

基于 shapelet 的时间序列分类研究

闫汶和 李桂玲

(中国地质大学(武汉)计算机学院 武汉 430074)

摘 要 时间序列是随时间次序变化的高维实值数据,广泛存在于医学、金融、监控等领域。因为传统的分类算法在时间序列上的分类效果不佳且不具备可解释性,而 shapelet 为时间序列中最具辨别性的连续子序列,具有可解释性,所以基于 shapelet 的时间序列分类已成为时间序列分类研究的热点之一。首先,通过归纳总结,将现有的时间序列 shapelet 发现算法分为空间搜索发现 shapelet 和目标函数优化学习 shapelet 两类,并介绍了 shapelet 的相关应用;然后,从分类的对象出发,重点阐述了基于 shapelet 的一元时间序列和多元时间序列的分类算法;最后,指出了基于 shapelet 的时间序列分类在未来的研究方向。

关键词 时间序列,分类,特征提取,shapelet

中图法分类号 TP391 文献标识码 A DOI 10.11896/j.issn.1002-137X.2019.01.005

Research on Time Series Classification Based on Shapelet

YAN Wen-he LI Gui-ling

(School of Computer Science,China University of Geosciences,Wuhan 430074,China)

Abstract Time series is high-dimensional real-value data changing with time order, and it appears extensively in the fields of medicine, finance, monitoring and others. Because the accuracy of conventional classification algorithms is not ideal for the time series and it doesn't possess the characteristic of interpretability, and shapelet is a discriminative continuous time-series subsequence, the time series classification based on shapelet has become one of the hot spots in the researches on time series classification. First, through analyzing the existing time series shapelet discovery methods, this paper classified them into two catalogues, namely shapelet discovery from shapelet candidates and learning shapelet by optimizing object function, and introduced the application of shapelet. Then, according to the classification object, this paper emphasized the univariate time series classification algorithms and multivariate time series classification algorithms based on shapelet. Finally, this paper pointed out the further research direction of time series classification based on shapelet.

Keywords Time series, Classification, Feature extraction, Shapelet

1 引言

时间序列是按照某种数据采集手段,在相同的时间间隔内获得的高维数值型序列。时间序列数据在实际领域中被广泛应用。例如,在金融领域,股票价格的变动是时间序列,可通过对股票时间序列进行分析来预测股票未来的走势,图 1 给出股票的时间序列;在医疗领域,心电图是时间序列,医生通过对病人的心电图时间序列进行监控,可及时发现病人的心脏疾病,并及时展开治疗,图 2 给出心电图的时间序列;在图像识别领域,可将图像转化为时间序列后再做进一步的分析处理,图 3 展示了将鸡蛋图像转化为时间序列的过程,得到图像的时间序列后,可根据不同类别图像的特点对图像进行分类。

时间序列能随着时间延长不断增长和动态更新,因此得到的时间序列维度高、数据量大,对时间序列数据进行分析、处理、应用时,使用传统的数据挖掘技术代价巨大。时间序列分类是时间序列数据挖掘的任务之一,其目标是通过训练集得到具有区分性的特征,在分类时,分类器根据提取的特征进行分类。与传统分类问题不同,时间序列强调了属性之间的次序,不同的属性次序会产生不同的特征提取结果。鉴于普通的分类方法不具备可解释性,Ye 和 Keogh^[1]提出了 shapelet 的概念,shapelet 是时间序列中最能代表该类时间序列的一段连续子序列,可以很好地解释分类结果。此外,普通的算法只能提供分类结果,不能说明类别间的差异,且不能了解数据的内部特性,但 shapelet 可以说明数据的差异性。除了以上优点外,一个高维时间序列在经过缩放变换之后,其基本特

到稿日期:2017-12-05 返修日期:2018-03-05 本文受国家自然科学基金(61702468),中国地质大学(武汉)教学实验室开放基金(SKJ2018286)资助。

闫汶和(1995—),男,硕士生,CCF 会员,主要研究方向为数据挖掘和机器学习、时间序列数据管理,E-mail:cugywh@163.com;李桂玲(1979—),女,副教授,硕士生导师,CCF 会员,主要研究方向为数据挖掘与机器学习、时间序列数据管理,E-mail:guiling@cug.edu.cn(通信作者)。

性不会发生很大的变化,因此 shapelet 在不同的维度下都能使用。

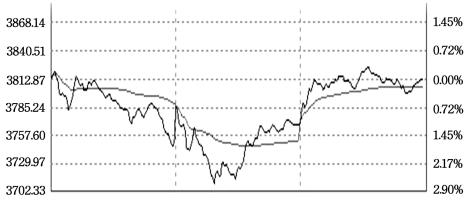


图 1 股票时间序列

Fig. 1 Stock time series

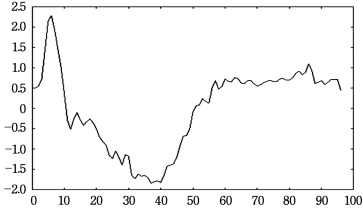


图 2 心电图时间序列

Fig. 2 ECG time series

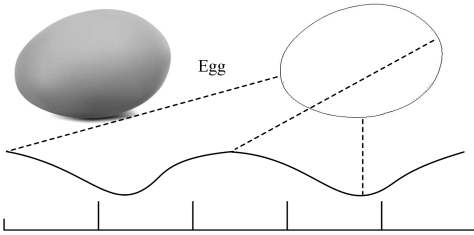


图 3 鸡蛋时间序列

Fig. 3 Egg time series

本文将介绍基于 shapelet 的时间序列分类方法,第 2 节给出了基于 shapelet 的时间序列分类中的相关定义;第 3 节对近年来的时间序列 shapelet 发现方法进行归纳,并介绍 shapelet 的应用;第 4 节重点介绍基于 shapelet 的一元时间序列分类和多元时间序列分类算法;第 5 节对今后的研究方向进行展望;最后进行总结。

