Select start point for ARF analysis

Kenichi YOSHIDA*
*University of Tsukuba, Ibaraki, Japan
Email: yoshida.kenichi.ka@u.tsukuba.ac.jp

***Abstract*—It is important to analyze online data whose characteristic changes over time, such as financial and coronavirus infection data. Many studies have been conducted. In general, ensemble-based learning methods perform well as analysis methods, and methods such as SEA, DWM, and ARF have been proposed. In addition, the change in characteristics over time is called concept drift, and its classification and detection methods have been studied. This paper reports the adverse effects of a characteristic that has yet to be considered in the classification of concept drifts and proposes a solution. The characteristic discussed in this paper is the ratio of explanatory variables whose characteristics change. This paper shows that when this ratio is large, it harms the ensemble-based learning methods. In addition, this paper evaluates a solution method to show that the method can improve the accuracy of the analysis. Since a naive implementation of the proposed method is inefficient, this paper also reports a beam search implementation.**

***Index Terms*—Online data, Concept drift, Ensemble learning, Exhaustive analysis**

I. Introduction

Analyzing online data, such as financial and coronavirus infection, is important. A typical characteristic of these online data is that the characteristics of these data change over time. This characteristic change over time is called concept drift, and many studies have been conducted (See [1] and [2] for surveys). In such studies, its classification methods [3] [4], drift detection method (DDM [5] and Adwin [6], etc.) are exhaustively studied. Furthermore, methods for comparing various analysis methods are also studied [7]. As far as we surveyed, the misclassification rate of ensemble learning methods, such as Adaptive random forests (ARF) [8], tends to be small.

This report will introduce a characteristic that needs better analysis in concept drift studies. This paper reports its adverse effects on ensemble learning methods and proposes a solution. Since a naive implementation of the proposed solution is inefficient, this paper also reports a beam search implementation.

The mainstream research has categorized concept drifts into four types: "Sudden", "Gradual", "Incremental", and "Reoccurring" [2]. "Sudden" changes the relationship between the explanatory and explained variables at some point. "Gradual" and "Incremental" change the relationship slowly. "Reoccurring" repeats the changes multiple times. Unlike previous studies, this paper argues that "the ratio of explanatory variables whose characteristics change" is important for classifying concept drift. No studies, as far as we surveyed, discussed the importance and adverse effects of similar classifications of concept drift.

In the discussion below, the explained variable y has a binary true or false value, and the explanatory variable $\mathbf{x}$ is an n-dimensional scalar vector. Among the scalar values composing the vector of explanatory variable $\mathbf{x}$, only n'($\leq$n) values are assumed to affect the value of explained variable y. The remaining (n-n') values are unrelated to the relationship between the explained variable y and the explanatory variable $\mathbf{x}$. Furthermore, concept drift is assumed to be caused by only one of n' scalar values changing its characteristics." For example, if we learn the relationship between y and $\mathbf{x}$ as an n-dimensional linear discriminant function, only one of the n regression coefficients is assumed to change.[1]

It is a strong assumption that only one regression coefficient changes. However, considering a situation where the scalar values that make up the vector of explanatory variables $\mathbf{x}$ are independent, it is natural to assume that they do not change their characteristics simultaneously. We think this is a reasonable assumption to start discussing for "The ratio of explanatory variables whose characteristics change is important for the classification of concept drift." In other words, "the ratio of explanatory variables whose characteristics change," which is the focus of discussion in this study, is 1/n'.

In this paper, after Section II surveys related works, Section III explains our hypothesis of the importance of "the ratio of explanatory variables whose characteristics change" on concept drift. Then, Section IV proposes a method to reduce the misclassification rate of ARF when "The ratio of explanatory variables whose characteristics change" is large. Since a naive implementation of the proposed method is inefficient, Section IV also reports a beam search implementation and confirms the proposed method's advantage through experiment. Finally, Section V summarizes our findings.

II. Related work

A. Classification of Concept Drift

As mentioned above, the characteristics of financial data, coronavirus infection data, etc., change over time. Analysis of changing online data is important, and numerous studies are being conducted (See [1] and [2] for surveys).

The classification of concept drift is also explained in [2]. The mainstream research has categorized concept drifts into

[1]It is assumed that only n'($\leq$n) values affect the value of the explained variable y. So, to be precise, "Only one of the n' regression coefficients changes, and (n'-1) regression coefficients do not change. The remaining (n-n') regression coefficients remain 0."

This work was supported by JSPS KAKENHI Grant Number 24K15082.

979-8-3503-5475-1/24/$31.00 ©2024 IEEE

four types: “Sudden”, “Gradual”, “Incremental”, and “Reoccurring” [2]. “Sudden” changes the relationship between the explanatory and explained variables at some point. “Gradual” and “Incremental” change the relationship slowly. “Reoccurring” repeats the changes multiple times.

Unlike previous studies, this paper argues that “the ratio of explanatory variables whose characteristics change” is important for classifying concept drift. As far as we surveyed, no studies discussed the importance of similar classifications of concept drift.

