基于多目标演化算法和改进概率分类的重尾时间序列预测

邹小云[1] 林文学[2]

[1](湖北职业技术学院科研处 湖北 孝感 432000)

[2](湖北职业技术学院继续教育学院 湖北 孝感 432000)

摘 要 金融、通信和气象等领域中高频时间序列的边际分布均为重尾分布,而传统时间序列预测算法大多将数据流考虑为正态分布,导致传统算法无法适用于重尾分布的时间序列。针对这种情况,提出一种基于演化算法和改进概率分类器的重尾时间序列预测算法。将预测提前量和预测准确率作为两个优化目标,利用演化算法对两个目标进行独立优化。对高斯过程分类进行改进,使其支持重尾时间序列的分类问题,并且优化了时间效率。实验结果表明,该算法有效地提高了时间序列的预测准确率。

关键词 多目标优化 风险预测 重尾分布 时间序列分类 概率分类器

中图分类号 TP391 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2020.12.043

PREDICTION OF HEAVY TAILED TIME SERIES BASED ON MULTI-OBJECTIVE EVOLUTIONARY ALGORITHM AND IMPROVED PROBABILISTIC CLASSIFICATION

Zou Xiaoyun[1] Lin Wenxue[2]

[1](*Department of Scientific Research, Hubei Polytechnic Institute, Xiaogan 432000, Hubei, China*)

[2](*School of Extended Education, Hubei Polytechnic Institute, Xiaogan 432000, Hubei, China*)

Abstract In the financial, communication and meteorological fields, the side distributions of high frequent time series are heavy tailed distributions, but most of the traditional time series prediction algorithms treat the time series as normal distribution, so that they are not suitable for heavy tailed distributed time series. In view of this, we propose a prediction algorithm of time series based on evolutionary algorithm and improved probabilistic classifier. We treated the prediction earliness and prediction accuracy as two optimization objectives and used evolutional algorithm to optimize both of objectives jointly. The classification of Gaussian processes was improved to support the classification of heavy tailed time series and optimize the time efficiency. The experimental results indicate that our algorithm improves the prediction accuracy of time series effectively.

Keywords Multi-objective optimization Risk prediction Heavy tailed distribution Time series classification Probabilistic classifier

0 引 言

随着互联网和计算机领域的普及和广泛应用,许多应用领域不断地产生新的数据流,如金融市场[1]、视频直播、语音通话和安全监控系统[2]等。数据流有别于传统的静态数据,所以传统数据挖掘技术无法直接用来分析数据流。数据流分析逐渐成为了数据挖掘领域的研究热点,频繁模式挖掘[3]、高效用模式挖掘[4]、概念漂移数据分类[5]、数据流异常检测是其中的重点方向。数据流分类方法要求时间序列的主要特征都出现后才能分类,但在天气预报、金融市场行情、网络安全等应用场景中,如果能尽早预测出数据流的趋势,能够为决策者带来巨大的效益。

收稿日期:2019-07-28。湖北省教育厅科学技术研究项目(B2017519);湖北职业技术学院校级课题"教育智能化背景下高职院校应用数学教学模式创新研究与实践"(2019A05)。**邹小云**,副教授,主研领域:大数据与数据挖掘,概率论与数理统计,高等数学教学及数学建模。**林文学**,副教授。

为了在获得部分特征的情况下对整个时间序列进行提前预测，Alonso 等[6]首次提出了时间序列早期分类的概念，利用基于谓词的分类器结合 boosting 算法可以达到早期分类的目的，但该方法的基分类器训练需要多次扫描样本。时间序列普通分类问题和早期分类问题之间的差别在于，前者的目标是最大化预测准确率，而后者存在预测准确率和早期性（时间提前量）两个冲突的目标。时间序列早期分类算法[7-8]主要分为两个类型：基于 Shapelet 的分类算法通过搜索时间序列中的 Shapelet 特征来判断样本的分类[9-10]；集成多个分类器评估不同时间戳的预测可靠性，确定预测分类的有效性。

在时间序列早期分类的问题中，早期性和分类准确率是两个冲突的目标，目前的主要方法是将这两个目标组合为一个总目标，再通过群体智能技术计算总目标的 Pareto 最优解[11]。如果使用 30% 的序列可获得 80% 的准确率，或者使用 50% 的序列可获得 90% 的准确率，难以判断这两种情况的优劣。在不同的应用场景下，对早期性和准确率两个目标的要求不同，如安全领域需要尽早检测出危险，而天气预报需要在指定天数前尽可能准确地预测出天气。现有的方法一般通过设置体现偏好的目标重要性参数来调节分类模型，但这种方法需要获取后验信息，并通过多次运行分类程序才能完成对参数的调节，显然无法适用于实时数据流预测的情况。

针对上述问题，本文将早期性和分类准确率作为两个独立的目标，采用演化算法同时优化两个目标，给出不同时间戳的预测准确率和早期性结果，供用户根据应用场景自行选择。此外，针对时间序列的重尾分布特点，对高斯过程分类进行了修改，提高对重尾分布时间序列的分类准确率。