2 相关定义

本节介绍文中所使用到的相关符号和定义。

定义 1(一元时间序列) 一元时间序列 $T = \{t_1, t_2, t_3, \dots, t_m\}$ 是一个长度为 m 的实值连续数据序列。其中, $t_1, t_2, t_3, \dots, t_m$ 按照时间的先后顺序排列, t_i 表示时间序列 T 的第 i 个点。

定义 2(一元时间序列数据集) D 为一个一元时间序列数据集,其中包含了 n 条一元时间序列, $D = \{T_1, T_2, \dots, T_n\}$, T_i 表示一元时间序列数据集的第 i 条时间序列。

定义 3(时间序列子序列) S 是时间序列中长度为 l 的一段连续子序列(Subsequence),记为 $S = \{t_p, \dots, t_{p+l-1}\}$, $1 \leq p \leq m-l+1$ 。

定义 4(时间序列相似性度量) 给定 2 个长度都为 m 的时间序列 T 和 R ,这两条时间序列之间的相似性度量表示为 $Dist(T, R)$,如式(1)所示:

$$Dist(T, R) = \sqrt{\frac{1}{m} \sum_{i=1}^m (x_i - y_i)^2} \quad (1)$$

定义 5(时间序列和子序列的距离) 完整的时间序列和子序列之间的距离可以利用 $SubDist(T, S)$ 计算,其中 $S_T^{|S|}$ 是序列 T 中所有长度为 $|S|$ 的子序列的集合。

$$SubDist = \min(Dist(S', S)), S' \in S_T^{|S|} \quad (2)$$

定义 6(熵) 一个时间序列数据集 D 中有 A 和 B 两类时间序列,类别 A 的概率为 $p(A)$,类别 B 的概率为 $p(B)$,则 D 的熵表示为:

$$Ent(D) = -p(A) \log p(A) - p(B) \log p(B) \quad (3)$$

定义 7(信息增益) 假设数据集 D 被划分为 2 个数据子集 D_1 和 D_2 ,则它的信息增益可表示为:

$$Gain = Ent(D) - \frac{|D_1|}{|D|} Ent(D_1) - \frac{|D_2|}{|D|} Ent(D_2) \quad (4)$$

定义 8(最佳分割点, Optimal Split Point, OSP) 计算 shapelet 和数据集中每一条时间序列的相似性,对结果进行非递减排序后会得到一个 order list,把 order list 中每两个相邻点的中间点当作分割点,然后计算这个分割点的信息增益,信息增益最大的分割点就是 OSP。由图 4 可以看出,第一个分割点的分割效果比第二个分割点的要好,因此第一个分割点就是当前的最佳分割点。

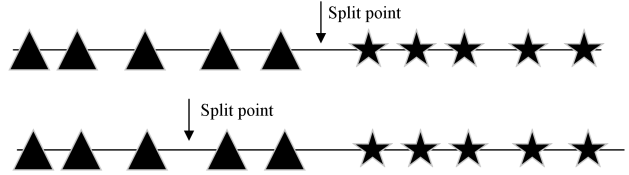


图 4 最佳分割点

Fig. 4 Optimal split point

定义 9(shapelet) shapelet 是时间序列中最能代表该类时间序列的一段连续子序列,每一个 shapelet 有一个自己的距离阈值 δ , δ 的取值为 OSP 的值。shapelet 用公式可以表示为:

$$Gain(shapelet(D), d_{(OSP(D, shapelet(D)))}) \geq Gain(S, d_{OSP(D, S)}) \quad (5)$$

图 5 给出不同类别曲线的 shapelet(实线部分)。

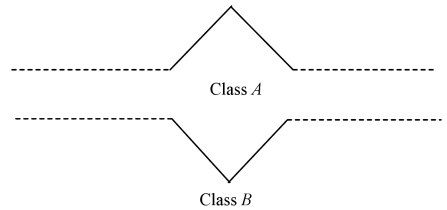


图 5 不同类别曲线的 shapelet

Fig. 5 Shapelet in different series

定义 10(多元时间序列) 多元时间序列是由多个一元时间序列组成的向量,记作 $\mathcal{T} = \{T_1, T_2, T_3, \dots, T_r\}$,其中 r 是多元时间序列的维度,每一条一元时间序列 T_i 的长度为 m 。

定义 11(多元时间序列数据集) $\mathcal{D} = \{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_s\}$, $\mathcal{D}$ 中有 s 条多元时间序列数据,每一条多元时间序列有相同的