B. Ensemble learning for Concept Drift Analysis

As far as we surveyed, the misclassification rate of ensemble learning methods on data with concept drift is small. Thus, various methods have been studied, such as Streaming ensemble algorithm (SEA) [9], Weighted ensemble classifiers [10], Dynamic weighted majority (DWM) [11], Adaptive classifiers ensemble system (ACE) [12], Dynamic streaming random forests [13], Adwin bagging and ASHT bagging [14], Leveraging bagging [15], Online smooth-boost (OSBoost) [16], Online accuracy updated ensemble (OAUE) [17], and Compacted object sample extraction (COMPOSE) [18]. Among them, there are many cases where the misclassification rate of ARF [8] is lower than that of other methods. Its low misclassification rate also has been reported in [7].

Figure 1 shows the algorithm of ARF. ARF creates a model like other online learning methods by ensembling multiple weak learners. As far as we understand, the characteristics of ARF are: 1) It is based on Random Forests [19] as an ensemble method (lines 10 and 16 of Figure 1); 2) While other methods change the corresponding weak learner to a new one when they detect concept drift for each weak learner, ARF creats a backup tree for the weak learner when concept drift is suspected (lines 11 and 12 of Figure 1). When concept drift is detected, ARF replaces the weak learner with the backup tree (lines 13 and 14 of Figure 1).

III. Preliminary experiments and hypotheses

A. Comparison of misclassification rate

Table I shows the results of preliminary experiments that compare misclassification rates of previously proposed drift detection methods: i.e., DDM, Adwin, ARF, and Without (analyze data using CART without detecting drift). ARF is a representative ensemble-based learning method, and DDM and Adwin are non-ensemble-based learning methods. Here, scikit-learn [20] and scikit-multiflow [21] were used for the experiments, and default values were used for parameters without optimization.

Souza et al. [7] collected the data used in the experiments to compare misclassification rates of previously proposed drift detection methods. The results in the table are those of re-experiments by the authors of this paper using CART (weka was used in the original paper [7]). Data exceeding 10,000 pieces in length were sampled at equal intervals to reduce experimental time, and the experiment was performed using a maximum of 10,000 pieces of data. As shown in Table I,

m: Maximum features evaluated per split
n: Total number of trees (n = |T|)
δ_w: Warning threshold
δ_d: Drift threshold
c(·): Change detection method
S: Data stream
B: Set of background trees
W(t): Tree t weight
P(·): Learning performance estimation function

1: **procedure** ARF(m, n, δ_w, δ_d)
2: T ← CreateTrees(n)
3: W ← Initialize Weights(n)
4: B← ∅
5: **while** Has Next(S) **do**
6: (x, y) ← Next(S)
7: **for** all t ∈ T **do**
8: $\hat{y}$ ← predict(t, x)
9: W(t) ← P($W(t), \hat{y}, y$)
10: RFTreeTrain(m, t, x, y)
11: **if** C(δ_w, t, x, y) **then** ▷ Warning detected?
12: B(t) ← Create background tree for t
13: **if** C(δ_d, t, x, y) **then** ▷ Drift detected?
14: t ← B(t) ▷ Replace t by background
15: **for** all b ∈ B **do**
16: RFTreeTrain(m, b, x, y)

Fig. 1. Algorithm of ARF (from [8])

TABLE I
Comparison of misclassification rates using published data

ID	Data Name	without	DDM	Adwin	ARF
1	insect-abrupt (bal)	53.9	47.9	41.2	**31.7**
2	insect-abrupt (imbal)	38.1	37.4	36.6	**26.7**
3	insect-gradual (bal)	56.2	40.9	33.4	**25.0**
4	insect-gradual (imbal)	40.2	35.3	33.1	**24.6**
5	insect-inc-abrupt (bal)	50.8	45.7	45.3	**30.3**
6	insect-inc-abrupt (imbal)	38.0	39.0	39.4	**29.3**
7	insect-inc-reoc (bal)	56.0	55.7	40.7	**27.7**
8	insect-inc-reoc (imbal)	38.7	38.8	40.1	**29.3**
9	insect-inc (bal)	53.6	53.2	52.3	**39.3**
10	insect-inc (imbal)	39.1	38.2	35.4	**26.1**
11	insect-out-of-control	41.9	41.9	41.9	**38.8**
12	NOAA	27.5	27.6	27.4	**22.8**
13	airlines	40.2	40.1	40.9	**38.5**
14	chess	**31.1**	**31.1**	31.2	33.0
15	covtype	32.9	32.5	31.1	**26.8**
16	elec	22.9	22.7	21.5	**18.3**
17	gassensor	37.3	21.5	16.4	**4.0**
18	kddcup99	1.4	1.4	1.4	**1.0**
19	poker-lsn	43.7	41.8	38.6	**34.3**
20	powersupply	86.0	85.5	85.9	**82.5**
21	rialto	67.6	67.6	68.7	**63.9**
22	sensorstream	93.0	93.0	92.3	**81.2**
23	keystroke	13.0	13.0	13.1	**6.1**
24	luxembourg	**0.00**	**0.00**	**0.00**	2.3
25	outdoor	88.8	23.0	77.9	**20.9**
26	ozone	7.2	7.6	7.6	**6.4**

Underline is the minimum misclassification rate (%)
The misclassification rates are the averages of 10 experiments.

the preliminary experiments obtained roughly the same results reported in [7]:

- The misclassification rate of ARF is the lowest for many problems (specifically, 24 out of 26 data).
- However, there are problems with DDM being better than ARF (Data 14 and 24).
- Also, as pointed out by [7], there were cases where the drift detection methods could not reduce the misclassification (Data 14 and 24).

