doi:10.3969/j.issn.1672-1292.2019.03.002

# 基于串行分类算法的不平衡时间序列多分类方法

陈俐名,黄诗茹,修保新,周　鋆

(国防科技大学信息系统工程重点实验室,湖南 长沙 410073)

[**摘要**]　提出了基于串行分类算法的不平衡时间序列多分类方法,并以“上证 50 指数”15 min 交易数据为例,进行了实验检验与结果分析. 结果表明,在多数情况下,串行分类算法比单一算法有更高的准确率、召回率和 F1 值,可以更有效解决不平衡时间序列多分类问题.

[**关键词**]　不平衡,时间序列,多分类,串行分类算法

[**中图分类号**]TP311　[**文献标志码**]A　[**文章编号**]1672-1292(2019)03-0008-07

# Imbalanced Time Series Multi-Classification Method Based on Two-Steps Algorithm

Chen Liming,Huang Shiru,Xiu Baoxin,Zhou Yun

(Science and Technology on Information Systems Engineering Laboratory,National University of Defense Technology,Changsha 410073,China)

**Abstract**:This paper proposes a multi-classification method of imbalanced time series based on two-steps classification algorithm,and takes the 15-minutes trading data of“Shanghai Stock Exchange 50 index”as an example to conduct an experimental test and analysis. The results show that,in most cases,the two-steps classification algorithm has a higher accuracy,recall rate and F1 value than the single algorithm,and that it can solve the problem of multiple classification of unbalanced time series more effectively.

**Key words**:imbalance,time series,multi-classification,two-steps algorithm

科学数据往往随着时间推移而测得. 这些按时间顺序排列的数据就是时间序列. 时间序列几乎无处不在. 温度变化、风能预测、移动跟踪、金融交易、心理和生理信号分析[1-2]、智能家居、天文观测[3-5]等都离不开时间序列. 而时间序列的顺序也不仅限于时间. Jason 等人[6]就提出任何实值序列都可视为时间序列. 因此,许多问题都可以借助时间序列的方法来解决,例如图像分类问题和语音识别问题.

目前,时间序列分类已成为科研和实践的热门话题. 时间序列包含研究对象在每个时间点的重要信息. 因此,许多研究人员致力于从中寻找特定模式,以期预测未来趋势. 时间序列的相关研究主要包括内容查询[7]、异常点检测[8]、预测[9]、聚类[10]和分类[11]. 其中,时间序列分类(time series classification,TSC)主要通过训练模型来区分给定序列. 在临床实践中,TSC 可用于判断疾病是否随时间恶化. 目前,基于 TSC 的实时预警系统已被应用于许多大型医院[1]. 可以说,时间序列从电子健康记录[12]和人类活动识别[13-14]到声学场景分类[15]和网络安全[16],早已无处不在. 数据类型丰富的 UCR/UEA 数据集[17](最大的时间序列数据集存储库)更是给出了 TSC 问题的诸多不同应用.

在 TSC 的诸多研究方向中,不平衡时间序列多分类问题(imbalabced time series multi-classification,ITSMC)指的是时间序列多分类问题中,某些类的样本数量远远少于其他类的情况. 疾病诊断[18]、欺诈检测[19]和异常识别[20]均属于 ITSMC 研究范畴. 数据的不平衡会导致分类器更倾向于多数类,甚至过拟合. 而在这种情况下,识别少数类通常有更重要的意义. 例如,将癌症患者诊断为健康,将欺诈性交易识别为正常.

**收稿日期**:2019-07-05.

**基金项目**:国家自然科学基金(61703416)、湖南省自然科学基金(2018JJ3614).

**通讯联系人**:修保新,博士,副研究员,研究方向:体系行为认知. E-mail:baoxinxiu@163.com

ITSMC 的相关研究通常分布于 3 个子领域:时间序列分类、不平衡分类和多分类,并已取得大量研究成果. 然而,针对 ITSMC 有效的整体性解决方案却十分罕见. 一个面向不平衡时间序列多分类问题的整体性解决方案亟待提出. 基于此,本文设计了串行分类算法,并以金融时间序列分类问题为例,通过"上证 50 指数"的 15 min 交易数据进行了实证检验,以期为 ITSMC 的有效解决提供一种新思路.