1 时间序列预测算法设计

1.1 问题模型

首先给出时间序列预测问题的相关定义。

定义 1 一个长度为 L 的时间序列定义为：

$$TS = \{(t_i, x_i) \mid i = 1,2,\cdots,L\} \tag{1}$$

式中：时间戳 $\{t_i\}_{i=1}^{L}$ 为升序排列的正实数；x_i 为多元变量。

定义 2 从时间 t 截断的时间序列定义为 TS_t，记为 $TS_t = \{(t_i, x_i) \mid i = 1,2,\cdots,t\}$。

基于定义 1 和定义 2，给出时间序列预测问题的定义。

定义 3 假设一个标记时间序列的训练集为 $X = \{(TS_1, CL_1),(TS_2,CL_2),\cdots,(TS_n, CL_n)\}$，其中：$TS_i$ 为时间序列；CL_i 为对应的类标签。时间序列预测分类问题定义为：根据一部分时间序列 TS_{t*} 建立从时间序列到类标签的映射，并能够尽早预测出新到达样本的类标签。

1.2 时间序列早期分类的优化方法设计

本文的目标是获得一个集合 $\{((h_1,h_2,\cdots,h_L), s_{\gamma 1}),((h_1,h_2,\cdots,h_L),s_{\gamma 2}),\cdots,((h_1,h_2,\cdots,h_L), s_{\gamma g})\}$，其中：$\{h_1,h_2,\cdots,h_L\}$ 为分类器序列；$s_{\gamma i}$ 为分类器对应的触发函数；$\{s_{\gamma 1}, s_{\gamma 2},\cdots, s_{\gamma g}\}$ 为优化的触发函数集。

本文方法主要分为 3 个步骤：训练概率分类器集，选择指定的触发函数，优化选择的触发函数集。

1.3 训练分类器

首先训练一个分类器集，负责预测每个时间戳样本的类标签，图 1 是训练分类器的主要流程。在训练时间戳 t 的分类器之前，提取 X 的所有时间序列，然后在时间戳 t 截断序列。时间戳 t 可以是绝对时间，也可以是序列长度的百分比。接着选择一个度量指标评估时间序列之间的距离，再组成一个距离矩阵。

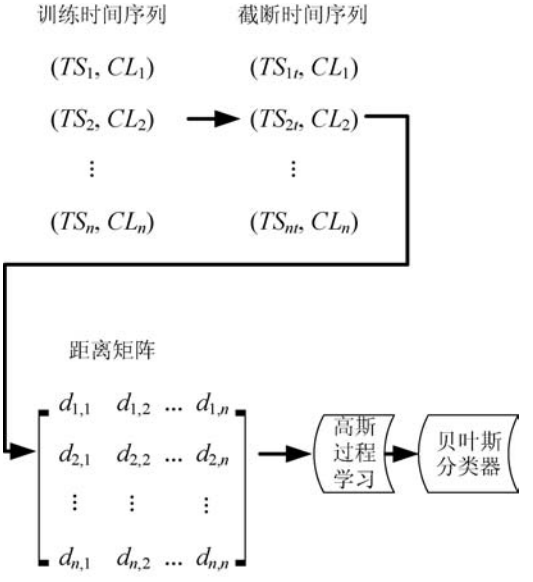

图 1 分类器训练程序的流程

不同的学习算法对输入数据的格式具有特定要求，动态时间归整（Dynamic Time Warping, DTW）采用动态规划方法来进行时间规整的计算，尤其适用于不同长度、不同节奏的时间序列，所以本文采用 DTW 度量时间序列之间的距离。采用改进的贝叶斯分类器输出时间戳 t 的样本关于每个类的隶属度。

1.4 定义触发函数

图 2 是预测时间序列分类的流程图，触发函数负

责判断分类器 h_t 在时间戳 t 所预测的类标签是否可靠,如果可靠度高,则采用其结果,如果可靠度较低,则等待更多的时间序列到达。本文根据分类器 h_t 输出的概率(隶属度)$p^t = \{p_1^t, p_2^t, \cdots, p_k^t\}$ 对时间序列进行预测,采用分类概率的理由是时间序列关于时间的分类概率分布信息对于预测的可靠性具有价值,而且本文对高斯过程分类进行了改进,提高了分类器对重尾分布序列的预测性能。