B. Interpretation of preliminary experiments

As described above, the ARF's misclassification rate is often lower than that of non-ensemble methods such as DDM. Based on this experimental result, this paper assumes the following hypothesis:

Hypothesis 1:

> In addition to the concept drift classification previously studied, i.e., Sudden, Gradual, Incremental, and Reoccurring [2], "The ratio of explanatory variables whose characteristics change" is essential. If this ratio is small, ensemble learning-based methods have the advantage of analyzing the data.

Hypothesis 2:

> If this ratio is large, the misclassification rate of ensemble learning-based methods might be high.

When 1) data that includes concept drift are to be analyzed, and 2) there are many explanatory variables, and 3) only one of them changes its characteristics and causes its concept drift; "The ratio of explanatory variables whose characteristics change" is small. If such data is analyzed using the procedure shown in Figure 1, the tree with the highest importance to the explanatory variable whose characteristic changed will be replaced by its backup. The learning results regarding the explanatory variables whose characteristics have not changed are maintained by continuously using other trees.

On the other hand, DDM and Adwin observe changes in the misclassification rate. When the deterioration of the misclassification rate exceeds a threshold, learning is restarted using data from that point.[2] Thus, DDM and Adwin must re-learn the explanatory variables whose characteristics have stayed the same. This may increase their misclassification rates.

Ensemble learning methods such as ARF eliminate the need to re-learn the discriminative aspects of explanatory variables whose characteristics have remained the same. This reduces the ensembled model's misclassification rate to less than that of the methods that re-learn from that point. On the other hand, if "the ratio of explanatory variables whose characteristics change" is large, the characteristics of many explanatory variables change at the same time. Restarting from that point is better than replacing only one part of the tree. In other words, in the ensemble learning methods, the presence of weak learners that still need to be replaced may reduce the learning speed.

[2]Strictly speaking, there are two types of thresholds: "warning" and "detection". At the time of "detection", the data is traced back to the point where the "warning" threshold is reached. However, explanatory variables remain whose characteristics have not changed before the "warning" point.

IV. Relationship between concept drift analysis and the ratio of explanatory variables whose characteristics change

This section reports the experiments to verify the hypothesis explained in the previous section and proposes a method designed based on that hypothesis.

A. Artificial data to verify hypothesis

If "the ratio of explanatory variables whose characteristics change" affects the misclassification rate of ensemble learning methods such as ARF, the effect can be verified by examining the misclassification rate of artificial data by changing the ratio. In addition, two types of "change in characteristics" are considered: A case in which the characteristic suddenly changes at a particular time and a case in which it changes slowly. Based on the above idea, the following artificial data 1-1∼5 and 2-1∼5 are generated.

Random Sequence 1∼5

> Random number sequence with mean 0, standard deviation 1, and length 600.

Explanatory variable sequences 1-2∼5 and 2-2∼5

> Random sequence 2∼5 itself.

Explanatory variable 1-1

> Add 0, 0.5, 1.0, 1.5, 2.0, 2.5 to the data from 1-100, 101-200, 201-300, 301-400, 401-500, 501-600 of Random Sequence 1.

Explanatory variable 2-1

> A random number sequence that adds "0.005*(number of positions in the sequence)" to Random Sequence 1.

data 1-n

> A data sequence comprises 600 pairs of five explanatory variables and a true/false explained variable. The value of the i-th explanatory variable is created from explanatory variable 1-i, and the explained variable is
>
> $$(\sum_{i=1}^{n} \text{variable in random sequence i}) > 0$$

data 2-n

> A data sequence comprises 600 pairs of five explanatory variables and a true/false explained variable. The value of the i-th explanatory variable is created from the explanatory variable 2-i, and the explained variable is
>
> $$(\sum_{i=1}^{n} \text{variable in random sequence i}) > 0$$

The true or false value of the explanatory variables is determined by whether the sum of the corresponding variables in the random sequence is positive or negative. In addition, data 1-1∼5 is divided into six sections. Since the difference between the explanatory variable given to the learning system and variables in random sequence changes step-wise, the identification aspect of explanatory variable 1-1∼5 changes step-wise. For data 2-1∼5, the identification aspect of explanatory variable 2-1∼5 changes gradually for each item.

In these data, the number of explanatory variables whose characteristics change due to the difference from the variable used to calculate the value of the explained variable is always 1. On the other hand, the number n' of explanatory variables related to the value of the explained variable changes to 1∼5. Thus, "the ratio of explanatory variables whose characteristics change" changes from 1/1, the maximum value by definition, to 1/5.

For example, the truth value of data 1-3 is determined depending on whether the sum of random variables 1, 2, and 3 is positive or negative (n'=3). Explanatory variables 2 and 3 of data 1-3 have the values used to calculate the truth value of the explained variable. Their identification aspect remains the same. On the other hand, the identification surface of explanatory variable 1-1 changes every 100 items. "The ratio of explanatory variables whose characteristics change" for this data 1-3 is 1/3.