维度, $\mathcal{D}_i^j$ 表示第 i 条多元时间序列的第 j 维数据。

3 时间序列 shapelet 的发现

shapelet 发现的目标是找到质量最佳的 shapelet, 而 shapelet 的质量就是子序列的可辨别性。在训练过程中得到的 shapelet 的质量越好, 分类准确率就越高。现有的 shapelet 发现算法可大致归为两类, 第一类是在 shapelet 候选者空间找到最好的 shapelet; 第二类是通过优化目标函数得到最好的 shapelet。

3.1 空间搜索发现 shapelet

时间序列数据集中的每一个时间序列的子序列都可能是最佳的 shapelet, 把每一个子序列看作一个 shapelet 候选者, 那么所有子序列就构成了 shapelet 候选者空间。Ye 和 Keogh^[1]首次基于 shapelet 候选者空间提出 shapelet 发现算法, 在候选者空间中找到使信息增益最大的 shapelet 候选者, 但此方法的时间复杂度对数据的维度和数据集的大小都很敏感, 仅适合序列长度较小且序列个数不多的数据集。

时间序列的维度高、数据量大, 在原始的时间序列上发现 shapelet 的时间复杂度很高。为了提高 shapelet 发现的速度, Rakthanmanon 等人^[2]提出了一种新的方法, 通过符号聚集近似(Symbolic Aggregate Approximation, SAX)对数据降维并将数据转换到符号空间中, 然后搜寻最佳的 shapelet, 该方法极大地加快了 shapelet 发现的速度。李祯盛等人^[3]随后提出了基于主成分分析(Principal Component Analysis, PCA)的 shapelet 发现方法, 通过 PCA 算法降低了数据的维度, 从而加快了运算速度。但是, 降维的方法会丢失原始数据中的有用信息, 影响分类准确率, 于是, Chang 等人^[4]使用并行计算的方法发现 shapelet, 从计算机硬件的角度出发, 提升了算法的效率。

因为 shapelet 候选者空间巨大, 有许多冗余的 shapelet, 所以可利用剪枝方法来降低算法的时间复杂度。He 等人^[5]先将数据符号化, 然后仅把出现频率较低的子序列看作候选者, 避免了许多无用计算, 加快了算法的速度。随后, Wei 等人^[6]通过一种基于最大相关性和最小冗余度的特征选择策略, 消除了部分相似 shapelet。Zhang 等人^[7]提出了一种带有时间序列关键点的数据预处理方法, 即缩小了 shapelet 候选者空间, 提高了算法的效率。另外, Ji 等人^[8]提出了基于重要数据点的快速 shapelet 发现算法, 仅把包含重要数据点的子序列看作候选者 shapelet, 剪去了大量无用的 shapelet。

与上述方法不同的是, Karlsson 等人^[9]利用随机的思想, 从随机选取的样本中随机选择 shapelet。随后, Renard 等人^[10]提出了一种新的随机方法, 即通过一个随机函数来决定当前 shapelet 是否应该被评价, 降低了算法的时间复杂度。虽然随机的方法避免了大量的计算, 但此方法得到的 shapelet 在可解释性方面存在不足。

相比一元时间序列, 从多元时间序列中发现 shapelet 则更为困难。Cetin 等人^[11]在 shapelet 发现过程中动态改变 shapelet 的长度, 缩小了 shapelet 的候选者空间, 加快了多元

时间序列 shapelet 发现的速度。另外, Wistuba 等人^[12]提出了一种极速 shapelet 发现算法, 虽然此方法使用了一些随机选取的 shapelet, 但分类准确率与一流的基于 shapelet 的分类器相当。为了提高 shapelet 的可解释性, Lin 等人^[13]在 shapelet 的选择过程中考虑了类别属性, 因此能得到质量更好的 shapelet。

3.2 目标函数优化学习 shapelet

通过目标函数来发现 shapelet 是一种新的思路, 将传统的 shapelet 发现过程转化为数学问题中的优化求解过程。这种方法虽然速度比较快, 但是需要设置很多参数, 在应对不同问题时需要根据实际情况选择最优参数。

Grabocka 等人^[14]第一次提出通过目标函数求解最佳的 shapelet, 相比于原始的 shapelet 发现算法, 此算法降低了时间复杂度。鉴于文献[14]的方法不能学习到随着时间轴弯曲的时间序列的 shapelet, Grabocka 等人^[15]又提出了基于动态时间规整(Dynamic Time Warping, DTW)的学习 shapelet 的方法, 但是该方法相对于文献[14]的方法在时间复杂度上有所增加。之后, Hou 等人^[16]结合广义特征向量, 设计了新的目标函数, 通过确定 shapelet 的位置来发现 shapelet, 此方法在准确率上稍逊于文献[14]的算法, 但速度更快, 在大的数据集上的效果也很好。另外, Yang 等人^[17]通过增量式神经网络的方法学习 shapelet, 能较快地学习到好的 shapelet。

3.3 shapelet 的应用

shapelet 是时间序列的连续子段, 是时间序列特有的局部特征。目前, shapelet 可以用在光谱分析、动作识别和心电信号处理等领域中的数据降维、分类和聚类方面。