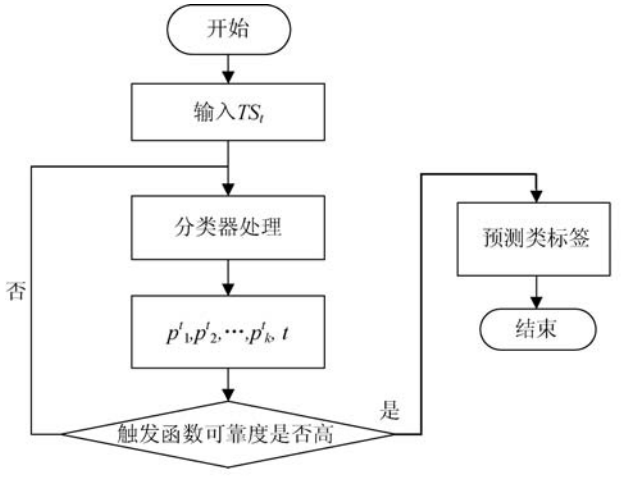

图 2　时间序列预测分类的流程图

图 2 中触发函数的输出为 0 或者 1。如果触发函数判断预测结果的可靠度高,则选择当前预测准确率最高的分类方案;如果触发函数判断预测结果不可靠,那么等待、收集更多的时间序列。线性触发函数定义为:

$$s_\gamma(p^t, t) = \begin{cases} 0 & \gamma_1 p^t_{1:k} + \gamma_1 (p^t_{1:k} - p^t_{2:k}) + \gamma_3 \dfrac{t}{L} \leqslant 0 \\ 1 & \text{其他} \end{cases} \quad (2)$$

式中:$p^t_{1:k}$ 和 $p^t_{2:k}$ 分别是在时间 t 获得的第一和第二的分类概率;$\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)$ 为参数向量,$\gamma_i \in [-1, 1]$。

触发函数集内所有函数的形状均相同,区别在于 $\boldsymbol{\gamma}$ 参数不同,因此需要计算优化预测准确率和预测提前量的 $\boldsymbol{\gamma}$ 参数。

1.5　最优触发函数集

$\boldsymbol{\gamma}$ 参数向量是决定触发函数(式(2))的关键,如果 γ_i 设为[-1,1]内的随机数,则可能出现重复值,并且其触发函数获得的预测准确率和早期性并非最优值。

1) 评价触发函数的质量。为了评价触发函数的预测准确率和早期性,采用两个被广泛使用的评价指标,评估早期性质量的方法为:

$$C_e(X, s_\gamma) = \frac{1}{|X|} \sum_{x \in X} \frac{t_x^*}{L_x} \cdot 100 \quad (3)$$

式中:t_x^* 为对序列 x 产生预测的最早时间戳;L_x 为 x 的长度。本文 t_x^* 为触发函数 s_γ 第一次输出 1 的时间戳。该指标计算了对目标序列成功预测的平均时间,定义为序列百分比的形式。一般通过分类错误率评估时间序列的预测准确率,本文将该指标定义为时间序列的预测错误百分比,计算式表示为:

$$C_a(X, s_\gamma) = \frac{1}{|X|} \sum_{x \in X} \mathrm{II}(\acute{CL}_x \neq CL_x) \cdot 100 \quad (4)$$

式中:$\acute{CL}_x$ 为在时间 t_x^* 对时间序列 x 的预测标签,CL_x 是其真实类标签;如果条件为真,则 $\mathrm{II}(\cdot)$ 值为1,否则为 0。然后基于时间序列 X 的训练集计算触发函数的 $C_e(X, s_\gamma)$ 和 $C_a(X, s_\gamma)$,计算 C_e 需要时间序列触发函数 s_γ 第一次输出 1 的时间戳,因此需要计算 X 中时间序列每个时间戳的分类概率。如果使用常规的分类器评价触发函数,可能出现过拟合和过度优化的情况,本文设计了概率分类器,通过对重尾分布的优化设计,获得预测的分类概率。

2) 优化触发函数。上文定义了预测准确率和早期性两个指标的评价方法,在此对两个目标进行联合优化处理。传统方法的目标是获得最优触发函数 s_γ^*,本文方法则输出一个非支配的触发函数解集 $\{s_{\gamma 1}, s_{\gamma 2}, \cdots, s_{\gamma g}\}$。求解的目标优化问题表示为:

$$\min_\gamma (C_e(X, s_\gamma), C_a(X, s_\gamma)) \quad (5)$$

然后使用元启发式算法求解多目标优化问题。采用多目标优化的理由主要有三点:(1) 如果将两个目标组合成一个优化问题,那么 C_a 和 C_e 必须缩放到相同的区间,才能使两个目标间保持平衡。(2) 两个目标组合优化问题需要预先分配权重,该参数难以确定。(3) 为了权衡两个目标的关系,需要多次运行程序,而多目标优化程序通过一次运行即可获得一组非支配解集。本文将非支配解集输出供用户根据具体的应用场景选择。

1.6　结构设计

本文目标是为用户提供一组分类器和触发函数的集合 $\{((h_1, h_2, \cdots, h_L), s_{\gamma 1}), ((h_1, h_2, \cdots, h_L), s_{\gamma 2}), \cdots, ((h_1, h_2, \cdots, h_L), s_{\gamma g})\}$。图 3 是本文时间序列预测方法的结构,用户根据实际需求选择满足其预测准确率和早期性的最佳触发函数 $s_{\gamma *}$,本文算法基于触发函数 $s_{\gamma *}$ 和分类器 $(h_1, h_2, \cdots, h_L)$ 来预测新时间序列的分类。