B. Proposed methods and Experiment for verification

To check the hypothesis, the misclassification rate of simple CART ("without" in Table II), DDM, Adwin, ARF, and enhancement of ARF ("SARF-b" and "SARF-e" in Table II) are measured using above artificial data. ARF (ARF) was chosen as a representative ensemble-based learning method. DDM and Adwin were chosen as representative non-ensemble-based learning methods.

If Hypothesis 1 is correct, the misclassification rate of ARF will increase for data 1-1 and data 2-1 where "the ratio of explanatory variables whose characteristics change" is large, i.e., 1. Also, the misclassification rate of ARF will decrease for data 1-5 and data 2-5 where "the ratio of explanatory variables whose characteristics change" is small, i.e., 1/5. Furthermore, the misclassification rates of DDM and Adwin are also investigated to compare them with ARF. If Hypothesis 1 is correct, the misclassification rates of DDM and Adwin would be small when "the ratio of explanatory variables whose characteristics change" is large, e.g., 1,

To check Hypothesis 2 and propose a counter-measurement, this paper introduces a mechanism to restart ARF's learning. SARF-e (Select start point for ARF - Exhaustively) and SARF-b (Select start point for ARF - by Beam search) are the methods that introduce the restart mechanism into ARF. At each time point t, SARF-e checks the misclassification rate of all the possible ARF models (See Figure 2 for the idea and Figure 3 for pseudo code). Here, ARF models are generated by changing the start point of learning from 0 to t-1 (Figure 3 line 5). If the assumed start point is t', the misclassification rate of the corresponding model is calculated based on the results from time t' to t-1 (Figure 3 line 7). Supposing the ARF model that starts learning from time point T has a minimum misclassification rate, data at time point t is analyzed using the ARF model that starts learning from time point T (Figure 3 line 13). Selection of the best time point T for learning works as the restart mechanism.

Since SARF-e's computing cost is O(t^2) (See double loop in Figure 3 line 5 & 6), SARF-b tries to reduce it using beam

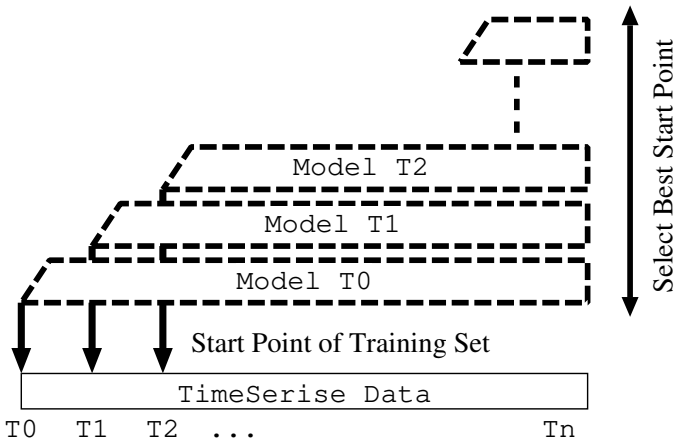


Fig. 2. Basic idea of SARF-e

```
1: function MODEL(y, x)
2:     return ARF Model for estimating y_t from x_{t-1}
                                                    ▷
3: function SELECT(y, x)
4:     remove last item from both y and x
5:     for s in [all possible start point in y] do
6:         for l in [positions in "subsequence of y
                                    starting from s"] do
7:             m = MODEL([subsequence of y from s to l],
                          [subsequence of x from s to l])
8:             est[s, l] = m(x[l])
9:     return s that gives minimum error rate in est[s,:]
                                                    ▷
10: function SARF-E(y, x)
11:     s = SELECT(y, x)
12:     m = MODEL([subsequence of y start from s],
                  [subsequence of x start from s])
13:     return m(x_t)          ▷ i.e., return estimation of y_{t+1}
```

Fig. 3. Algorithm of SARF-e

```
1: FIFO: array of ARF models
2: BEAM: array of ARF models
3: BEST: best ARF model in BEAM
                                                    ▷
4: function SARF-B PREDICTION(NewData)
5:     return prediction of NewData based on BEST
                                                    ▷
6: procedure SARF-B LEARNING(NewData)
7:     Make new ARF model NEW using NewData.
8:     Update ARF models in FIFO and BEAM using NewData.
9:     Exclude oldest ARF models in FIFO as OLD
10:    Add NEW into FIFO
11:    Add OLD into BEAM
12:    Select best ARF models among those in BEAM as BEST
13:    Remove worst ARF model from BEAM
```

