齐志轩, 徐伟. 基于随机森林的 Apache Flink 水位线策略[J]. 智能计算机与应用,2026,16(3):73-78. DOI:10.20169/j.issn.2095-2163.24051201

# 基于随机森林的 Apache Flink 水位线策略

齐志轩[1,2], 徐 伟[1,2]
(1 中国科学技术大学 先进技术研究院, 合肥 230088; 2 平安科技(深圳)有限公司, 上海 200030)

**摘 要**: Apache Flink 是一款出色的流式数据处理引擎,尤其擅长处理实时数据。在 Flink 框架中,水位线(Watermarks)是一个关键机制,用于衡量事件时间(Event Time)的进展,以解决实时数据中的乱序问题,对保证数据处理的实时性具有至关重要的作用。传统的水位线策略通常通过设置固定的延迟时间来应对事件的乱序到达,这种方法无法灵活地根据窗口内事件的具体到达情况和乱序程度动态调整水位线。本研究通过引入随机森林模型,根据窗口内已到达事件的特征动态地设定水位线,有效提高了水位线的准确性和系统的处理实时性。

**关键词**: Apache Flink; 水位线; 随机森林
**中图分类号**: TP181 **文献标志码**: A **文章编号**: 2095-2163(2026)03-0073-06

## Apache Flink watermarks strategy based on Random Forest

QI Zhixuan[1,2], XU Wei[1,2]
(1 Institute of Advanced Research, University of Science and Technology of China, Hefei 230088, China;
2 Ping An Technology (Shenzhen), Shanghai 200030, China)

**Abstract**: Apache Flink is an exceptional stream processing engine, particularly adept at handling real-time data. Within the Flink framework, Watermarks serve as a crucial mechanism to measure the progression of Event Time, addressing the issue of out-of-order data arrivals essential for ensuring timely data processing. Traditional watermarks strategies often rely on fixed delay times to manage out-of-order events, in which the flexibility is lacked in dynamically adjusting watermarks based on the specific arrival patterns and disorder within the window. This study introduces a Random Forest model to dynamically set watermarks based on the characteristics of events within the window, significantly enhancing the accuracy of watermarks and the real-time processing capabilities of the system.

**Key words**: Apache Flink; watermarks; Random Forest

## 0 引 言

Apache Flink 是一个高性能、可扩展的流处理和批处理框架,专门设计用于处理大规模的数据流。可支持包括实时数据分析、事件驱动应用程序以及复杂数据流管道等多种数据处理功能[1]。在 Flink 中,水位线(Watermarks)是一种用来跟踪和管理事件时间(Event Time)进展的机制,主要用于处理实时数据流中常见的事件乱序问题[2-5]。水位线通过帮助系统确定何时可以安全地处理或结束特定的时间窗口,确保了数据计算的正确性和完整性。

目前,常用的水位线策略包括有序水位线、延迟水位线和周期性水位线。这些策略根据数据流的不同特性,提供了多样化的处理方法,但同时也存在各自的优势和局限[6-8]。有序水位线假设事件按严格的时间顺序到达,适用于时间顺序明确的数据流。延迟水位线则允许一定程度的延迟以应对事件的乱序到达,而周期性水位线通过定期发出水位线,以周期性方式推进时间窗口的处理。

尽管这些策略各有适应场景,但在处理分布式系统的复杂多变性时仍有不足。因此,引入机器学习模型-特别是随机森林,成为提高系统适应性、处理效率和准确性的关键[9-12]。通过基于窗口内事件的特征数据,随机森林模型能够动态地预测合适的水位线。这种方法利用了机器学习的预测能力,根据实时数据流的实际变化动态调整水位线,从而优化整个数据处理流程[13-14]。

本文将详细介绍随机森林模型在 Apache Flink

作者简介: 齐志轩(1997—), 男,硕士研究生,主要研究方向:软件开发。Email:1473841455@qq.com。

收稿日期: 2024-05-12

水位线策略中的训练与应用。

## 1 问题分析

在流式数据处理系统中,正确设置水位线对于实现高效且精确的数据处理至关重要。Apache Flink 作为一个广泛使用的流式数据处理框架,依赖其水位线机制来处理乱序事件,并确保事件时间语义的正确性[15-16]。水位线设置过低可能增加处理延迟,降低系统的实时性;而设置过高则可能导致尚未处理的数据被误丢,影响结果的完整性和准确性。因此,水位线的设定需要尽可能准确。