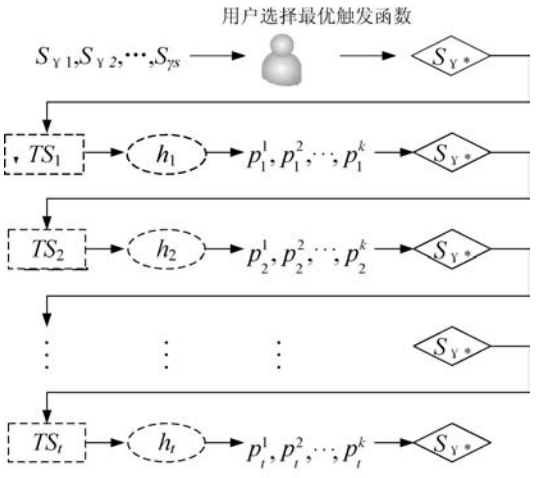

图 3 时间序列预测方法的结构

2 概率分类器设计

设计开发了高斯过程分类器预测时间序列的分类概率,并针对重尾分布型时间序列进行了改进。将贝叶斯分类器和半参数模型结合,能够自适应调节模型的参数,并提高对时间序列中风险数据的预测能力。

2.1 时间序列的贝叶斯分类器模型

设 $y_i(s)$ 表示子集 S_i 的训练函数,设 $c_i \in \{1,2,\cdots,G\}$ 为对应的类标签。假设类 g 的观察样本满足相同的平均函数 $\eta_g(s)$ 和协方差函数 $\gamma_g(s,t)$,观察样本可以是动态或静态的。设 π_g 表示观察样本属于类 g 的先验概率,分类器的目标是学习一个预测模型,将新观察样本 y 分配到 G 中的一个类。贝叶斯分类器的最优规则是最大化类 g 的后验概率:

$$P(c=g \mid y)=\frac{p(y \mid g)\pi_g}{\sum_{j=1}^{G} p(y \mid j)\pi_j} \tag{6}$$

式中:$p(y \mid g)$ 表示样本 y 的类标签为 g 的似然概率。

以网格形式记录序列函数曲线,设 $y_{it}=y(s_{it})$ 表示位置 $s_{it}(t=1,2,\cdots,n_i)$、第 i 个样本的观察函数。新的观察样本表示为向量形式 $\boldsymbol{y}$,将式(6)的 $p(y \mid g)$ 替换为 $f(y \mid g)$,其中 $f(\cdot \mid g)$ 表示类 g 的分布密度。式(6)的分类隶属度后验信息提供了分类不确定性的重要信息,该信息对于风险预测具有巨大的价值。传统的高斯过程分类生成每个曲线的极限后验概率,对于错误分类表现出过度置信,导致对重尾分布存在过度置信的问题。为了解决这个问题,本文对高斯过程分类做出修改,简称为改进的高斯过程分类器(Improved Gaussian Classifier,IGC),IGC 不仅降低了重尾分布数据的分类错误率,同时所估计的后验概率分布能够准确地反映样本在类中的不确定性。

采用分类错误率和对数损失 LogLoss 从不同角度评价分类器的性能,LogLoss 通过惩罚错误分类实现对分类器准确性的量化。LogLoss 定义为:

$$LogLoss=-\frac{1}{N}\sum_{i=1}^{N}\sum_{g=1}^{G} I_{ig}\log(p_{ig}) \tag{7}$$

式中:i 表示观察样本,如果 i 属于类 g,那么 I_{ig} 等于 1,否则等于 0;p_{ig} 表示第 i 个观察样本分配到类 g 的预测概率。

2.2 改进的高斯过程

为了提高式(6)的鲁棒性和后验密度,将式(6)修改为半参数的混合高斯训练模型,定义为:

$$\begin{cases}\boldsymbol{y}_i=\boldsymbol{B}_i\beta+\boldsymbol{R}_i\gamma_i+\varepsilon_i\\ \gamma_i\sim t_q(0,\boldsymbol{\Gamma},v)\quad \varepsilon_i\sim t_{n_i}(0,\boldsymbol{\Lambda}_i,v)\end{cases} \tag{8}$$

式中:$i=1,2,\cdots,M_g$,M_g 为类 g 的观察样本数量;$\boldsymbol{y}_i$ 表示第 i 个观察样本的离散响应(长度为 n_i);$\boldsymbol{B}_i$ 和 $\boldsymbol{R}_i$ 分别为 $n_i\times p$ 和 $n_i\times q$ 的样条基矩阵;β 和 γ_i 分别为 $\boldsymbol{B}_i$ 和 $\boldsymbol{R}_i$ 的系数因子;ε_i 为高斯模型的位移因子;t_q 和 t_{n_i} 分别表示范围为 q 和 n_i 的高斯分布;采用多变量的 t-分布建模随机系数和度量误差,两者的自由度均设为 v;$\boldsymbol{\Gamma}$ 和 $\boldsymbol{\Lambda}_i$ 分别表示两者的缩放矩阵。