Fig. 4. Algorithm of SARF-b

TABLE II
COMPARISON OF MISCLASSIFICATION RATES USING ARTIFICIAL DATA

ID	without	DDM	Adwin	ARF	SARF-e	SARF-b
Data1-1	25.4	**7.2**	21.8	19.2	8.6	10.2
Data1-2	20.3	20.6	20.3	19.0	**15.9**	17.3
Data1-3	21.7	21.7	21.6	18.7	18.8	**18.3**
Data1-4	23.7	23.8	23.8	**19.6**	21.9	21.4
Data1-5	29.8	30.0	29.9	**20.7**	23.2	23.0
Data2-1	27.8	11.5	22.4	21.6	**9.7**	10.7
Data2-2	20.1	20.0	20.0	20.2	**15.6**	16.4
Data2-3	23.0	22.9	22.9	**18.4**	19.7	19.6
Data2-4	22.5	22.3	22.7	**18.7**	21.7	22.4
Data2-5	29.9	29.9	29.8	**20.9**	23.4	22.9

Bold double underline is the minimum misclassification rate (%)
Single underline is runner-up.
The misclassification rates are the averages of 10 experiments.

search (See Figure 4 for pseudo code). SARF-b selects the best n ARF models at each time point t (Figure 4 line 13). Then, it checks those n ARF models at the next time point t+1 (Figure 4 line 8). As shown in Figure 4, SARF-b also checks the most recent n' models (Handling of FIFO in Figure 4) since the misclassification accuracy of the recent model tends to be unstable. Also, our implementation of SARF-b does not exclude the oldest model from BEAM in line 13 Figure 4. Since the oldest model in BEAM and the model of ARF are the same, this implementation makes comparing SARF-b with ARF easy.

C. Results and considerations of experiments using artificial data

Table II shows the experimental results. The misclassification rates in the table are the averages of 10 experiments. The table shows that the results generally support Hypotheses 1 and 2. That is:

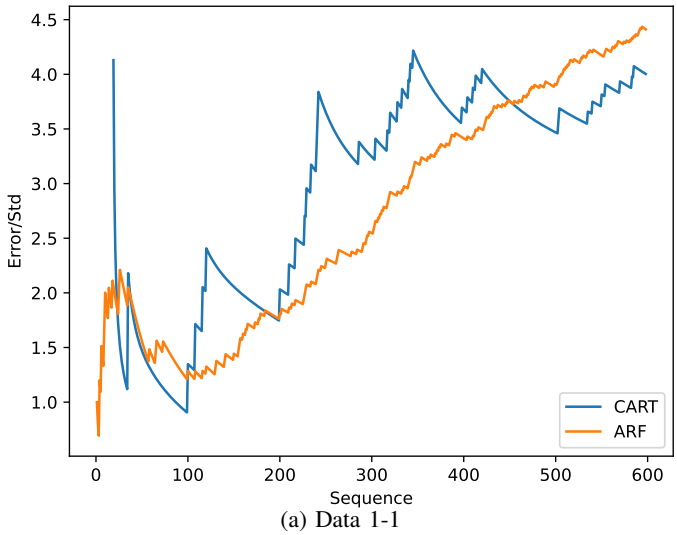
- As Hypothesis 1 suggests, the misclassification rate of ARF is lower than that of DDM and Adwin for data 1-3∼5 and data 2-3∼5, which have a small "ratio of explanatory variables whose characteristics change".
- For data 1-1 and 2-1, which have a large "ratio of explanatory variables whose characteristics change", the misclassification rate of ARF is higher than that of DDM.

Furthermore

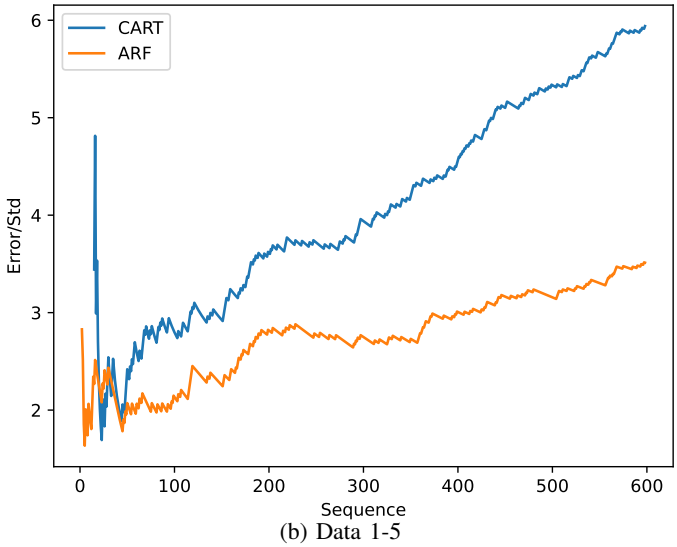
- Both SARF-e and SARF-b reduce the misclassification rates of ARF for data 1-1∼2 and data 2-1∼2, which have a large "ratio of explanatory variables whose characteristics change."

In addition to the proposition set as a hypothesis, the following can be gleaned from the results.

- The effect of reducing the DDM misclassification rate is large for data 1-1 and data 2-1, where "the ratio of explanatory variables whose characteristics change" is large. This is because the drift significantly changes the misclassification rate. Thus, the drift is easily detected by DDM, which depends on changes in the misclassification rate.



