Online 기반 Random forest 알고리즘

20510082 전우진

20510083 한재웅

CONTENTS

01. 소개

- Online 기반 Random forest 알고리즘 연구 선정
  - Online random forests regression with memories
  - Calculating Feature Importance in Data Streams with Concept Drift using Online Random Forest
  - Online adaptive decision trees based on concentration inequalities
  - Enhanced-Online-Random-Forest Model for Static Voltage Stability Assessment Using Wide Area Measurements

- Online 기반 Random forest 알고리즘
  - Online 학습 기반 알고리즘은 지속적으로 생성되는 방대한 양의 데이터를 mining 하는 부분에 적합함
  - Online 학습 기반 알고리즘으로 OnlineTree 기법을 기반한 다양한 알고리즘이 제안됨

01. 소개

- 기존 방법의 한계점
  - Online random forests regression with memories
    - Online 학습에 대한 대부분의 기존 연구들은 model의 업데이트가 너무 자주 일어나 비효율적인 연산이 많음
    - Data stream의 샘플 사이의 continuing dependencies를 간과함
  - Calculating Feature Importance in Data Streams with Concept Drift using Online Random Forest
    - 초기엔 Concept drift가 일어나는 data stream에 대해 분류기에 대해 re-training, re-weighting 시킴
    - 다른 해결방안으로는 주기적으로 새 분류기를 만드는 것이기 때문에 cost가 높음

01. 소개

- 기존 방법의 한계점
  - Online adaptive decision trees based on concentration inequalities
    - 기존 알고리즘인 data stream은 시간의 지남에 따라 변화가 있을 수 있고 이전 학습 모델이 시대에 뒤떨어질 수 있음
    - 예를 들어, 사용자는 뉴스나 의류 선호도가 변경되면 학습시스템은 적절한 제안을 찾기 위해 실시간으로 내용을 수집해야 추적할 수 있음
    - Data stream 중 하나인 OnlineTree2는 의사결정 트리 구조를 조정하고 로컬 매개변수를 조정하는 방법으로 로컬 성능 향상, 저장된 예시 수 최적화, 처리 시간 단축 등의 측면에서 효율적인 학습을 제시하지만, 더 정교한 것을 제안하고자 함

  - Enhanced-Online-Random-Forest Model for Static Voltage Stability Assessment Using Wide Area Measurements
    - 산업적인 면에서 규모가 커지고 데이터 분석이 갈수록 복잡해지기 때문에, 기존 알고리즘 마이닝으로는 불안정성을 보임
    - 산업적인 부분에서의 데이터 안정성을 적용하자면 예를 들어서 전력 모니터링 상황 인식에 한계점을 가지고 있음
    - 시스템 정보에서 주기적으로 시스템이 업데이트가 되고 모델 성능을 개선시켜야 할 필요성이 있음

01. 소개

- 한계점을 해결하기 위한 연구 방법
  - Online random forests regression with memories
    - ➢ Off-line으로 훈련된 RF의 구조에서 leaf-level weight를 훈련데이터가 입력되는 것과 동시에 훈련된 RF의 구조변화 없이 업데이트함
  - Calculating Feature Importance in Data Streams with Concept Drift using Online Random Forest
    - ➢ Off-line feature importance 측정방식을 online 분류기에 적용함
    - ➢ Mean Decrease in Gini Impurity (MDG), Mean Decrease in Accuracy (MDA)

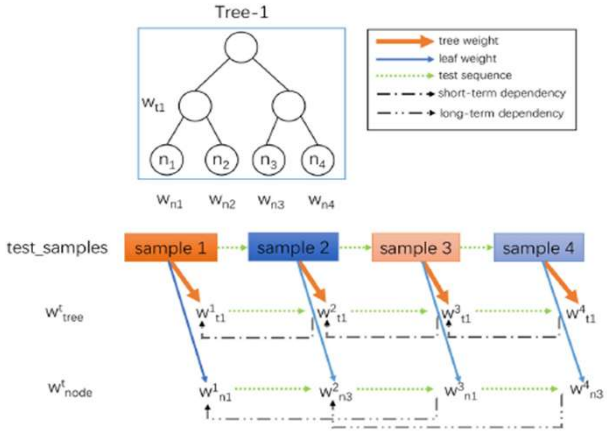
01. 소개

- 한계점을 해결하기 위한 연구 방법
  - Online adaptive decision trees based on concentration inequalities
    - iadem-3 알고리즘을 제안하고 있음
    - iadem-3 는 각 트리 노드의 오류율 추적을 지속적으로 점검하여 하위 트리의 일관성을 모니터링하는 방법을 제안함
    - 점진적인 변화에서의 오류를 측정하여 변화의 영향을 살펴보고 학습함

  - Enhanced-Online-Random-Forest Model for Static Voltage Stability Assessment Using Wide Area Measurements
    - EORF(강화 온라인 랜덤 포레스트)라는 새로운 알고리즘을 제안함
    - 전력 산업에서의 사례기반을 통해 PMU 데이터를 이용하여 전압 안정성을 시간적으로 평가를 하고 성능을 예측함 (부하 예상 예측)
    - 기존 오프라인 모드에서는 다시 학습하는데 많은 시간이 걸리므로 제안하는 알고리즘으로 개선하고자 함