在降维方面, 因为 shapelet 是能代表该类时间序列的子序列, 所以可以把 shapelet 看作时间序列降维处理的结果。此外, 由于时间序列具有顺序性, 可以在获取多个 shapelet 的同时计算序列与这些 shapelet 的距离, 这些距离组成的新序列就是降维处理的结果, 这种降维方式打破了时间序列顺序性的限制, 因此可以使用传统的算法对数据做进一步分析。

在分类方面, 使用 shapelet 让结果更具解释性, 使研究员更了解数据。shapelet 作为一种局部特征, 可以保证分类的结果更准确, 且具有一定的抗噪能力。在速度上, shapelet 的分类速度也具有一定的优势。

在聚类方面, 传统的时间序列聚类方法要求序列等长, 依赖序列到整个数据集的距离计算, 且易受噪声的干扰。使用 shapelet 能对不等长的序列进行聚类, 消除部分噪声的影响, 提高聚类的效果。

shapelet 不仅可以用于静态时间序列的挖掘任务, 也可以用于动态环境, 即时间序列流。时间序列流是一种动态的时间序列, 流中的元素是按时间顺序的、快速变化的、海量的和潜在无限的^[18]。

朱利鹏等人^[19]利用时间序列 shapelet 研究基于大数据的暂态电压稳定评估, 从故障后相量测量单元(Phasor Measurement Unit, PMU)测得的动态序列中提取 shapelet 特征, 并将其用于在线监测及稳定评估, 当判断系统失稳时, 可以及

时采取措施使电压恢复稳定。Mousheimish 等人^[20]利用时间序列 shapelet 对复杂事件处理 (Complex Event Processing, CEP) 中的事件流自动学习预测规则。

4 基于 shapelet 的时间序列分类

时间序列分类问题是众多分类问题的子问题,在信息发展如此迅速的今天,对时间序列数据进行挖掘显得尤为重要。时间序列分类问题广泛存在于心电图分类^[21]、地震波分析^[22]、动作识别^[23]、图像识别^[23]等领域。为促进时间序列的研究,Keogh 等人收集了一个时间序列的 UCR 数据集¹⁾,UCR 数据集中的大部分数据是一元时间序列,随着研究的深入,Karlsson 等人还专门收集了用于多元时间序列数据挖掘的数据集²⁾。

在基于 shapelet 的时间序列分类问题中,对时间序列进行分类时,需计算 shapelet 与待分类时间序列的距离,当距离小于 shapelet 的距离阈值时,这条时间序列就被归为 shapelet 所属的类别,否则被归为其他类。因为时间序列是一个有序的序列,它依赖于属性的次序,当属性的次序发生改变时,时间序列就会变化,从而影响 shapelet 的发现。shapelet 代表了某一个类别的时间序列的重要特征,使结果具有可解释性,并且使用 shapelet 分类的准确率不低于传统分类方法,因此基于 shapelet 的时间序列分类成为近几年的研究热点之一。

时间序列分为一元时间序列和多元时间序列,基本定义见第 2 节,下面将详细介绍针对这两种时间序列的分类算法。

4.1 一元时间序列分类

时间序列是一种高维的数据序列,处理过程中的计算代价非常高,最初针对时间序列分类的研究主要集中在一元时间序列。一元时间序列分类器的构造方法又分为两类,第一类是将 shapelet 发现与分类器的构造相结合,第二类是将 shapelet 发现与分类器的构造相分离,由此,可将一元时间序列的分类算法分为两类。

4.1.1 shapelet 发现与分类器的构造相结合

shapelet 发现与分类器的构造相结合就是在发现 shapelet 的同时构造分类器。Ye 等人^[1]最先提出了与 1NN-Euclidean 和 1NN-DTW 准确率相当的基于 shapelet 的决策树的分类方法,通过暴力搜索发现 shapelet,然后根据 shapelet 构造决策树,最后利用决策树进行分类。鉴于暴力搜索算法的时间复杂度太高,其还提出了两种加速策略,即提前终止和熵剪枝,但是这种方法构造分类器的时间代价很大,不适合大的数据集。由于单个 shapelet 无法有效区分相似类别的序列,Mueen 等人^[24]提出了逻辑 shapelet 的方法,通过逻辑操作将多个 shapelet 组合起来,提高了分类的准确率;并提出了计算重用和剪枝策略,加快了算法的速度。但是,此方法处理多分类问题的效果不理想。

在分类模型的构造上,王树英等人^[25]提出一种增量式的方法,动态地选择 shapelet 并调整已经构造好的决策树,节省

了每次重新构造决策树的开销。Gordon 等人^[26]提出一种随机模型产生方法,该方法可以快速收敛并且可以用于时间序列分类中。在此之后,Gordon 等人^[27]又提出了一种快速并且内存占用较小的分类方法,其能快速、准确地构建分类模型。

质量评价是分类中的重要环节,Lines 等人^[28]使用 3 种不同的 shapelet 质量评估标准选取 shapelet,然后对构造的分类器进行性能评估。鉴于大多数分类方法并未考虑变量之间的联系,Yan 等人^[29]提出了基于趋势特征表示的 shapelet 分类方法,该方法考虑了序列的变化趋势,分类结果更为准确。在相似性度量方面,Chouakri 等人^[30]提出了一种新的时间序列相似性度量方法,整合了趋势相似与数值相似,并采用新的度量方法建立决策树。因为曼哈顿距离考虑了数据项之间的相关性,并且具有尺度不变性,所以 Arathi 等人^[31]使用曼哈顿距离来度量序列间的相似性,提高了分类的准确率,并且采用悲观统计相关性测试设计了悲观剪枝策略,提高了算法的速度。