(a) Data 1-1



(b) Data 1-5

Fig. 5. Misclassification rate of CART and ARF

Figure 5 shows the change in misclassification rate when CART and ARF are used as learning methods for data 1-1 and 1-5. When data 1-1 is analyzed using CART (Figure 5(a)), the misclassification rate changes significantly at the time when drift occurs.
On the other hand, when data is analyzed using ARF or "the ratio of explanatory variables whose characteristics change" is small as data 1-5, the misclassification rate changes slowly. It makes the detection by DDM difficult. Also, ARF supports drift handling, which makes drift points unclear.

- The exhaustive search method, i.e., SARF-e, is a method that exhaustively tests the learning start point and makes the prediction based on the model learned using the found best start point (Figure 2).
Initially, we thought that "the learning start point where the result of one point before is the best" is the same as the "starting point of the drift". However, SARF-

e sometimes fails to decrease the misclassification of ARF. We believe this is because "the learning start point where the result of one point before is the best" is not the "starting point of the drift", i.e., "best learning start point." This, i.e., the difference between "the learning start point where the result of one point before is the best" and "best learning start point", is left as a future research issue.

In the experiments, the beam size of SARF-b is set to 50. Precisely, the length of FIFO (See Figure 4) is set to 10, and the length of BEAM is set to 40. The length of FIFO was selected to stabilize the misclassification of ARF. While the period is short, the misclassification of ARF varies. However, ten seems enough in our experiments to get a stable misclassification rate. Figure 6 shows the effect of the length of BEAM. In Figure 6, the horizontal axis is the length of BEAM, and the vertical axis is the number of misclassifications in the preliminary experiments. Although the number of misclassifications varies depending on the BEAM length, 40 was chosen in the experiments shown in Table II.

D. Experimental Results on published data

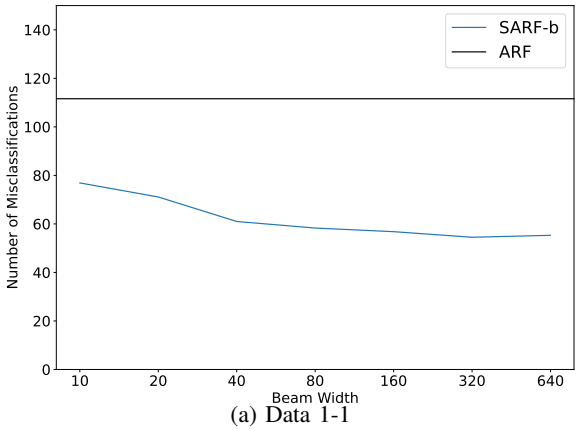
Table III compares the misclassification rates of SARF-e and SARF-b with previous methods on public data [7]. SARF-e has 10 out of 26 data with the lowest misclassification rate, and SARF-b has 9 out of 26 data with the lowest misclassification rate. SARF-e or SARF-b reduced the misclassification rate of ARF on 20 data. These results clearly show the effect of the mechanism to restart ARF's learning.

Here, unlike artificial data, the "ratio of explanatory variables whose characteristics change" in public data is unknown. It is also unclear whether the assumption made at the beginning of this paper that "only one explanatory variable has changed" characteristics is satisfied. In the published data, independence among variables is not mentioned, and the possibility remains that multiple explanatory variables are changing their characteristics simultaneously.

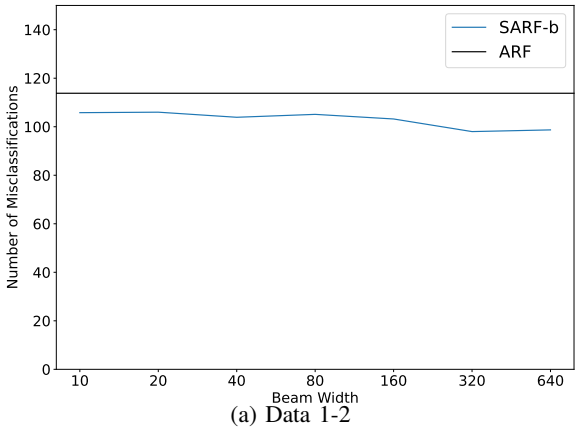
In this sense, the significance of experimental results using public data is limited in verifying the hypothesis that is the subject of this paper. The characteristics of public data require further analysis. However, experimental results show that there is public data that SARF-e and SARF-b improve the misclassification rate of ARF. This is considered a result that reinforces the validity of the research.

Furthermore, SARF-e, which combines ARF and the exhaustive search method, must improve its long execution and processing times. Table IV shows ARF, SARF-e, and SARF-b elapsed time. While ARF completes its processing in seconds, SARF-e requires several hours, depending on the data. This study purposely investigated the misclassification rate of SARF-e in order to verify Hypothesis 2, However, SARF-e is difficult to use when solving large-scale problems practically.