# 1 金融时间序列分类问题

## 1.1 问题描述

"低买高卖"是金融市场赚取价差收益的基本方式,一直备受关注. 而"低买高卖"的关键在恰当选择买卖时机. 按照机器学习思路,这是一个将金融时间序列分为"买入"、"卖出"、"持有"、"空仓"的四分类问题. 为便于具体说明和详细研究,本文参照投资组合理论[21],选取了兼具规模性和流动性的上证 50 指数为研究对象. 具体地,本文参照"低买高卖"的原则,以每 15 min 作为一个时间窗口,通过计算最高价、最低价、开盘价和收盘价,将其转化为一条条数据样本. 最终,本文共收集到 1 427 个交易日(2010 年 1 月 19 日—2015 年 12 月 7 日)的 22 832 条原始数据样本.

结合专家建议的"买入"、"卖出"、"持有"、"空仓"区间,本文首先对原始样本进行了类别比例的统计(图 1). 其中,"买入"和"卖出"为少数类样本,二者占比相差低于 2%. "持有"和"空仓"为多数类样本,二者比例同样十分接近. 因此,这是一个典型的不平衡时间序列多分类问题. 而且,多数类内部和少数类内部的样本是近似平衡的.

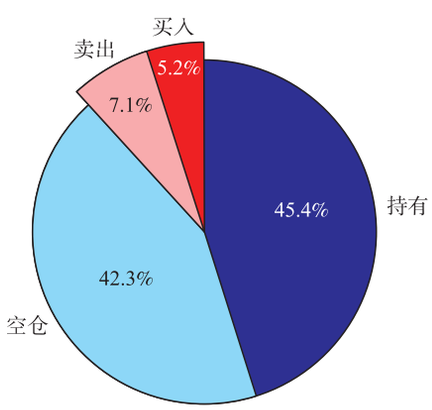


**图 1 原始样本类别比例图**

**Fig. 1 Pie chart of the original sample category**

## 1.2 问题建模

对于不平衡时间序列多分类问题,已有研究主要分布于 3 个子问题领域,而鲜有整体性解决方法. 而本文发现不平衡时间序列,在少数类和多数类的类内分布是平衡的. 鉴于此,本文提出将少数类和多数类分开训练的设想,以期能整体解决不平衡问题(图 2).

这为不平衡时间序列多分类问题提供了新的可能性. 然而,如果无法区分样本归属于多数类还是少数类,就无法确定使用少数类分类器还是多数类分类器. 而时间序列的连贯性带来的类别边界模糊问题,又使得处于"分类边界"的样本难以区分. 对此,我们尝试通过串行分类和并行分类两种机制解决样本的类别归属问题(图 3).

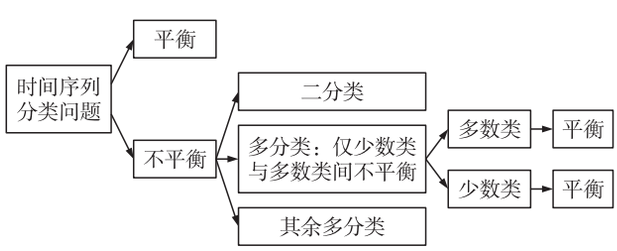


**图 2 不平衡时间序列多分类问题的求解思路**

**Fig. 2 The solution of ITSMC**

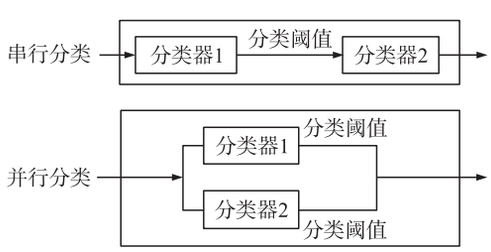


**图 3 串行分类和并行分类解决机制示意图**

**Fig. 3 Schematic diagram of two-steps classification and parallel classification**

按照并行分类,训练集样本同步进入两个分类器,得到两组分类概率. 之后根据分类阈值便可完成样本的分类. 对于串行分类,训练集样本首先进入第 1 个(强)分类器,得到分类概率. 而不满足分类阈值的样本将进入第 2 个分类器,分类结果即为最后结果. 注意,两种机制都需要结合训练过程不断调整分类阈值,以寻求最优值.

考虑到分类器的性能差异显著,本文基于串行分类的解决机制,提出了基于串行分类算法的不平衡时间序列多分类方法.