02. 알고리즘별 핵심 제안

- 알고리즘
  - Online random forests regression with memories
    - Online Weight Learning Random Forest Regression(OWL-RFR)
    - 예를 들어 temporal sequence의 sample이 4개가 있고, sample 3이 sample 1에 비슷하고, sample 4가 sample 2에 비슷하다고 함. 그럼 sample 1, 3과 2, 4는 각각 같은 node에 속할 것임(node n_1, node n_3)
    - Tree weight 방식과 달리 leaf weight 방식은 현재의 sample이 이전의 sample과 비슷해서 같은 node에 속하면 이전 sample을 바탕으로 weight를 업데이트함
      - n_1의 sample 3에 대한 weight는 sample 1에 대한 weight를 바탕으로 업데이트됨 → current prediction에 long-term memory를 제공함



02. 알고리즘별 핵심 제안

- 알고리즘
  - Calculating Feature Importance in Data Streams with Concept Drift using Online Random Forest
    - Tree discarding
      - Concept drift가 일어나면 old concept으로 훈련된 트리는 분류 정확도가 떨어지고 ORF는 해당 트리를 훈련되지 않은 트리 (“stump”)로 교체함
    - Mean Decrease in Gini Impurity (MDG)
      - 트리에서 똑같은 feature를 사용하는 노드들의 평균을 구하고 해당 feature가 data를 분류하는데 필요 없어지면 그 feature와 관련된 tree를 교체함
    - Mean Decrease in Accuracy (MDA)
      - 훈련데이터의 특정 feature를 permutation 한 뒤, forest의 평균 accuracy 변화를 측정하고 변화가 거의 없으면 해당 feature가 중요하지 않다고 결정하고 관련된 tree를 제거함

02. 알고리즘별 핵심 제안

- 알고리즘
  - Online adaptive decision trees based on concentration inequalities
    - ➢ IADEM-3 induction algorithm
      - ❖ 주된 tree와 병렬로 유도되는 대체 하위 트리를 생성함
      - ❖ 점진적인 변화가 종종 오류율을 증가시키므로 하위 트리의 변화도 확인하면서 반복적으로 수행함
      - ❖ 하위 트리는 학습 모델의 정확도를 현저히 향상시킬 때만 수행함 (각 하위 트리의 오류율을 추정)
      - ❖ 점진적인 학습을 통해 정확도를 향상시키는 것이 주된 목표

Algorithm 1: IADEM-3 induction algorithm.

Procedure LearnFromInstance:
- *instance*: Training instance
- *mainNode*: Tree node
- δ: Confidence to compare subtrees
- τ: Sample size threshold to break ties by model complexity

Result: The tree rooted at *mainNode* is updated with *instance*

```
1  begin
2      Perform test-then-train to update the change detector and
       statistics of mainNode with instance
3      if mainNode is a split node then
4          CheckForPruning(mainNode, δ)
5          if not pruned then
6              if mainNode.changeDetector estimates a concept drift
               then
7                  add a new alternative subtree in mainNode
8              forall the altNode rooted at mainNode do
9                  LearnFromInstance(instance, altNode, δ, τ)
10                 if IsMoreAccurate(altNode, mainNode, δ) then
11                     promote altNode
12                 else if IsLikelyToBeUseless(altNode, mainNode, δ)
                   then
13                     remove altSubtree
14                 else
15                     // altNode has not been promoted nor
                       deletedBreakTieByModelComplexity(altNode,
                       mainNode, τ)
16                 if altNode was promoted then
17                     exit procedure
18             sort instance into the corresponding child of
               mainNode
19             LearnFromInstance(instance, child, δ, τ)
20     else
21         // mainNode is a leaf node
22         if mainNode.changeDetector estimates a concept drift
           then
23             reset all statistics in mainNode
24         try to split mainNode
```

02. 알고리즘별 핵심 제안

- 알고리즘
  - Enhanced-Online-Random-Forest Model for Static Voltage Stability Assessment Using Wide Area Measurements
    - ➢ Enhanced-Online-Random-Forest Algorithm
      - ❖ 기존 Online Random Forest에서 성능이 좋게 나오는 구간을 가중치를 두어서 성능 측정함
      - ❖ PMU 데이터에서 실시간 시스템 위상 변경과 전압 변경성을 고려하고 전압을 예측함
      - ❖ 기존 Offline과 Online에서의 Accuracy, Security, Reliability를 비교함
      - ❖ 최종적으로 다른 알고리즘과도 성능 비교함