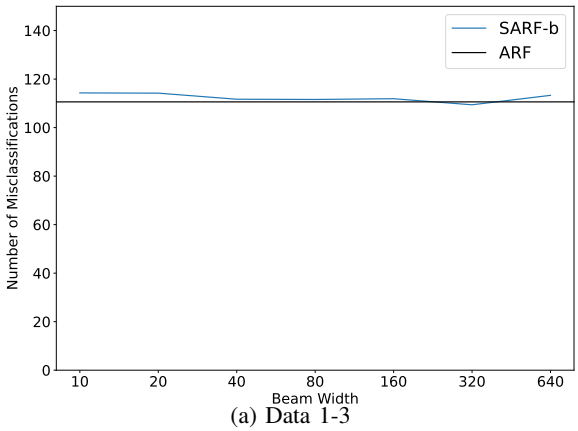
In contrast, the elapsed time of SARF-b is much shorter due to multiprocessing (Table IV). However, even though the CPU used in the experiment supports 64 threads and is larger



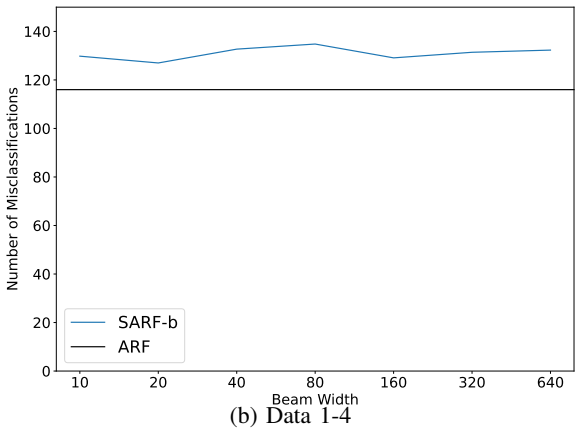
(a) Data 1-1



(a) Data 1-2



(a) Data 1-3



(b) Data 1-4

Fig. 6. Effect of BEAM width on SARF-b's misclassification

TABLE III
COMPARISON OF MISCLASSIFICATION RATES USING PUBLISHED DATA

ID	without	DDM	Adwin	ARF	SARF-e	SARF-b
1	53.9	47.9	41.2	31.7	**29.6**	29.7
2	38.1	37.4	36.6	26.7	26.6	**26.5**
3	56.2	40.9	33.4	25.0	**23.4**	23.8
4	40.2	35.3	33.1	24.6	24.3	**24.1**
5	50.8	45.7	45.3	30.3	**28.7**	29.6
6	38.0	39.0	39.4	**29.3**	29.9	29.4
7	56.0	55.7	40.7	27.7	**26.6**	26.9
8	38.7	38.8	40.1	**29.3**	30.3	29.6
9	53.6	53.2	52.3	39.3	40.2	**39.1**
10	39.1	38.2	35.4	26.1	26.7	**26.0**
11	41.9	41.9	41.9	38.8	38.2	**37.8**
12	27.5	27.6	27.4	**22.8**	25.1	26.0
13	40.2	40.1	40.9	**38.5**	40.9	41.1
14	**31.1**	**31.1**	31.2	33.0	32.8	32.6
15	32.9	32.5	31.1	26.8	**24.2**	24.6
16	22.9	22.7	21.5	18.3	**17.0**	17.2
17	37.3	21.5	16.4	4.0	3.9	**3.8**
18	1.4	1.4	1.4	1.0	0.9	**0.8**
19	43.7	41.8	38.6	**34.3**	36.7	37.7
20	86.0	85.5	85.9	82.5	**71.7**	73.8
21	67.6	67.6	68.7	63.9	59.3	**59.0**
22	93.0	93.0	92.3	81.2	**78.7**	79.6
23	13.0	13.0	13.1	6.1	**5.2**	5.7
24	**0.00**	**0.00**	**0.00**	2.3	1.2	0.6
25	88.8	23.0	77.9	**20.9**	22.5	23.0
26	7.2	7.5	7.6	6.4	**6.2**	**6.2**

Underline is the minimum misclassification rate (%)
The misclassification rates are the averages of 10 experiments.

TABLE IV
ELAPSED TIME

ID	Length	Attributes	ARF(sec)	SARF-e(sec)	SARF-b(sec))
1	10000	33	176.3	26708.9	687.5
2	10000	33	154.8	24391.3	705.3
3	10000	33	136.3	24214.7	626.5
4	10000	33	151.9	23883.1	665.9
5	10000	33	145.8	23113.4	658.1
6	10000	33	162.6	23905.8	712.6
7	10000	33	149.8	21084.5	645.5
8	10000	33	152.6	24387.3	718.5
9	10000	33	185.9	26320.9	769.9
10	10000	33	160.2	23730.0	712.8
11	10000	33	458.0	65203.2	1421.9
12	10000	8	57.9	11471.7	346.4
13	10000	7	63.0	11046.0	316.0
14	534	7	4.6	26.5	18.9
15	10000	54	90.3	15925.9	459.8
16	10000	8	56.0	10771.8	347.9
17	10000	128	172.6	27544.6	775.9
18	10000	41	72.9	11101.8	434.9
19	10000	11	91.8	12718.4	360.3
20	10000	2	65.4	8672.6	337.1
21	10000	27	182.9	26075.2	766.0
22	10000	5	344.7	43534.2	1149.2
23	1600	10	13.8	212.5	57.7
24	1901	30	13.7	317.7	77.4
25	4000	21	119.2	6315.8	363.5
26	2534	72	26.2	787.8	115.5

AMD Ryzen Threadripper 2990WX
128GByte memory (32 Core 64 Thread)
The Elapsed times are the averages of 10 experiments.

than the BEAM size, SARF-b requires 3∼4 times more time than ARF. This is because the processing time of ARF at each point in SARF-b varies greatly.