水位线本质上是一个特殊的时间戳,用于标识所有时间戳小于或等于此水位线的事件已经被数据处理系统接收。这使得系统能够对时间戳早于或等于水位线的数据进行处理和聚合,确保数据处理的完整性和精度。利用随机森林模型分析事件的到达顺序和事件发生时间,可以预测更适当的水位线。此方法旨在动态调整水位线,以适应数据流的变化,减少因不恰当的水位线设置导致的数据延迟或丢失,同时提高处理效率和吞吐量,确保数据处理的准确性[17-19]。

### 1.1 最优水位线的定义

为了训练随机森林模型,首先必须明确最优水位线的定义。在理想状态下,如果窗口内的事件能按事件发生的时间顺序准时到达,则可将水位线设置为最新到达事件的时间戳。然而,在分布式系统中,由于传输和处理的延迟,较早的事件有可能会在后来的事件之后到达。因此,最优水位线应选为所有已到达事件中时间戳最大者,且此时间戳小于任何未到达事件的时间戳,这样的设置能保证不丢弃任何未到达的数据,同时尽可能降低处理延迟,确保系统的实时性。

在实际的数据处理过程中,由于无法预先获知未到达事件的确切时间戳,需要通过模型预测水位线。利用过往数据中已知的事件时间和到达顺序,可以基于窗口内的完整事件时间序列训练随机森林模型来进行水位线的预测[20-21]。

### 1.2 乱序度的定义和应用

随着时间的推移,事件的时间戳逐渐增大,这引发了一个问题:直接使用模型预测当前运行窗口内的水位线可能因训练数据的时效性问题而不够准确。具体来说,训练数据集中的最优水位线值必然低于当前正在运行窗口的最优水位线值,因为当前的时间戳肯定大于过去任何时刻的时间戳。

为了解决这一挑战,本研究引入了乱序度(Disorder Degree)的概念,作为水位线预测的辅助调整机制。乱序度是根据已到达事件的时间戳分布计算出的衡量指标,用于动态调整基于事件发生时间的水位线预测,以准确地反映实时数据流的当前状态。定义乱序度 $DD(n)$ 为基于 $n$ 个已到达事件的乱序度,最优水位线的计算公式为:

$$\text{Watermark}(n) = \max(e_1, e_2, e_3, \cdots, e_n) - DD(n) \quad (1)$$

其中,$\max(\cdot)$ 函数用于确定所有已到达事件中的最大事件时间戳,$e_i$ 表示第 $i$ 个到达事件的时间戳。通过预测乱序度、而非直接预测水位线的时间戳,可以更有效地适应实时数据流的动态变化和乱序到达的特性。

## 2 数据处理

### 2.1 数据集介绍

在 Apache Flink 引擎运行过程中,初期采取基于传统方法的水位线策略。为了收集用于模型训练的数据集,在数据处理引擎运行初期捕获各个计算窗口内的事件数据,并记录每个事件的发生时间戳,再按照事件到达的顺序进行整理。将这些时间戳信息格式化为 JSON 格式,并被持续地写入到一个预先指定的文件中。通过这种方法,就能够累积收集到一个足够随机森林模型训练的数据集,该数据集包含了事件时间的时间戳信息。具体的数据集示例如图 1 所示。

**图 1 数据集**

**Fig. 1 Data set**

### 2.2　数据预处理

在 Apache Flink 引擎的实际应用中，对每个到达的事件进行实时处理，并根据已到达事件的时间戳更新水位线至关重要。这一过程要求系统能够基于当前已知的信息实时调整。尽管如此，系统在处理每次事件到达时并不能预知未来事件的到达时间及其时间戳。这一实际需求与数据集的现状存在显著差异，因为数据集中的记录一开始就包含了整个计算窗口内所有事件的完整时间戳信息，这与实际逐步到达的事件处理情况不符。