本文的改进模型保留了高斯模型原有的优点,如支持无参数的拟合平均函数,支持通过无结构的协方差矩阵 $\boldsymbol{\Gamma}$ 来近似曲线内的协方差结构,支持多种数据分布类型。高斯模型也具有一定的鲁棒性,对重尾分布也具有一定的处理能力,这也是本文选择以高斯分类器为基础的原因。

对每个类训练一个鲁棒模型,计算新观察样本对类的似然来近似式(6)。使用伽玛-正态混合分布表示多变量 t 模型,$\boldsymbol{y}_i$ 的边缘分布属于多变量的 t-分布,定义为:

$$\boldsymbol{y}_i\sim t_{ni}(\boldsymbol{B}_i\beta,\boldsymbol{R}_i\boldsymbol{\Gamma}\boldsymbol{R}_l^{\mathrm{T}}+\boldsymbol{\Lambda}i,v_d) \tag{9}$$

式中:v_d 为密度函数;设 $\gamma_i \mid \tau_i\sim N_q(0,\boldsymbol{\Gamma}/\tau_i)$ 和 $\varepsilon_i \mid \gamma_i$ 两者之间条件独立,新到达样本概率 τ_i 服从正态分布:$\tau_i\sim N_{ni}(0,\boldsymbol{\Lambda}_i/\tau_i)$,也服从伽玛分布:$\tau_i\sim$ Gamma($v/2$,$v/2$)。

参考文献[12]的方法将数据投影到一个特征函数的序列来近似式(6)的密度概率,采用基样条曲线来拟合离散的度量指数,本文方法能够处理不规则采样的数据和多曲线的分类问题。时间序列中存在许多高维数据的情况,因此分类器可能使用高维信息来建模密度,一般通过粗糙惩罚(L1 泛化)或者稀疏性条件(L2 泛化)来解决高维问题。因为本文的分类器在实数数据流的情况下工作,分类效率是一个关键的要素,所以本文采用 B 样条基进行无参数建模,从而使本文

方法的计算效率高于 L1 泛化和 L2 泛化高维近似方法。

2.3　改进的高斯过程分类

在时间序列的实际应用中,存在大量重复的观察样本,所以观察样本之间存在相关性,通过分析相关的观察样本,能够加快时间序列预测的速度。为了考虑观察样本间的依赖性,对式(8)进行修改,增加类级别的随机函数,其系数服从以 v 为密度函数的 t-分布。设 $\boldsymbol{y}_{ij}$ 为第 i 个类、第 j 次重复的 n_{ij} 维响应向量,如果每个类包含多次重复,那么将多级分类模型重写为:

$$\begin{cases}\boldsymbol{y}_{ij}=\boldsymbol{B}_{ij}\beta+\boldsymbol{R}_{ij}\gamma_{ij}+\boldsymbol{D}_{ij}\boldsymbol{\delta}_i+\varepsilon_{ij}\\ \gamma_{ij}\sim t_q(0,\boldsymbol{\Gamma},v)\\ \boldsymbol{\delta}_i\sim t_r(0,\boldsymbol{\Psi},v)\quad \varepsilon_{ij}\sim t_{nij}(0,\boldsymbol{\Lambda}_{ij},v)\end{cases}\tag{10}$$

式中:$\boldsymbol{D}_{ij}$ 为 $n_{ij}\times r$ 的样条基矩阵;$\boldsymbol{\delta}_i$ 为一个随机向量。根据多变量 t-分布将每个样本分类;$\boldsymbol{\psi}$ 表示高斯分布的偏差。

因为重复观察样本的类标签应当相同,所以通过结合多个度量信息进行类级别的再分类。设 $\boldsymbol{y}_i^*=[\boldsymbol{y}_{i1}^{\mathrm{T}},\boldsymbol{y}_{i2}^{\mathrm{T}},\cdots,\boldsymbol{y}_{im_i}^{\mathrm{T}}]^{\mathrm{T}},\boldsymbol{B}_i^*=[\boldsymbol{B}_{i1}^{\mathrm{T}},\boldsymbol{B}_{i2}^{\mathrm{T}},\cdots,\boldsymbol{B}_{im_i}^{\mathrm{T}}]^{\mathrm{T}},n_i^*=\sum_{j=1}^{m_i}n_{ij}$,假设($\boldsymbol{\delta}_i$ 与 τ_i)、(γ_{ij} 与 τ_i)、(ε_i 与 τ_i) 之间彼此条件独立,可将 $\boldsymbol{y}_i^*$ 的边缘分布改写为:

$$\boldsymbol{y}_i^*\sim t_{n_i^*}(\boldsymbol{B}_i^*\beta,\mathrm{Diag}(\{\boldsymbol{R}_{ij}\boldsymbol{\Gamma}\boldsymbol{R}_{ij}^{\mathrm{T}}+\boldsymbol{D}_{ij}\boldsymbol{\Psi}\boldsymbol{D}_{ij}^{\mathrm{T}}+\Lambda_{ij}\}_{j=1}^{m_i})+off\_Diag(\{\boldsymbol{D}_{ij}\boldsymbol{\Psi}\boldsymbol{D}_{ij'}^{\mathrm{T}}\}_{j,j'=1}^{m_i}),v)\tag{11}$$