Algorithm 1: Enhanced Online Random Forest Algorithm

Input: $\mathbf{x}, n_u, n_t$.
Initialize: $c_d \leftarrow 0$; $D^*_{\text{learn}} \leftarrow \varnothing$.
$T_{\text{new}} \leftarrow$ Select n_u trees randomly from $T = \{t_k\}_{k=1}^{n_t}$;
$T_{\text{old}} \leftarrow T - T_{\text{new}}$;
$c_d \leftarrow$ Detect drift via the drift detection method;
if $c_d \neq 0$ **then**
  $D^*_{\text{learn}} \leftarrow$ Update it via (5);
  for $k = 1, \ldots, n_u$ **do**
    $n_r \leftarrow \text{Poisson}(1)$;
    if $n_r > 0$ and $c_d = 1$ **then**
      for 1 to n_r
        $t_k^{\text{new}} \leftarrow$ Update it by tree growth strategy;
      end for
    else
      $t_k^{\text{new}} \leftarrow$ Replace it with a new tree by D^*_{learn};
    end if
  end for
$T_{\text{new}} \leftarrow \{t_k^{\text{new}}\}_{k=1}^{n_u}$;
end if
Output: $T \leftarrow T_{\text{old}} \cup T_{\text{new}}$.

PMU: Power Management Unit

03. 실험 방법

- 실험 방법
  - Online random forests regression with memories
    - 일반적인 RFs regression(RFR), OWL-RFR, Mondrian Forests(MF)의 성능을 Mean Squared Error(MSE)와 R^2 Score로 비교함
    - Leaf-level weight vs. tree-level weight
      - OWL-RFR, RFR and Online Tree Weight Learning for RFs Regression(OTWL-RFR)의 MSE들을 비교함.
    - 제안된 모델의 stability를 확인하기 위해 동일한 testing data set을 재배치하여 5개의 순서만 다른 testing set들을 만들고 OWL-RFR과 RFR의 accumulated loss의 차이를 비교함

  - Calculating Feature Importance in Data Streams with Concept Drift using Online Random Forest
    - Streaming algorithm의 성능 테스트에 많이 쓰이는 rotating hyperplane dataset을 사용함
    - ORF의 MDA, MDG와 sliding window 기법으로 훈련된 RF의 MDA, MDG를 비교함
    - ANOVA와 Pairwise t-test로 유의성을 검증함

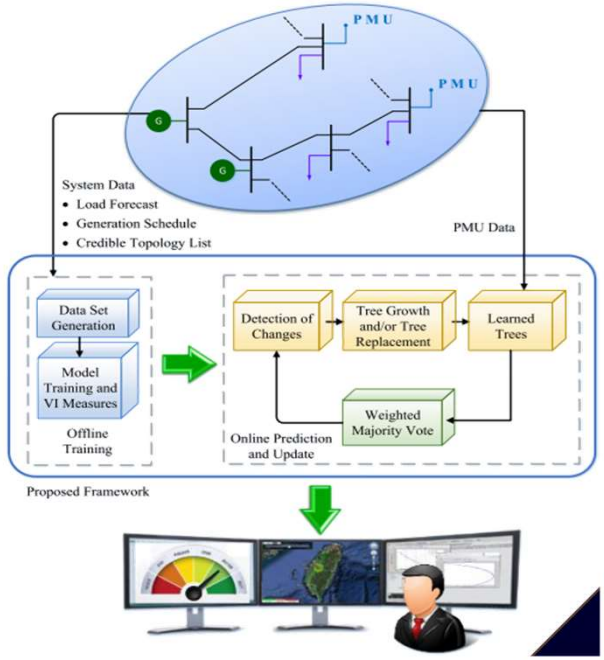
03. 실험 방법

- 실험 방법
  - Online adaptive decision trees based on concentration inequalities
    - 모델의 안정성, 시간에 따른 학습의 진화, 알고리즘의 오류율을 비교하면서 평가함
    - Moa software를 통해 vfdt* 패키지 알고리즘 시스템을 활용해서 성능을 평가함 (평가 속도를 올리기 위해)
    - 기존 idaem-2의 가지치기 방법을 이용해서 하위 분류까지 오류율을 평가함
    - iadem-3의 성능을 알아보기 위해 기존에 제시된 VFDT, iadem-2, HAT, OnlineTree2 방법과의 성능 비교
      - iadem-3의 성능의 정확도가 가장 높게 측정됨

vfdt: Very Fast Decision Tree

03. 실험 방법

- 실험 방법
  - Enhanced-Online-Random-Forest Model for Static Voltage Stability Assessment Using Wide Area Measurements
    - ➢ 기존 offline training으로 나온 데이터에서 성립된 알고리즘에서 실시간 데이터를 통해 점진적으로 개선시킴
    - ➢ 점진적으로 개선할 때, 가지치기로 나온 구간의 정확도에 따른 가중치를 두면서 개선시킴
    - ➢ 가중치를 통해 트리를 개선하면서 가중화 및 교체를 함
    - ➢ 최종적으로 kNN, SVM, DT, AdaBoost와 같이 비교 성능을 함
      - ❖ 제안된 EORF Model이 PMU Dataset 예측 성능에서 가장 우수하게 측정됨



Thank you