单一的分类器总是存在某方面的不足,于是研究者提出了集成的方法。为了缩短训练时间,Mitzev 等人^[32]通过组合多个分类器的决策方式来对时间序列进行分类。另外,Karlsson 等人^[9]提出通过 shapelet 森林进行分类,并且还提出了一种可以找出时间序列中比较重要的区域的方法。之后,Deng 等人^[33]提出了 Entrance Gain 分裂评估指标,提高了时间序列森林的分类准确率。Raza 等^[34]通过集成多个基于 shapelet 的分类器来构造一个新的分类器,并用新的分类器来对序列进行分类。其中,每一个基于 shapelet 的分类器都是通过对 shapelet 候选者进行随机抽样而构造生成的。由于分类结果具有多样性,因此集成分类器的方法提高了分类器的准确性,并且降低了计算成本。

早期分类是在获得部分时间序列数据时就可以提前对其分类,而不必等到收集到完整的时间序列数据时再分类。Xing 等人^[35]提出了从 shapelet 的早期性和准确率两方面来选取 shapelet,使分类器具有较好的早期性和准确率,并设计了两种计算距离阈值的方法和一种加速技术。之后,Ghalwash 等人^[36]引入了不确定性参数,并且提出了新的 shapelet 剪枝方法,该算法的准确率与文献^[35]的算法相当,但早期性更好。Karlsson 等人^[37]提出了用于早期分类的随机森林的方法,当暂时得到的数据序列的长度小于 shapelet 的长度而无法计算距离时,其提出了两种不同的处理策略,并通过集成的方法提高了早期分类的准确率。

4.1.2 shapelet 发现与分类器的构造相分离

shapelet 发现与分类器的构造相分离就是先利用得到的 shapelet 对时间序列进行数据空间的转换,然后构造分类器,转换后的数据可以结合传统的分类策略,提高了算法的灵活性。Lines 等人^[38]首次提出 shapelet 转换的方法,此方法在保留 shapelet 解释性的同时使数据更具灵活性,并且算法的

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

2) <https://people.dsv.su.se/~isak-kar/grsf/#sec-1-2>

速度比传统的决策树的方法更快。该算法虽然有着突出的优点,但发现 shapelet 的过程非常耗时,而且 shapelet 转换的方法并不适合所有问题。之后,Bostrom 等人^[39]提出了一种快速 shapelet 转换方法和一种加速技术,该方法在速度上比原始的 shapelet 转换算法有了较大的提升,且准确率相当。Bostrom 等人^[40]还提出了针对多分类问题的 shapelet 转换方法和一种新的加速技术,使得多分类问题的准确率和搜寻的速度均得到提升。

在 shapelet 转换中如何确定用于 shapelet 转换的 shapelet 数量是一个重要问题,数量太多会使模型变得复杂,数量太少会降低数据的质量,进而影响分类的结果。原继东等人^[41]提出了基于 shapelet 剪枝和覆盖的时间序列分类算法,即在训练阶段去除掉相似的 shapelet,并提出了可以确定用于转换的 shapelet 的数量的方法。Zalewski 等人^[42]利用 shapelet 转换构建了一个用于时间序列分类的符号模型,并通过特征选择提高了 shapelet 转换表示方法的质量,这种方法避免了估计最小长度、最大长度和用于 shapelet 转换的 shapelet 的数量,在一定程度上简化了算法模型。

在 shapelet 转换过程中,传统的算法直接使用选取的 shapelet 而没有考虑 shapelet 之间的逻辑关系,从而降低了分类的准确率,因此原继东等人^[43]提出了基于 shapelet 连接性和分离性的 shapelet 转换方法,提高了算法的性能。为了使转换后的数据更具表达力,孙其法等人^[44]提出多样化的 top- k 的 shapelet 转换方法,去除了相似的 shapelet。另外,Yan 等人^[45]将超限学习机用于时间序列 shapelet 的转换中,提高了分类器的效率和有效性。

4.2 多元时间序列分类

在医学、生物信息学等领域,需同时收集多条数据序列,这样就生成了多元时间序列。在心脏疾病监测中,通过多个传感器收集病人的心跳信息,图 6 给出一个二维的多元时间序列。多元时间序列的数据量更大,并且不同维度的序列之间会互相影响,加大了分类的难度。

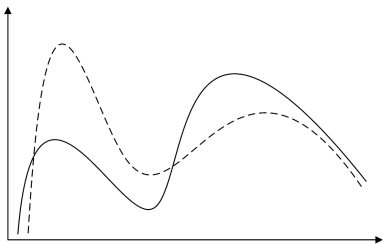


图 6 二维的多元时间序列

Fig. 6 Two-dimensional multivariate time series

Karlsson 等人^[21]通过分割和组合的方法对多元时间序列进行分类,并提出几种组合策略,但使用此方法时需要选择合适的基准分类器和组合策略。Cetin 等人^[11]提出将每一个维度的分类器进行集成的方法,并设计了剪枝策略,加快了运算速度,并能保持较高的分类准确率。之后,He 等人^[46]在类不平衡的多元时间序列早期分类问题上提出了新的解决方法,即通过欠采样和动态子空间生成的方法得到初始的训练