# 2　串行分类算法

## 2.1　算法流程

串行分类算法流程如图 4 所示. 按照串行分类算法,样本首先进入分类器 1,此时得到分类结果及对应概率. 结合分类阈值和类别概率,可以判断样本是否进入分类器 2 再分类. 若类别概率达到分类阈值,则直接给出相应类别,结束分类;若类别概率未达到分类阈值,则样本进入分类器 2 再分类. 样本通过分类器 2 得到分类结果和对应概率,结束全部分类. 即分类阈值的作用在于判别经过分类器 1 的样本是否需要进入分类器 2 再分类.

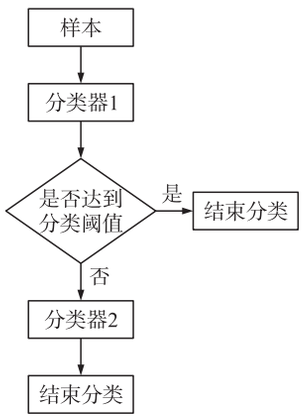


**图 4　串行分类算法流程图**

**Fig. 4　Flow chart of two-steps algorithm**

串行分类算法的两个关键点是分类器的顺序和分类阈值的设置.

分类器顺序的确定需要基于对分类器能力的强弱以及留下正确样本的数量综合判断. 不同问题的训练样本数量的大小与分类效果之间的关系是无法直接进行比较的,但对于同一个问题的不同类别,数据量越大的类别涵盖的特征越多,使得该类分类效果可能更加突出. 而且,若少数类分类器优先,那么经过第一个分类器后进入第二个分类器的样本是占绝大多数的,而这需要少数类分类器极强的分类能力和极高的置信度. 因此,此串行分类算法默认按照首先进行多数类分类,后进行少数类分类的顺序进行. 但在具体问题求解中要注意具体分析. 本文的串行分类算法的实验同样以多数类分类器为首个分类器.

分类阈值可以是一个数,也可以是一个函数,需根据具体情况设计. 在串行分类算法里,分类阈值需要结合两个分类器各自的分类效果,来得到整体的分类效果. 好的分类阈值要做到:

(1)尽量让多数类样本都留在多数类分类器,而尽量少地进入少数类分类器. 因为一旦进入少数类分类器,必然错分.

(2)尽量让少数类样本都离开多数类分类器,而尽量多地进入少数类分类器.

如此,少数类的分类效果才能由少数类分类器决定. 按照阈值对称性多数类样本数量相当的原则,本文在多次试验后将分类阈值设定为 0.75 和 0.25. 即在第一个分类器预测值大于 0.75 和小于 0.25 的样本以多数类分类器结果为准,其余样本进入少数分类器进行二次分类.

## 2.2　评价指标

合适的评价指标是模型客观评价的关键所在. 考虑到不平衡时间序列多分类问题的特殊性,本文采用的评价指标[22]主要有:精度(Precision)、召回率(Recall)、F1 得分(F1-score,F1)、ROC 曲线下面积(area under ROC curve,AUC).

# 3　实验设计与结果分析

## 3.1　实验数据集

本文原始样本为上文中的上证 50 指数共 1 427 个交易日(2010 年 1 月 19 日—2015 年 12 月 7 日)的 22 832 条 15 分钟交易数据.

证券的买入和卖出往往基于对未来一段时间价格走势的预判. 价格走势包括短期、中期、长期等不同时间期限的趋势. 这种不同期限的趋势又往往通过不同期限数据的移动平均值得以刻画. 而收盘价又最能反映一定时间价格的最终状态. 因此,本文在原始样本收盘价的基础上,结合我国证券市场每周 5 个交易日的现实,扩展出 7 个特征:MA_5、MA_10、MA_20、MA_30、MA_60、MA_120、MA_180,分别为 5、10、20、30、60、120、180 条连续数据样本收盘价的移动平均值. 考虑到单个交易日的数据完整性,本文实验数据集为上证 50 指数共 1 415 个交易日(2010 年 2 月 4 日—2015 年 12 月 7 日)和 11 个特征的 22 640 条样本.