为了使数据集更好地模拟实际事件逐个到达的情形，并为模型训练提供更接近实际的数据，研究采取了一系列预处理步骤来转换原始数据集。具体而言，对每个计算窗口内包含的事件序列进行拆分，从而生成了一系列子数据集，每个子数据集代表了窗口内从第一个事件到当前事件的累积状态。因此，对于一个包含 $n$ 个事件的窗口，将得到 $n$ 个独立的记录，每个记录包含从第一个事件到第 $i$ 个事件的时间戳信息（$i = 1, 2, 3, \cdots, n$）。这种方法能够模拟每个事件到达时基于当前已到达事件所能进行的水位线更新过程。

为了进一步计算每个窗口中事件到达时的最优水位线，根据 1.1 节中定义的最优水位线标准，并结合式（1）计算乱序度。乱序度计算流程如图 2 所示。图 2 展示了从原始事件时间戳到计算最优水位线，再到计算乱序度的整个流程。这一过程不仅为模型的训练和测试准备了精确的输入数据，也确保了后续模型预测的准确性和实用性。预处理后的数据示例如图 3 所示。

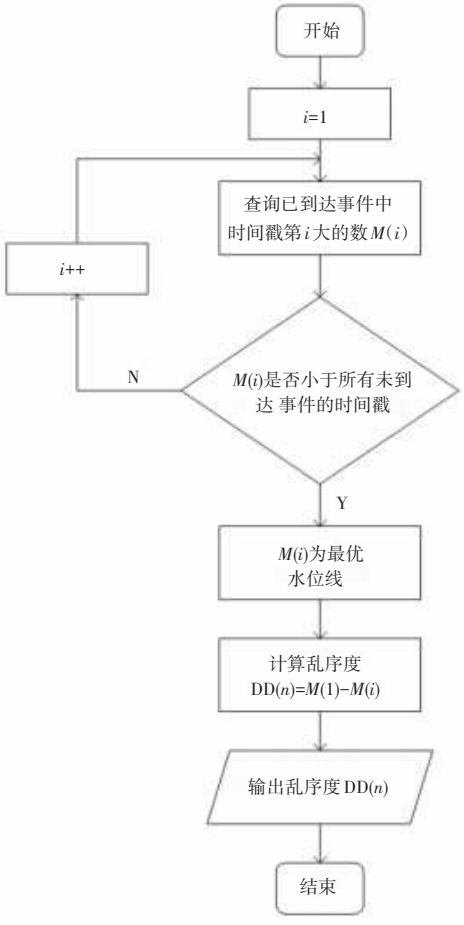


**图 2　乱序度计算流程图**

**Fig. 2　Flowchart of disorder degree calculation**

**图 3　预处理数据**

**Fig. 3　Preprocessing data**

### 2.3　特征值提取

在随机森林模型的构建过程中，合理的特征值选取对于解决回归问题至关重要。考虑到机器学习模型的性能直接受到输入特征选择的影响，本研究在数据预处理的基础上进一步进行了特征提取工作。由于输入数据的数量可能会有所变化，直接使用原始数据作为模型输入并不可行。因此，本研究通过人工筛选，确定了一系列代表性的特征值，以此作为随机森林模型的输入。表 1 是基于预处理后数据选取的特征值列表，及其相应的类型说明。

表 1 特征表

Table 1 Feature table

| 特征 | 类型 |
|---|---|
| 事件总数(SUM) | INT |
| 时间跨度(TS) | INT |
| 时间间隔绝对值平均数(MAVTI) | FLOAT |
| 时间间隔最大绝对偏差(MAXTI) | INT |
| 时间间隔最小绝对偏差(MINTI) | INT |
| 时间间隔标准差(SDTI) | FLOAT |
| 时间间隔中位数(MTI) | FLOAT |
| 逆序对数量($I$) | INT |
| 逆序度(ID) | FLOAT |

在提取特征值的过程中，首先计算了数据集中每组数据的事件总数(SUM($n$) = $n$)，以衡量输入数据的规模。接着，时间跨度 TS 被定义为窗口内最大和最小事件时间戳之差，用于反映数据在宏观上的分布特征。具体公式化：

$$\mathrm{TS}(n) = \max(e_1, e_2, e_3, \cdots, e_n) - \min(e_1, e_2, e_3, \cdots, e_n) \tag{2}$$