数据,最后通过组合分类器的方法进行分类。为了提高效率,Grabocka 等人^[47]提出针对多元时间序列的快速分类算法,其通过在线聚类的方法避免了对相似的 shapelet 进行准确率评估,加快了算法的速度。另外,Wang 等人^[48]通过逻辑回归的方法学习 shapelet,并且提出了在多元时间序列上集成 boost 的分类方法。Baydogan 等人^[49]则提出了一种新的多元时间序列的符号表示方法,并且使用带有时间节点的决策树对多元时间序列进行分类。Patri 等人^[50]考虑了多元时间序列各维度之间的相互影响,并在此基础上对多元时间序列进行分类。

在多元时间序列的早期分类问题中,Ghalwash 等人^[51]第一次将隐马尔可夫模型(Hiden Markov Model, HMM)和支持向量机(Support Vector Machine, SVM)模型进行整合,提出了用于多元时间序列早期分类的模型。之后,Ghalwash 等人^[52]通过设计目标函数来选取 shapelet,然后通过不断判断当前序列是否与候选集中的 shapelet 相匹配来对序列进行分类,这种方法能在比较短的时间内得到结果,但训练的时间比较长。随后,He 等人^[53]提出了新的计算 shapelet 实用性分数的方法,此方法综合考虑了 shapelet 的早期性、查准率和查全率,并提出了两种多元时间序列分类策略。

5 进一步的研究方向

时间序列 shapelet 发现和基于 shapelet 的时间序列分类方法的研究虽然取得了一些成果,但仍然存在着不足,未来可以从如下方面进行进一步的研究。

(1)研究时间序列表示方法。时间序列的高维性是处理时间序列数据的一大挑战,需要设计一种高效的数据降维方法,在保留数据重要特性的同时降低数据的维度,加快算法的速度。可从研究时间序列中的重要特征点出发,发现时间序列中一些比较重要的特征点,将这些特征点组成新的时间序列,然后发现它们的内在特性。此外,因为时间序列具有时序性,在对时间序列进行分类时,传统分类器的效果有限,所以可以研究数据转换的方法,将时间序列数据转化为普通的数据,以便进一步处理。

(2)改进基于 shapelet 的时间序列分类算法。可研究 shapelet 的发现方法,加速 shapelet 的发现速度,提高发现的 shapelet 的质量;可研究将 shapelet 融合到其他传统分类器中,尝试不同分类器的分类效果;可研究 shapelet 之间的不同之处。此外,还可研究算法中的参数,如最小长度、最大长度、质量评价标准对 shapelet 选取的影响。

(3)研究集成的分类方法。不同的分类方法具有不同的优点和缺点,单一的算法往往只在具有一定特点的数据集上具有较好的效果,可以将不同的算法进行集成,找出一种比较通用的分类算法。

(4)研究早期分类问题。以前的时间序列分类问题研究主要是基于完整的时间序列进行的,对早期分类问题的研究较少,而当今处于流媒体时代,对流数据进行早期分类十分重要,有着很重要的应用价值。如果能够在越早的时间对目标

进行分类,就能越早地发现目标中的价值,因此早期分类的研究有着重要的发展潜力。

(5)研究多元时间序列的分类问题。多元时间序列目前多用于医疗、生物信息学方面,一元时间序列的分类算法并不都适合于多元时间序列的分类,因此需要研究新的分类算法。

结束语 本文首先对 shapelet 的发现方法进行分类,然后介绍了基于 shapelet 的一元时间序列分类方法和基于 shapelet 的多元时间序列分类方法。在实际应用中,现有的方法仍存在不足之处。如何加快 shapelet 的发现速度,如何提高分类器在时间序列多分类问题上的准确率等,都需要更深入的研究。