式中:$\mathrm{Diag}(\{\boldsymbol{A}_j\boldsymbol{A}_j^{\mathrm{T}}\}_{j=1}^{J})$ 包含($n_j\times k$) 的矩阵 $\boldsymbol{A}_j$,$\boldsymbol{A}_j$ 表示以 $\boldsymbol{A}_{jj}$ 为对角的块对角矩阵,$off\_Diag(\{\boldsymbol{A}_j\boldsymbol{A}_{j'}^{\mathrm{T}}\}_{j,j'=1}^{J})$ 定义为:

$$off\_Diag(\{\boldsymbol{A}_j\boldsymbol{A}_{j'}^{\mathrm{T}}\}_{j,j'=1}^{J})=\begin{bmatrix}0_{n_1\times n_1} & A_{12} & \cdots & A_{1J}\\ A_{21} & 0_{n_2\times n_2} & \cdots & A_{2J}\\ \vdots & \vdots & & \vdots\\ A_{J1} & A_{J2} & \cdots & 0_{n_J\times n_J}\end{bmatrix}\tag{12}$$

式中:$0_{nj\times nj}$ 表示0矩阵;$\boldsymbol{A}_{jj'}=\boldsymbol{A}_j\boldsymbol{A}_{j'}^{\mathrm{T}}$。类级别的随机函数 $off\_Diag()$ 有助于度量观察样本间的相关性,并改进了对曲线协方差结构的估计。

3　实　验

3.1　实验数据集

目前主流的时间序列早期分类算法和预测算法均采用 UCR 时间序列数据库作为 benchmark 数据库。本文从 UCR 数据库中选择 17 个不同应用场景的数据集,便于与其他算法的结果作比较。这 17 个数据集的应用场景不同、维度不同且分类数量也不同。将分类器的学习步长定义为时间序列总长度的 5%,采用 DTW 度量时间序列之间的距离。

3.2　参数设置

本文方法包含两个重要的模块:双目标优化算法和分类器的学习算法。分类器采用本文所设计的改进高斯过程分类器。此外,需要通过双目标优化方法计算非支配解集,多目标优化方法并非本文的研究重点,因此选择 3 个被广泛应用的多目标演化优化算法分别与本文方法集成,评估本文时间序列预测方法的性能。3 个双目标优化算法分别为非支配排序遗传算法(NSGA-II)、多目标萤火虫算法(MOFA)和多目标粒子群优化(MOPSO)。3 个优化算法与本文方法的结合版本分别简称为 NSGA_IPA、MOFA_IPA 和 MOPSO_IPA。表 1 所示是 3 个优化算法的参数设置。

表 1　优化算法的参数设置

优化算法	参数	值
NSGA II	交叉率	0.7
	变异率	0.2
	编码方案	向量编码
	交叉算子	两点交叉
	选择算子	精英机制
	变异算子	随机变异
	交叉分布指数	5
	变异分布指数	10
	种群大小	56
	最大迭代次数	100
MOFA	收率速度参数 L	2
	光强吸收系数	5
	变异系数	0.15
	Δ	0.96
	种群大小	10
	最大迭代次数	100
MOPSO	网格数量	5
	当前粒子的速度和方向	1.5
	衰减因子 w	0.97
	局部最优位置的系数 C_1	1.5
	全局最优位置的系数 C_2	1.5
	粒子数量	10
	最大迭代次数	100

3.3　对比方法选择

本文方法的特殊之处是独立优化预测准确率和早

期性两个目标,而其他许多方法均将多个目标组合为一个单目标的优化算法。

本文方法与最近相关的2个时间序列早期分类方法做比较:CAEC 算法[13]、ESC 算法[14]。CAEC 算法是一种广义的时间序列预测分类算法,该算法也是一种基于概率的分类程序,而本文方法也是基于概率的分类程序。ESC 算法是一种基于 Shapelet 的时间序列分类算法,该算法与本文方法属于不同的类型,但其分类准确率好于其他基于 Shapelet 的算法。表2是这2个算法在实验中的参数设置。

表2 对比方法的参数设置

对比方法	参数	值
CAEC	最小支持度	0.1,0.2,0.4,0.8
	切比雪夫边界	2.5,3,3.5
ESC	可靠度阈值	0.001,0.1,0.5,0.9
	准确率阈值	99%