对于每个时间间隔($t_i = e_{i+1} - e_i$)，计算了其绝对值的平均数(MAVTI)，以及最大和最小绝对偏差(MAXTI 和 MINTI)，时间间隔的标准差(SDTI)以及中位数(MTI)，用以更细致地反映时间戳的波动情况。逆序对数量($I$) 及其归一化形式，即逆序度(ID)，也被选为特征值，以反映数据的乱序情况。可用下式来计算：

$$\mathrm{ID}(n) = \frac{I(n)}{n} \tag{3}$$

特征值的选取不仅基于统计学原理，也考虑了数据的实际意义和对模型预测性能的潜影响。通过这些特征，模型能够更准确地捕捉到数据的内在规律，特征提取的数据如图 4 所示。

图 4 特征值数据

Fig. 4 Feature data

## 3 评估方式分析

为了全面评估随机森林模型预测 Apache Flink 窗口乱序度的性能，本研究采用了 4 种主要的评估指标：均方误差(MSE)、均方根误差(RMSE)、平均绝对误差(MAE)、以及决定系数($R^2$)。以下分别介绍这 4 种评估方法的计算公式以及各自的侧重点。

### 3.1 均方误差

均方误差(MSE)是衡量模型预测值与实际值偏差的平方的平均值，其公式如下：

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4}$$

其中，$n$ 表示样本数量；$y_i$ 表示第 $i$ 个实际值；$\hat{y}_i$ 表示第 $i$ 个预测值。MSE 的值越小，表示模型的预测准确度越高。该指标侧重于惩罚较大的误差。

### 3.2 均方根误差

均方根误差(RMSE)是均方误差的平方根，提供了误差的平均水平，公式如下：

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)} \tag{5}$$

RMSE 同样侧重于对较大预测误差的惩罚，且其单位与原数据保持一致，易于解释。

### 3.3 平均绝对误差

平均绝对误差(MAE)是预测值偏离实际值的

绝对值的平均数，其公式如下：

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6}$$

MAE 提供了预测误差的平均大小，对所有大小的误差都给予相同的权重，更为稳健。

### 3.4　决定系数

决定系数（$R^2$），又称为拟合优度，是衡量模型解释变量波动的能力。$R^2$ 的值越接近 1，表示模型的解释能力越强，其公式如下：

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{7}$$

其中，$\bar{y}_i$ 表示所有实际值的平均值。$R^2$ 侧重于评估模型对数据变异性的解释程度。

通过运用上述 4 种评估指标，本研究将全面评估随机森林模型在预测 Flink 窗口内已到事件的乱序度的效果。

## 4　模型训练

随机森林模型通过 Bagging 算法进行训练，即对原始数据集重复地有放回抽样产生若干子集，并且每个子集被用来训练一棵决策树。模型训练的具体配置和环境详见表 2，以下是详细的模型训练流程。

（1）按照 1.2 节的方法对数据进行预处理，计算乱序度。

（2）将预处理后的数据分为训练集和测试集，其中 80% 用作训练集，20%用作测试集。

（3）对训练数据集按照 2.3 节介绍的方法进行特征提取。

（4）随机森林模型使用训练数据集进行训练。

（5）完成训练后，利用测试集进行评估，并根据第 3 节介绍的指标评估其性能。

在对随机森林模型的参数设定进行优化时，需要特别关注决策树的数量，这一参数对模型的预测准确度具有显著影响。鉴于数据集的特征数量较少，最大特征数参数设为不受限制，以利用全部特征。通过实验比较，发现当决策树数量少于 340 时，增加树的数量能有效降低均方误差，从而提升模型的预测精度。然而，当树的数量超过 340 后，进一步增加决策树数量不再显著改善模型的准确度，均方误差表现出波动状态。因此，为了既提高乱序度的预测效率、又节省计算资源，决定将决策树数量定为 340。此外，其他参数也通过交叉验证方法来确定最优值，具体设置详见表 3。

**表 2　实验配置表**

**Table 2　Experimental configuration table**

| 项目 | 项目配置 |
|---|---|
| 操作系统 | Windows 11 |
| 编程语言 | Python 3.11 |
| 处理器 | 英特尔十三代 Core i5 |