参 考 文 献

- [1] YE L,KEOGH E. Time series shapelets:a new primitive for data mining[C]//ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009:947-956.
- [2] RAKTHANMANON T,KEOGH E. Fast shapelets: A scalable algorithm for discovering time series shapelets[C]//Proceedings of the 2013 SIAM International Conference on Data Mining, 2013:668-676.
- [3] LI Z S,HE Z F. Time Series Shapelet Extraction Based on Principal Component Analysis[J]. Computer Systems & Applications, 2014, 23(11):145-149. (in Chinese)
李祯盛,何振峰. 基于主成分分析的时间序列 Shapelet 提取方法[J]. 计算机系统应用, 2014, 23(11):145-149.
- [4] CHANG K,DEKA B,HWU W M W, et al. Efficient Pattern-Based Time Series Classification on GPU[C]//IEEE International Conference on Data Mining. IEEE Computer Society, 2012:131-140.
- [5] HE Q,DONG Z,ZHUANG F, et al. Fast Time Series Classification Based on Infrequent Shapelets[C]//International Conference on Machine Learning and Applications. IEEE, 2013:215-219.
- [6] WEI Y,JIAO L,WANG S, et al. Time Series Classification with Max-Correlation and Min-Redundancy Shapelets Transformation[C]//International Conference on Identification, Information, and Knowledge in the Internet of Things. IEEE, 2016:7-12.
- [7] ZHANG Z,ZHANG H,WEN Y, et al. Accelerating time series shapelets discovery with key points[C]//Asia-Pacific Web Conference. Springer International Publishing, 2016:330-342.
- [8] JI C,ZHAO C,PAN L, et al. A Fast Shapelet Discovery Algorithm Based on Important Data Points[J]. International Journal of Web Services Research, 2017, 14(2):67-80.
- [9] KARLSSON I,PAPAPETROU P,BOSTRÖM H. Forests of Randomized Shapelet Trees[C]//International Symposium on Statistical Learning and Data Sciences. Cham: Springer, 2015: 126-136.
- [10] RENARD X,RIFQI M,ERRAY W, et al. Random-shapelet: An algorithm for fast shapelet discovery[C]//IEEE International Conference on Data Science and Advanced Analytics. IEEE, 2015:1-10.
- [11] CETIN M,MUEEN A,CALHOUN V. Shapelet Ensemble for Multi-dimensional Time Series[C]//SDM, 2015:307-315.
- [12] WISTUBA M,GRABOCKA J,SCHMIDT-THIEME L. Ultra-fast shapelets for time series classification[J]. arXiv: 1503.05018, 2015.
- [13] LIN Y F,CHEN H, TSENG V S, et al. Reliable Early Classification on Multivariate Time Series with Numerical and Categorical Attributes[M]//Advances in Knowledge Discovery and Data Mining. Springer International Publishing, 2015:199-211.
- [14] GRABOCKA J,SCHILLING N,WISTUBA M, et al. Learning time-series shapelets[C]//ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2014: 392-401.
- [15] SHAH M,GRABOCKA J,SCHILLING N, et al. Learning DTW-shapelets for time-series classification[C]//Proceedings of the 3rd IKDD Conference on Data Science. ACM, 2016:13-16.
- [16] HOU L,ZURADA J M,ZURADA J M. Efficient learning of timeseries shapelets[C]//Thirtieth AAAI Conference on Artificial Intelligence. AAAI Press, 2016:1209-1215.
- [17] YANG Y,DENG Q,SHEN F, et al. A Shapelet Learning Method for Time Series Classification[C]//IEEE International Conference on TOOLS with Artificial Intelligence. IEEE Computer Society, 2016:423-430.
- [18] LI X G,SONG B Y,YU G, et al. Wavelet-Based pseudo period detection on time series stream[J]. Journal of Software, 2010, 21(9):2161-2172. (in Chinese)
李晓光,宋宝燕,于戈,等. 基于小波的时间序列流伪周期检测方法[J]. 软件学报, 2010, 21(9):2161-2172.
- [19] ZHU L P,LU C,HUANG H, et al. Wide-Area Time Series Data Mining Based Transient Voltage Stability Assessment[J]. Power System Technology, 2016, 40(1):180-185. (in Chinese)
朱利鹏,陆超,黄河,等. 基于广域时序数据挖掘策略的暂态电压稳定评估[J]. 电网技术, 2016, 40(1):180-185.
- [20] MOUSHEIMISH R,TAHER Y,ZEITOUNI K. autoCEP: Automatic Learning of Predictive Rules for Complex Event Processing[C]//International Conference on Service-Oriented Computing. Springer International Publishing, 2016:586-593.
- [21] KARLSSON I,PAPAPETROU P,ASKER L. Multi-channel ECG classification using forests of randomized shapelet trees[C]//ACM International Conference on Pervasive Technologies Related To Assistive Environments. ACM, 2015:1-6.
- [22] YE H C C M,ZHU Y,ULANOVA L, et al. Time series joins, motifs, discords and shapelets: a unifying view that exploits the matrix profile[J]. Data Mining & Knowledge Discovery, 2018, 32(1):83-123.
- [23] YE L,KEOGH E. Time series shapelets:a novel technique that allows accurate, interpretable and fast classification[J]. Data Mining and Knowledge Discovery, 2011, 22(1):149-182.
- [24] MUEEN A,KEOGH E, YOUNG N. Logical-shapelets: an expressive primitive for time series classification[C]//ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2011:1154-1162.

[25] DING J, WANG S Y. Incremental Time Series Classification Algorithm Based on Shapelets[J]. Computer Science, 2016, 43(5): 257-260. (in Chinese)
丁剑, 王树英. 一种使用 shapelets 的增量式时间序列分类[J]. 计算机科学, 2016, 43(5): 257-260.

[26] GORDON D, HENDLER D, ROKACH L. Fast randomized model generation for shapelet-based time series classification[J]. arXiv:1209.5038, 2012.

[27] GORDON D, HENDLER D, ROKACH L. Fast and space-efficient shapelets-based time-series classification[J]. Intelligent Data Analysis, 2015, 19(5): 953-981.

[28] LINES J, BAGNALL A. Alternative Quality Measures for Time Series Shapelets[M]// Intelligent Data Engineering and Automated Learning-IDEAL 2012. Springer Berlin Heidelberg, 2012: 475-483.

[29] YAN X, MENG F, YAN Q. Shapelet classification method based on trend feature representation[J]. Journal of Computer Applications, 2017, 37(8): 2343-2348.