3.4 实验结果分析

每隔5%的时间序列训练一组分类器,分别使用 NSGA-II、MOFA 和 MOPSO 算法作为优化算法,3个优化算法均采用随机的初始化种群。采用式(3)和式(4)分别计算早期性 C_e 和预测准确率 C_a。

采用超体积指标[15]评估3个优化算法解空间的质量,该指标度量了解空间的体积。解集的超体积指标越大,表示其支配的解空间和覆盖的区域越大,其性能越好,该指标同时反映了解集的质量和多样性。统计了3个优化算法对17个数据集最小化 C_a 和 C_e 的超体积结果图,如图4所示,以(100,100)坐标为参考点,该坐标是最差的可能解。

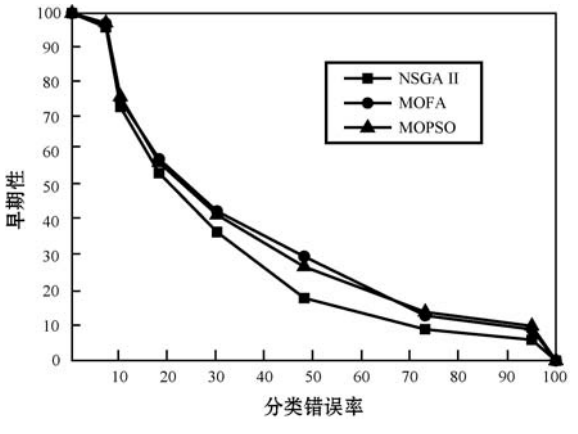

图4 3个优化算法最小化成本 C_a 和 C_e 的超体积结果比较

图4并未统计其他3个对比方法,因为它们的目标是以分类准确率为支配目标,在大多数情况下仅获得了极少的解。从图中可看出,本文方法和3个优化算法的组合方法均获得了较多的解集,其中 NSGA_IPA 的求解质量略好于其他 MOFA 和 MOPSO,因此选择 NSGA_IPA 作为实验方法。

3.5 时间序列的预测实验

(1)搜索的解集。表3所示是3个预测方法对于17个数据集获得的解集数量统计,表中"i/j"中 i 表示非支配解的数量,j 表示计算的所有解数量。可看出 CAEC 计算了固定的总候选解的数量为8,ESC 也计算了固定的总候选解的数量为12,本文方法则提供了较多的候选解数量,并且从解集中提取出非支配解供用户选择。总体而言,本文计算的解集数量远高于其他2个算法。

表3 3个预测算法的解集数量

序号	数据集	CAEC	ESC	NSGA_IPA
1	Beef	0/8	0/12	34/51
2	CBF	4/8	0/12	28/49
3	Coffee	5/8	0/12	5/12
4	Computers	3/8	0/12	16/26
5	ECG200	3/8	0/12	20/37
6	FaceAll	3/8	6/12	0/25
7	Gun_Point	4/8	0/12	6/13
8	Haptics	1/8	0/12	34/50
9	IPD	0/8	0/12	37/50
10	Lighting2	2/8	0/12	40/52
11	Lighting7	2/8	2/12	5/12
12	MALLAT	0/8	0/12	7/13
13	MedicalImage	2/8	3/12	9/14
14	MoteStrain	0/8	0/12	39/48
15	OliveOil	0/8	0/12	7/12
16	Trace	0/8	0/12	37/50
17	TwoLeadECG	0/8	0/12	25/42

(2)预测准确率的结果。虽然本文的目标是为用户提供非支配解集,使用户按照应用需求选择最合适的解,但本文方法对高斯过程分类进行了改进,有效地提高了重尾分布时间序列的分类性能。CAEC 和 ESC 2个算法均以最大化分类准确率为支配目标,以早期性为约束条件,所以这2个算法的分类准确率应当较为理想。此外将本文方法和普通高斯过程分类方法结合,组成的算法简称为 NSGA_GC,比较本文对高斯过程分类的改进效果,NSGA_GC 和 NSGA_IPA 的支配解作为该组实验的性能结果。4个时间序列早期分类算法对于17个数据集分别进行了预测分类实验,每组实验独立运行30次,统计30次的平均准确率作为实验

结果,如图5 所示。比较 NSGA_GC 和 NSGA_IPA 两个算法,NSGA_IPA 对于 17 个数据集的分类准确率均高于 NSGA_GC,反映出本文对于高斯过程分类的改进效果较好。ESC 的预测准确率与本文算法较为接近,ESC 通过检测 Shapelet 实现了较好的分类准确率。但综合 17 个数据集的结果,NSGA_IPA 的预测分类质量最高,并且提供了丰富的解集供用户选择。