**表 3　随机森林参数设置**

**Table 3　Random Forest parameters setting**

| 参数 | 参数说明 | 参数值 |
|---|---|---|
| n_estimators | 决策树数量 | 340 |
| max_features | 最大特征数 | 全部 |
| max_depth | 最大深度 | 17 |
| min_samples_split | 最小样本分割数 | 5 |
| min_samples_leaf | 叶子节点最少样本数 | 3 |
| bootstrap | 自助采样 | True |
| criterion | 分割标准 | MSE |

在其他应用场景中，当 Apache Flink 数据处理引擎使用随机森林模型进行水位线预测时，所需的训练参数可能与上述参数不完全相同，因此可能需要进行适当的调整。

## 5　模型测试

训练好的随机森林模型的预测结果与实际值显示，如图 5 所示。研究可知，随机森林模型在预测乱序度方面具有较高的准确性。

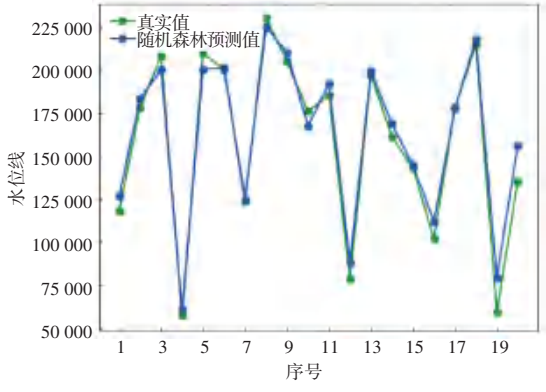


**图 5　模型预测结果**

**Fig. 5　Prediction results of the model**

利用事先分割好的测试数据集，计算了随机森林模型的均方误差、均方根误差、平均绝对误差和决定系数。结果见表 4。

**表 4　模型性能测试结果**

**Table 4　Test results of model performance**

| MSE | RMSE | MAE | $R^2$ |
|---|---|---|---|
| 77 600 023 | 8 809 | 6 915 | 0.97 |

随机森林模型能够更准确地预测乱序度，这归功于其能够有效处理不确定大小的数据组，并利用

提取的特征进行准确预测。特征值权重的计算结果如图 6 所示,进一步证明了随机森林模型的适用性。

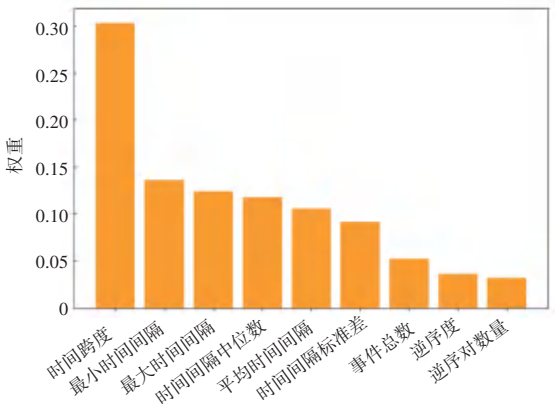


**图 6 随机森林特征值权重**

**Fig. 6 Eigenvalue weights of Random Forest**

为进一步验证随机森林模型在提升 Apache Flink 数据处理实时性方面的效果,本研究采用了控制变量法,比较了传统水位线策略与随机森林水位线策略在相同吞吐量条件下的数据处理延迟。

实验结果见表 5。由表 5 对比分析表明,随机森林水位线策略能够在保持相同吞吐量的前提下,有效地减少数据处理延迟。特别是在处理大规模数据的场景中,随机森林水位线策略在减少处理延迟方面展现出了明显的优势。

**表 5 对比测试结果**

**Table 5 Comparison test results**

| 吞吐量/(records · $s^{-1}$) | 传统水位线策略下的时延/ms | 随机森林水位线策略下的时延/ms |
|---|---|---|
| 10 000 | 88 | 85 |
| 15 000 | 153 | 145 |
| 20 000 | 189 | 171 |
| 25 000 | 360 | 311 |
| 30 000 | 689 | 570 |
| 40 000 | 986 | 701 |