[30] DOUZAL-CHOUAKRIA A, AMBLARD C. Classification trees for time series[J]. Pattern Recognition, 2012, 45(3): 1076-1091.

[31] ARATHI M, GOVARDHAN A. An Efficient and Accurate Time Series Classification Using Shapelets[J]. International Journal of Information and Electronics Engineering, 2014, 4(5): 347-353.

[32] MITZEV I, YOUNANN H. Concatenated Decision Paths Classification for Time Series Shapelets[J]. International Journal of Instrumentation and Control Systems (IJICS), 2016, 1(6): 15-25.

[33] DENG H, RUNGER G, TUV E, et al. A time series forest for classification and feature extraction[J]. Information Sciences, 2013, 239(4): 142-153.

[34] RAZA A, KRAMER S. Ensembles of Randomized Time Series Shapelets Provide Improved Accuracy while Reducing Computational Costs[J]. arXiv:1702.06712, 2017.

[35] XING Z, PEI J, YU P S, et al. Extracting Interpretable Features for Early Classification on Time Series[C]// Eleventh Siam International Conference on Data Mining. DBLP, 2011: 247-258.

[36] GHALWASH M F, RADOSAVLJEVIC V, OBRADOVIC Z. Utilizing temporal patterns for estimating uncertainty in interpretable early decision making[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2014: 402-411.

[37] KARLSSON I, PAPAPETROU P, BOSTRÖM H. Early Random Shapelet Forest[M]// Discovery Science. Springer International Publishing, 2016: 261-276.

[38] HILLS J, LINES J, BARANAUSKAS E, et al. Classification of time series by shapelet transformation[J]. Data Mining & Knowledge Discovery, 2014, 28(4): 851-881.

[39] BOSTROM A, BAGNALL A, LINES J. Evaluating Improvements to the Shapelet Transform[C]// Knowledge Discovery in Database 2016 Time Series Workshop. KDD, 2016.

[40] BOSTROM A, BAGNALL A. Binary Shapelet Transform for Multiclass Time Series Classification[M]// Big Data Analytics and Knowledge Discovery. 2015: 257-269.

[41] YUAN J D, WANG Z H, HAN M. Shapelet Pruning and Shapelet Coverage for Time Series Classification[J]. Journal of Software, 2015, 26(9): 2311-2325. (in Chinese)
原继东, 王志海, 韩萌. 基于 Shapelet 剪枝和覆盖的时间序列分类算法[J]. 软件学报, 2015, 26(9): 2311-2325.

[42] ZALEWSKI W, SILVA F, MALETZKE A G, et al. Exploring shapelet transformation for time series classification in decision trees[J]. Knowledge-Based Systems, 2016, 112: 80-91.

[43] YUAN J D, WANG Z H, HAN M. A logical shapelets transformation for time series classification[J]. Chinese Journal of Computers, 2015, 38(7): 1448-1459. (in Chinese)
原继东, 王志海, 韩萌. 基于逻辑 shapelets 转换的时间序列分类算法[J]. 计算机学报, 2015, 38(7): 1448-1459.

[44] SUN Q F, YAN Q Y, YAN X M. Diversified top-k shapelets transform for time series classification[J]. Journal of Computer Applications, 2017, 37(2): 335-340. (in Chinese)
孙其法, 闫秋艳, 闫欣鸣. 基于多样化 top-k shapelets 转换的时间序列分类方法[J]. 计算机应用, 2017, 37(2): 335-340.

[45] YAN Q, SUN Q, YAN X. Adapting ELM to Time Series Classification: A Novel Diversified Top-k Shapelets Extraction Method[C]// Australasian Database Conference. Springer International Publishing, 2016: 215-227.

[46] HE G, DUAN Y, QIAN T, et al. Early prediction on imbalanced multivariate time series[C]// Acm International Conference on Information & Knowledge Management. ACM, 2013: 1889-1892.

[47] GRABOCKA J, WISTUBA M, SCHMIDT-THIEME L. Fast classification of univariate and multivariate time series through shapelet discovery[J]. Knowledge and Information Systems, 2016, 49(2): 429-454.

[48] WANG H, WU J. Boosting for Real-Time Multivariate Time Series Classification[C]// AAAI Conference on Artificial Intelligence (AAAI). 2017: 4999-5000.

[49] 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.

[50] PATRI O P, KANNAN R, PANANGADAN A V, et al. Multivariate time series classification using inter-leaved shapelets[C]// Time Series Workshop in Neural Information Processing Systems. NIPS, 2015.

[51] GHALWASH M F, RAMLJAK D, OBRADOVIC Z. Early classification of multivariate time series using a hybrid HMM/SVM model[C]// IEEE International Conference on Bioinformatics and Biomedicine. IEEE, 2012: 1-6.

[52] GHALWASH M F, RADOSAVLJEVIC V, OBRADOVIC Z. Extraction of Interpretable Multivariate Patterns for Early Diagnostics[C]// IEEE International Conference on Data Mining. IEEE, 2013: 201-210.

[53] HE G, DUAN Y, PENG R, et al. Early classification on multivariate time series[J]. Neurocomputing, 2015, 149(PB): 777-787.