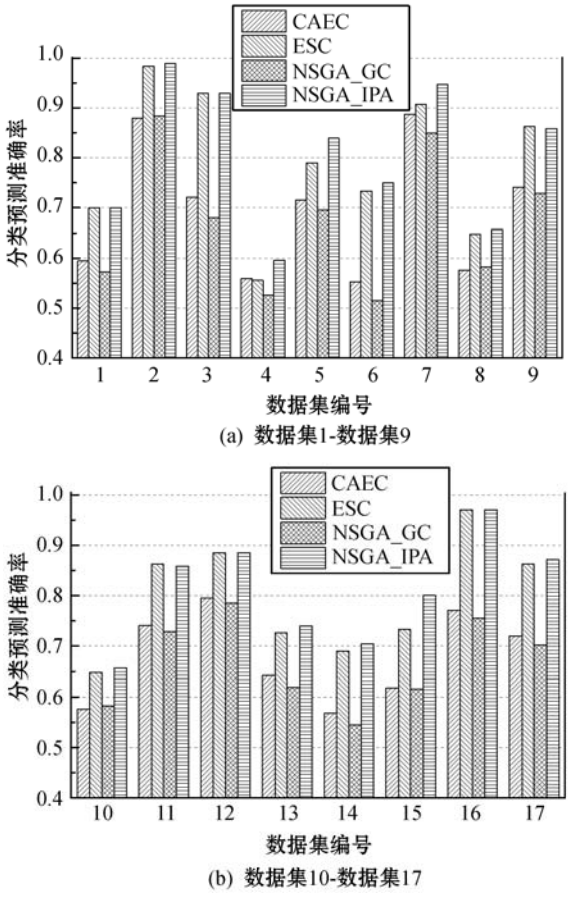

图 5　时间序列早期分类的平均准确率

4　结　语

本文将早期性和分类准确率作为 2 个独立的目标,采用演化算法同时优化 2 个目标,给出不同时间戳的预测准确率和早期性结果,供用户根据应用场景自行选择。此外,针对时间序列的重尾分布特点,对高斯过程分类进行了修改,提高对重尾分布时间序列的分类准确率,并且优化了时间效率。实验结果表明,本文算法有效地提高了时间序列的预测准确率,并为用户提供了丰富的非支配解集。

本文采用了经典的 3 个多目标优化算法完成了仿真实验,优化算法是本文时间序列早期分类的一个关键部分。未来将针对早期性和预测准确率 2 个指标研究更加合适的多目标优化算法,以提高算法的总体性能。

参　考　文　献

[1] Sun J,Fujita H,Chen P,et al. Dynamic financial distress prediction with concept drift based on time weighting combined with Adaboost support vector machine ensemble[J]. Knowledge-Based Systems,2017,120:4 – 14.

[2] 韩德志,陈旭光,雷雨馨,等. 基于 Spark Streaming 的实时数据分析系统及其应用[J]. 计算机应用,2017,37(5):1263 – 1269.

[3] 张亚玲,王婷,王尚平. 增量式隐私保护频繁模式挖掘算法[J]. 计算机应用,2018,38(1):176 – 181.

[4] 吴倩,王林平,罗相洲,等. 基于 MapReduce 的 top-k 高效用模式挖掘算法[J]. 计算机应用研究,2017,34(10):2897 – 2900,2932.

[5] Kim Y,Park C H. An efficient concept drift detection method for streaming data under limited labeling[J]. IEICE Transactions on Information and Systems,2017,100(10):2537 – 2546.

[6] Alonso C J,Juan G,Rodríguez J J,et al. Boosting interval-based literals:Variable length and early classification[M]// Data Mining In Time Series Databases,2003.

[7] 马超红,翁小清. 基于 PAA 的时间序列早期分类[J]. 计算机科学,2018,45(2):291 – 296,317.

[8] Hao P,Zhan Y,Wang L,et al. Feature selection of time series MODIS data for early crop classification using random forest:A case study in kansas,USA[J]. Remote Sensing,2015,7(5):5347 – 5369.

[9] Karlsson I,Papapetrou P,Boström H. Early random shapelet forest[C]//International Conference on Discovery Science,2016.

[10] Ji C,Liu S,Yang C,et al. A shapelet selection algorithm for time series classification:New directions[J]. Procedia Computer Science,2018,129:461 – 467.

[11] Ma C H,Weng X Q,Shan Z N. Early classification of multivariate time series based on piecewise aggregate approximation[C]//International Conference on Health Information Science. Springer,2017.

[12] Buta E,Doss H. Computational approaches for empirical bayes methods and bayesian sensitivity analysis[J]. Annals of Statistics,2011,39(5):2658 – 2685.

[13] Tavenard R,Malinowski S. Cost-Aware early classification of time series[C]//ECML PKDD 2016:European Conference on Machine Learning and Knowledge Discovery in Databases,2016.

[14] He G,Zhao W,Xia X,et al. An ensemble of shapelet-based classifiers on inter-class and intra-class imbalanced multivariate time series at the early stage[J]. Soft Computing,2019,23:6097 – 6114.

[15] 邱飞岳,莫雷平,王丽萍,等. 周期性变量分解的多目标进化算法研究[J]. 小型微型计算机系统,2016,37(6):1318 – 1322.