### 3.2 算法设计

Wang 等[1]指出,距离度量相似性的方法同样适用于时间序列分类问题. 只不过常用的欧氏距离(Euclidean distance,ED)不再完全适用. 而 ED 改进之后的动态时间规整(dynamic time warping,DTW)则十分适合于时间序列分类问题相似性的度量(如图 5 所示)[23]. 除此之外,循环神经网络(recurrent neural network,RNN)应用于时间序列分类也由来已久,但一直受限于梯度消失问题而停滞不前. 而长短期记忆网络(long short-term memory,LSTM)的出现,不仅解决了梯度消失问题,还在缩短训练时间的同时,大大提高了准确率. 因此,LSTM 更适合金融时间序列的分类问题. 鉴于此,本文以基于 DTW 的 KNN 算法和 LSTM 算法为基准算法,进行了分类实验,检验了串行分类算法的有效性.

结合问题背景和预实验结果,本文将 KNN 的 $k$ 值设为买入与卖出 F1 之和最高的 32,分类阈值设为 0.75 和 0.25. 为防止 LSTM 过拟合,评价过程还引入 DropOut 和 Early Stopping. 本文设定 DropOut 参数为生成网络最多的 0.5,将 Early Stopping 参数设为 8,即测试集损失值在 8 个训练内没有提高,则认为过拟合,并输出评价结果.

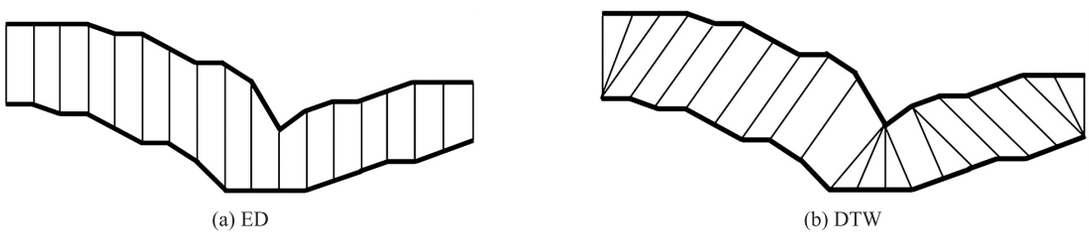


**图 5　两种距离度量示意图**

**Fig. 5　Schematic diagram of two distance measurements**

为比较串行分类算法与单一分类算法的性能,本文首先通过“随机欠采样”的方法将不平衡分类问题转化为平衡分类问题,进行了基于 DTW 的 KNN 算法和时间序列 LSTM 算法的分类实验. 其中,LSTM 参数设置以均线系统策略和“T+1”交易限制为指导. 以时间步长(time_step length)为例,本文设置为 32,即 2 个交易日(1 个最小交易周期),而这刚好符合我国股票市场“T+1”的交易限制. 之后,本文进行了串行分类算法的分类实验,具体实验流程(图 6)为:

(1)切分数据集:将前 5/6 的数据作为训练集,后 1/6 的数据作为测试集.

(2)训练分类器:分别用训练集的多数类样本和少数类样本,利用相关算法分别训练多数类分类器和少数类分类器.

(3)多数类分类:测试集样本进入多数类分类器,得到类别概率值,若达到阈值,结束分类;否则,进入

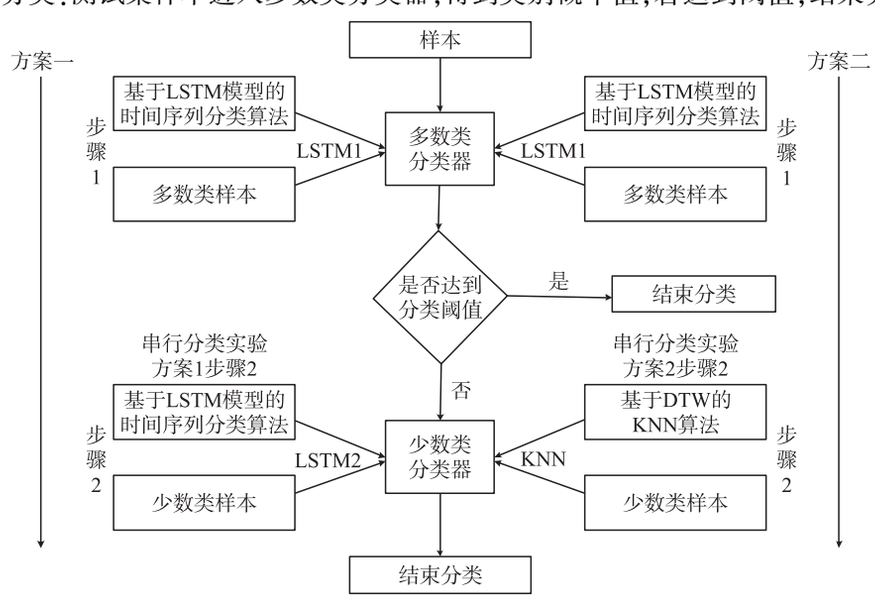


**图 6　串行分类实验流程图**

**Fig. 6　Two-steps classification experiment flow chart**

下一步.