## 6 结束语

本篇论文详细探讨了随机森林模型在 Apache Flink 水位线策略中的应用。研究包括了数据处理、特征提取、模型训练、性能评估及对比实验的各个方面。基于实验结果,本文证实了随机森林模型在 Apache Flink 水位线策略中的应用能显著降低数据处理延迟,并有效提升系统的实时性。本论文需要感谢中国科学技术大学先进技术研究院在项目研发和论文撰写过程中提供的支持和帮助。

## 参考文献

[1] CARBONE P, KATSIFODIMOS A, EWEN S, et al. Apache flink: Stream and batch processing in a single engine[J]. The Bulletin of the Technical Committee on Data Engineering, 2015, 38(4): 28-38.

[2] 曹云柯. 一种基于 Flink 实时数仓的系统设计及功能实现研究[J]. 电子技术与软件工程,2022(6):219-222.

[3] CHINTAPALLI S, DAGIT D, EVANS B, et al. Benchmarking streaming computation engines: Storm, flink and spark streaming[C]//Proceedings of 2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW). Piscataway, NJ: IEEE, 2016: 1789-1792.

[4] 安超广. 基于 Flink 的分布式并行逻辑回归算法的研究[J]. 长江信息通信,2023,36(4):65-67.

[5] 吕鹤轩, 黄山, 艾力卡木·再比布拉, 等. Flink 水位线动态调整策略[J]. 计算机工程与科学, 2023, 45(2):237-245.

[6] 徐海霞. Apache Flink 流式计算模型在数据处理中的应用与性能优化研究[J]. 电脑知识与技术, 2024, 20(7):71-73.

[7] 曹张宇, 钟原, 周静. 基于 Flink 的分布式在线集成学习框架研究[J]. 计算机应用研究, 2023, 40(6):1784-1788.

[8] 孙国璋, 黄山, 艾力卡木·再比布拉, 等. 基于 Flink 的 k-支配 skyline 体并行求解算法[J]. 计算机工程与科学, 2023, 45(1):17-27.

[9] RIGATTI S J. Random Forest[J]. Journal of Insurance Medicine, 2017, 47(1): 31-39.

[10] 方匡南, 吴见彬, 朱建平, 等. 随机森林方法研究综述[J]. 统计与信息论坛, 2011, 26(3): 32-38.

[11] BIAU G, SCORNET E. A random forest guided tour[J]. Test, 2016, 25: 197-227.

[12] PAL M. Random forest classifier for remote sensing classification[J]. International Journal of Remote Sensing, 2005, 26(1): 217-222.

[13] QI Yanjun. Random forest for bioinformatics[M]//ZHANG C, MA Y. Ensemble Machine Learning: Methods and Applications. Cham: Springer, 2012: 307-323.

[14] BIAU G. Analysis of a random forests model[J]. The Journal of Machine Learning Research, 2012, 13(2): 1063-1095.

[15] 艾力卡木·再比布拉, 甄妞, 黄山, 等. 基于深度学习的容器化 Flink 上下游负载均衡策略研究[J]. 大连民族大学学报, 2023,25(1):47-52.

[16] 冯鹏,胡佳丽,黄山,等.一种基于 Flink 的流关联挖掘算法[J]. 大连民族大学学报,2022,24(1):58-62.

[17] CHEN Yanyu, ZHENG Wenzhe, LI Wenbo, et al. Large group activity security risk assessment and risk early warning based on random forest algorithm[J]. Pattern Recognition Letters, 2021, 144: 1-5.

[18] SCHONLAU M, ZOU R Y. The random forest algorithm for statistical learning[J]. The Stata Journal, 2020, 20(1): 3-29.

[19] IWENDI C, BASHIR A K, PESHKAR A, et al. COVID-19 patient health prediction using boosted random forest algorithm[J]. Frontiers in Public Health, 2020, 8: 357.

[20] KABIRAJ S, RAIHAN M, ALVI N, et al. Breast cancer risk prediction using XGBoost and random forest algorithm[C]// Proceedings of 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT). Piscataway, NJ: IEEE, 2020: 1-4.

[21] SHEYKHMOUSA M, MAHDIANPARI M, GHANBARI H, et al. Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 13: 6308-6325.