge Disco-very and Data Mining (KDD 1998). 1998: 80-86

[44] McGovern A, Rosendahl D, Brown R, et al. Identifying predictive multi-dimensional time series motifs: an application to severe weather prediction[J]. Data Mining and Knowledge Discovery, 2011, 22: 232-258

[45] Mueen A, Keogh E, Zhu Q, et al. Exact discovery of time series motifs[C]// Proceedings of the SIAM International Conference on data mining (SDM 2009). 2009: 473-484

[46] Mueen A, Keogh E, Shamlo N. Finding time series motifs in disk-resident data[C]// Proceedings of the 9th IEEE International Conference on Data Mining (ICDM 2009). 2009: 367-376

[47] Mueen A, Keogh E J, Young N. Logical-shapelets: an expressive primitive for time series classification[C]// Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2011). 2011: 1154-1162

[48] Nanopoulos A, Alcock R, Andmanolopoulos Y. Feature-based classification of time-series data[J]. International Journal of Computer Research, 2001, 10: 49-61

[49] Popivanov I, Miller R J. Similarity search over time-series data using wavelets[C]// Proceedings of the 18th International Conference on Data Engineering (ICDE 2002). 2002: 212-221

[50] Povinelli R J, Johnson M T, Lindgren A C, et al. Time series classification using Gaussian mixture models of reconstructed phase spaces[J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 6(16): 779-783

[51] Rakthanmanon T, Campana B, Mueen A, et al. Searching and mining trillions of time series subsequences under dynamic time warping[C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2012). 2012: 262-270

[52] Rakthanmanon T, Keogh E. Fast shapelets: a scalable algorithm for discovering time series shapelets[C]// Proceedings of the 13th SIAM International Conference on Data Mining (SDM13). 2013: 668-676

[53] Ratanamahatana C A, Keogh E, Bagnall A J, et al. A novel bit level time series representation with implications for similarity search and clustering[C]// Proceedings of 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD 2005). 2005: 771-777

[54] Ting J, Fu T C, Chung F L. Mining of stock data: intra- and inter-stock pattern associative classification[C]// Proceedings of the 2006 International Conference on Data Mining (ICDM 2006). 2006: 30-36

[55] Veloso A, Meira W, Zaki M J. Lazy associative classification[C]// Proceedings of the 6th International Conference on Data mining (ICDM 2006). 2006: 645-654

[56] Wang J, Karypis G. HARMONY: Efficiently mining the best rules for classification[C]// Proceedings of the Fifth SIAM International Conference on Data Mining. 2005: 205-216

[57] Wang X, Mueen A, Ding H, et al. Experimental comparison of representation methods and distance measures for time series data[J]. Journal of Data Mining and Knowledge Discovery, 2013, 26: 275-309

[58] Xi X, Keogh E, Shelton C, et al. Fast time series classification using numerosity reduction[C]// Proceedings of the 23th International Conference on Machine Learning (ICML2006). 2006: 1033-1040

[59] Xing Z, Pei J, Keogh E J. A brief survey on sequence classification[J]. SIGKDD Explorations, 2010, 12(1): 40-48

[60] Xing Z, Pei J, Yu P, et al. Extracting interpretable features for early classification on time series[C]// Proceedings of the 11th SIAM International Conference on Data Mining (SDM 2011). 2011: 247-258

[61] Xing Z, Pei J, Yu P. Early classification on time series [J]. Knowledge-based Information Systems, 2012, 31(1): 105-127

[62] Yamada Y, Suzuki E, Yokoi H, et al. Decision-tree induction from time-series data based on a standard-example split test[C]// Proceedings of the Twentieth International Conference (ICML 2003). 2003: 840-847

[63] Yan X F, Han J W. gSpan: Graph-based substructure pattern mining[C]// Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002). 2002: 721-724

[64] Ye L, Keogh E J. Time series shapelets: a new primitive for data mining[C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2009). 2009: 947-956

[65] Ye L, Keogh E J. Time series shapelets: a novel technique that allows accurate, interpretable and fast classification[J]. Data Mining and Knowledge Discovery, 2011, 22(1/2): 149-182

[66] Yin X, Han J. CPAR: Classification based on predictive association rules[C]// Proceedings of the SIAM International Conference on Data Mining. 2003: 369-376

[67] Zakaria J, Mueen A, Keogh E. Clustering time series using unsupervised-shapelets[C]// Proceedings of the 12th IEEE International Conference on Data Mining (ICDM 2012). 2012: 785-794

[68] Zhong S, Andghosh J. HMMs and coupled HMMs for multi-channel EEG classification[C]// Proceedings of the IEEE International Joint Conference on Neural Networks. 2002: 1154-1159