(4)少数类分类:部分测试集样本进入少数类分类器,得最终分类结果,结束全部分类.

串行分类实验以时间序列分类的基准算法为基础,设计了两种串行方案. 其中,方案一采用 LSTM1+LSTM2 的模型组合,方案二采用 LSTM1+KNN-DTW 的模型组合.

### 3.3　结果分析

不同实验的分类结果如图 7、表 1、表 2 所示. 其中,LSTM1+KNN 模型和 LSTM1+LSTM2 模型均以 LSTM1 的分类效果 0.75 和 0.25 为分类阈值进行分类实验. 由图表可知,串行分类算法的各项指标,在多数情况下,整体优于单一的 KNN 算法或 LSTM 算法. 在精度、召回率、F1 值等指标上,LSTM1+LSTM2 略胜 LSTM1+KNN 一筹. 然而,KNN 算法过高的计算复杂度,使得分类的时间成本非常高,高达数十小时. 而同样的数据,LSTM 则只需几分钟甚至数十秒钟即可输出结果.

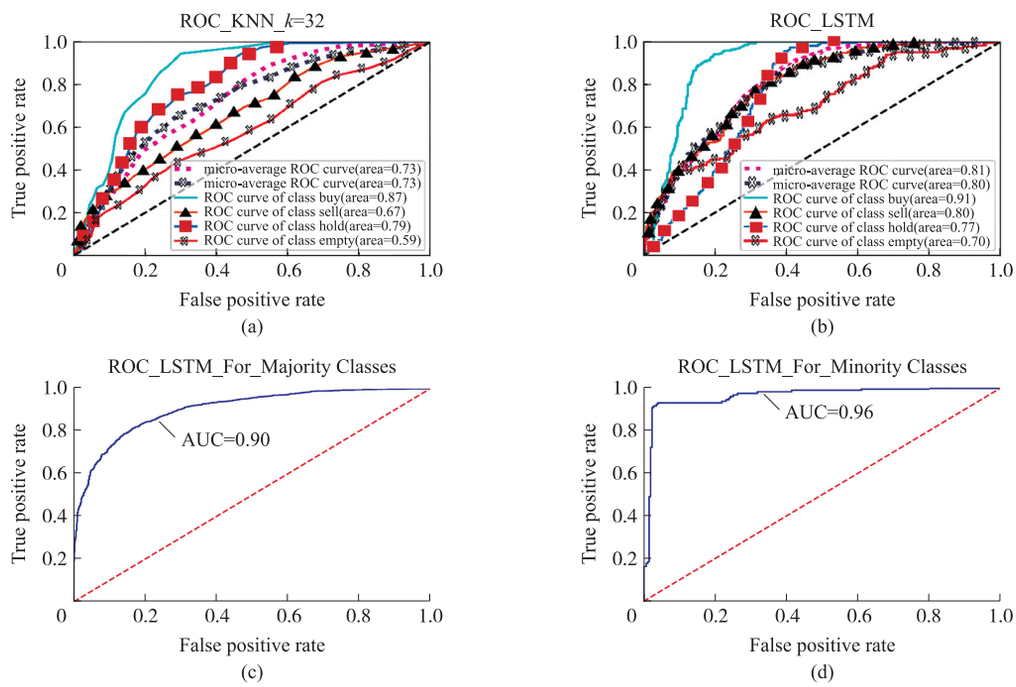


**图 7　ROC 曲线图**

**Fig. 7　ROC curve**

**表 1　具体分类结果对比**

**Table 1　Comparison of specific classification results**

| | 实验 | precision | recall | F1-score |
|---|---|---|---|---|
| 持有 | 单一 KNN | 0.28 | 0.62 | 0.39 |
| | 单一 LSTM | 0.51 | 0.69 | 0.59 |
| | 串行 LSTM1+KNN | 0.79 | 0.84 | 0.81 |
| | 串行 LSTM1+LSTM2 | 0.79 | 0.84 | 0.81 |
| 空仓 | 单一 KNN | 0.39 | 0.23 | 0.29 |
| | 单一 LSTM | 0.46 | 0.38 | 0.42 |
| | 串行 LSTM1+KNN | 0.61 | 0.76 | 0.67 |
| | 串行 LSTM1+LSTM2 | 0.61 | 0.76 | 0.67 |
| 买入 | 单一 KNN | 0.50 | 0.62 | 0.39 |
| | 单一 LSTM | 0.57 | 0.70 | 0.62 |
| | 串行 LSTM1+KNN | 0.20 | 0.18 | 0.19 |
| | 串行 LSTM1+LSTM2 | 0.30 | 0.21 | 0.25 |
| 卖出 | 单一 KNN | 0.65 | 0.42 | 0.51 |
| | 单一 LSTM | 0.53 | 0.33 | 0.40 |
| | 串行 LSTM1+KNN | 0.09 | 0.03 | 0.04 |
| | 串行 LSTM1+LSTM2 | 0.06 | 0.02 | 0.03 |

尽管串行分类算法的效果明显优于单一算法,但指标仅略超 0.6. 归结起来,本文认为主要原因在于数据集规模太小,导致分类器效果不高. 须注意,两次完全不同的 AUC 值,不可一起计算. 因此,使用串行分类算法,难以直接比较两种模型的 AUC 值.

表 2 总体分类结果对比

Table 2 Comparison of overall classification results

| 实验 | precision | recall | F1-score |
|---|---|---|---|
| 单一 KNN | 0.48 | 0.44 | 0.44 |
| 单一 LSTM | 0.52 | 0.52 | 0.51 |
| 串行 LSTM1+KNN | 0.60 | 0.66 | 0.62 |
| 串行 LSTM1+LSTM2 | 0.61 | 0.66 | 0.63 |

由以上结果,本文初步得到如下结论:

(1)实验验证了串行分类算法在处理不平衡时间序列多分类问题上的有效性. 分别为少数类和多数类训练分类器,通过"串行连接"两个分类器,可有效解决不平衡分类问题.

(2)无论是否使用串行分类算法,在多数情况下,LSTM 算法的分类效果优于基于 DTW 的 KNN 算法.

(3)串行分类算法的效果很大程度上取决于第一个分类器. 提升整体分类效果,须重点优化串行的首个分类器.

## 4 结语

本文通过串行分类算法,先训练少数类分类器和多数类分类器,再结合分类阈值分类的方法,可有效解决不平衡时间序列多分类问题. 实验结果证明了串行多分类算法的有效性,但仍存在以下提升可能:

(1)扩充数据集规模,改善分类效果

串行分类算法的有效性虽然得到验证,但受限于数据规模太小,难以训练出很好的深度学习模型,分类效果仍有待提高. 鉴于此,一方面可以考虑扩充数据集,另一方面可以将串行分类算法的两个分类器替换为两组集成[22]分类器. 通过集成学习[22]的方法,或许可以得到效果更优的组合分类器. 例如,在串行分类算法的基础之上,可以使用 LGB 模型解决不平衡时间序列分类问题.

(2)引入投资策略[24-25],优化评价指标

本文只是采用准确率、召回率、F1 得分、AUC 等常用评价指标进行分类器效果评价,这与问题背景和实际应用,并不十分吻合. 因此,后续研究将结合价差收益的立足点,设计投资策略,引入风险-收益评价模型[15],以获得更接近应用的评价结果.

(3)串行多分类算法的验证推广

本文建立的串行多分类算法有效性在金融时间序列已得到验证,具备进一步推广的可能性. 但该算法是否对时序数据更有针对性,能否将其应用于其它类型的数据,需进一步探讨.

串行分类算法是利用"多分类"解决"不平衡"问题,凡是满足框架设定条件的问题,都可再设计使用. 因此,该算法框架的潜在应用价值,有待增加数据集和基准算法进一步验证(交叉验证[26]、假设检验[22]等).

## [参考文献](References)


[1] WANG Z, YAN W, OATES T. Time series classification from scratch with deep neural networks: a strong baseline[C]//2017 International Joint Conference on. Alaska, USA: IEEE, 2017.

[2] KARIM F, MAJUMDAR S, DARABI H, et al. LSTM fully convolutional networks for time series classification[J]. IEEE access, 2017, 6(99): 1662-1669.

[3] AHN J, LEE J H. Clustering algorithm for time series with similar shapes[J]. KSII transactions on internet and information systems, 2018, 12(7): 3112-3127.

[4] JERZAK Z, ZIEKOW H. The DEBS 2014 grand challenge[C]//ACM Press the 8th ACM International Conference. Mumbai, India, 2014.

[5] MUTSCHLER C, ZIEKOW H, JERZAK Z. The DEBS 2013 grand challenge[C]//ACM Press the 7th ACM International Conference. California, USA, 2013.

[6] 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. Beijing, China, 2012.

[7] LIU X Y, REN C L. Fast subsequence matching under time warping in time-series databases[C]//2013 International Conference on Machine Learning and Cybernetics(ICMLC). Tianjin, China, 2013.
[8] WEISS G M. Mining with rarity: a unifying framework[J]. ACM sigkdd explorations newsletter, 2004, 6(1): 7-19.
[9] PURWANTO P, CHIKKANNAN E. Enhanced hybrid prediction models for time series prediction[J]. The international arab journal of information technology, 2018, 15(5): 866-874.
[10] KEOGH E, LIN J, TRUPPEL W. Clustering of time series subsequences is meaningless: implications for previous and future research[J]. Knowledge & information systems, 2005, 8(2): 154-177.
[11] ERGEZER H, LEBLEBICIOGLU K. Time series classification using point-wise features[C]//IEEE 2017 25th Signal Processing and Communications Applications Conference(SIU). Antalya, Turkey, 2017.
[12] RAJKOMAR A, OREN E, CHEN K, et al. Scalable and accurate deep learning for electronic health records[J]. NPJ digital medicine, 2018, 18: 1-18.
[13] NWEKE H F, TEH Y W, AL-GARADI M A, et al. Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: state of the art and research challenges[J]. Expert systems with applications, 2018, 105: 233-261.
[14] WANG J, CHEN Y, HAO S, et al. Deep learning for sensor-based activity recognition: a survey[J]. Pattern recognition letters, 2019, 119: 3-11.
[15] NWE T L, DAT T H, MA B. Convolutional neural network with multi-task learning scheme for acoustic scene classification[C]//9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Kuala Lumpur, 2017.
[16] SUSTO G A, CENEDESE A, TERZI M. Big data application in power systems[M]. Canada: Elseview, 2018: 179-220.
[17] BAGNALL A, LINES J, BOSTROM A, et al. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances[J]. Data mining and knowledge discovery, 2017, 31(3): 606-660.
[18] YI M, CHEN W, CHEN Y, et al. An integrated data mining approach to real-time clinical monitoring and deterioration warning[C]//ACM Sigkdd International Conference on Knowledge Discovery & Data Mining. Not Wiki, 2012.
[19] KRAWCZYK B, GALAR M, JELEN L, et al. Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy[J]. Applied soft computing, 2016, 38: 714-726.
[20] WEI W, LI J, CAO L, et al. Effective detection of sophisticated online banking fraud on extremely imbalanced data[J]. World wide web, 2013, 16(4): 449-475.
[21] MARKOWITZ H M. Portfolio selection[J]. Journal of finance, 1952, 7(1): 77-91.
[22] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.
ZHOU Z H. Machine learning[M]. Beijing: Tsinghua University Press, 2016.(in Chinese)
[23] 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.
[24] 王小川, 陈杰, 卢威等. Python与量化投资: 从基础到实战[M]. 北京: 电子工业出版社, 2018.
WANG X C, CHEN J, LU W, et al. Python and quantitative investing: from basics to practice[M]. Beijing: Publishing House of Electronics Industry, 2018.(in Chinese)
[25] 蔡立耑. 量化投资: 以Python为工具[M]. 北京: 电子工业出版社, 2017.
CAI L D.Quantitative investing: use Python as a tool[M]. Beijing: Publishing House of Electronics Industry, 2017.(in Chinese)
[26] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012.
LI H. Statistical learning methods[M]. Beijing: Tsinghua University Press, 2012.(in Chinese)

[责任编辑: 陈 庆]