V. CONCLUSION

This paper investigates the influence of the "ratio of explanatory variables whose characteristics change" on online data analysis, including concept drift.

- Previous studies classify concept drift into four classes: e.g. Sudden, Gradual, Incremental, and Reoccurring. In addition to this traditional concept drift classification, this paper reports a new aspect of concept drift classification; the "ratio of explanatory variables whose characteristics change".
- Ensemble learning methods are excellent at analyzing online data with concept drift where "the ratio of explanatory variables whose characteristics change" is small.
- Ensemble learning methods may have a high misclassification rate when analyzing online data with a large ratio of "the explanatory variables whose characteristics change".

This was confirmed through experiments using artificial data.

In addition, this paper proposes methods to search for the start point of ARF analysis as a countermeasure against the increase in the misclassification rate when "the ratio of explanatory variables whose characteristics change" is large. SARF-e exhaustively searches for the best start point, and SARF-b searches for it by beam search. SARF-e and SARF-b reduced the misclassification rate of ARF on 20 out of 26 public data.

REFERENCES

[1] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM computing surveys (CSUR)*, vol. 46, no. 4, pp. 1–37, 2014.

[2] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, "Learning under concept drift: A review," *IEEE transactions on knowledge and data engineering*, vol. 31, no. 12, pp. 2346–2363, 2018.

[3] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern recognition*, vol. 45, no. 1, pp. 521–530, 2012.

[4] G. I. Webb, L. K. Lee, B. Goethals, and F. Petitjean, "Analyzing concept drift and shift from sample data," *Data Mining and Knowledge Discovery*, vol. 32, pp. 1179–1199, 2018.

[5] J. Gama, P. Medas, G. Castillo, and P. Rodrigues, "Learning with drift detection," in *Advances in Artificial Intelligence–SBIA 2004: 17th Brazilian Symposium on Artificial Intelligence, Sao Luis, Maranhao, Brazil, September 29-Ocotber 1, 2004. Proceedings 17*. Springer, 2004, pp. 286–295.

[6] A. Bifet and R. Gavaldà, "Learning from time-changing data with adaptive windowing," vol. 7, 04 2007.

[7] V. M. A. Souza, D. M. Reis, A. G. Maletzke, and G. E. A. P. A. Batista, "Challenges in benchmarking stream learning algorithms with real-world data," *Data Mining and Knowledge Discovery*, vol. 34, pp. 1805–1858, 2020.

[8] H. M. Gomes, A. Bifet, J. Read, J. P. Barddal, F. Enembreck, B. Pfharinger, G. Holmes, and T. Abdessalem, "Adaptive random forests for evolving data stream classification," *Machine Learning*, vol. 106, pp. 1469–1495, 2017.

[9] W. N. Street and Y. Kim, "A streaming ensemble algorithm (sea) for large-scale classification," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 2001, pp. 377–382.

[10] H. Wang, W. Fan, P. S. Yu, and J. Han, "Mining concept-drifting data streams using ensemble classifiers," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003, pp. 226–235.

[11] J. Z. Kolter and M. A. Maloof, "Dynamic weighted majority: An ensemble method for drifting concepts," *The Journal of Machine Learning Research*, vol. 8, pp. 2755–2790, 2007.

[12] K. Nishida and K. Yamauchi, "Adaptive classifiers-ensemble system for tracking concept drift," in *2007 International Conference on Machine Learning and Cybernetics*, vol. 6, 2007, pp. 3607–3612.

[13] H. Abdulsalam, D. B. Skillicorn, and P. Martin, "Classifying evolving data streams using dynamic streaming random forests," in *Database and Expert Systems Applications*, S. S. Bhowmick, J. Küng, and R. Wagner, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 643–651.

[14] A. Bifet, G. Holmes, B. Pfahringer, R. Kirkby, and R. Gavalda, "New ensemble methods for evolving data streams," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 139–148.

[15] A. Bifet, G. Holmes, and B. Pfahringer, "Leveraging bagging for evolving data streams," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part I 21*. Springer, 2010, pp. 135–150.

[16] H.-T. L. Shang-Tse Chen and C.-J. Lu, "An online boosting algorithm with theoretical justifications," in *ICML'12: Proceedings of the 29th International Coference on International Conference on Machine Learning*, 2012, pp. 1873–1880.

[17] D. Brzezinski and J. Stefanowski, "Combining block-based and online methods in learning ensembles from concept drifting data streams," *Information Sciences*, vol. 265, pp. 50–67, 2014.

[18] J. Sarnelle, A. Sanchez, R. Capo, J. Haas, and R. Polikar, "Quantifying the limited and gradual concept drift assumption," in *2015 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2015, pp. 1–8.

[19] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[21] J. Montiel, J. Read, A. Bifet, and T. Abdessalem, "Scikit-multiflow: A multi-output streaming framework," *Journal of Machine Learning Research*, vol. 19, no. 72, pp. 1–5, 2018. [Online]. Available: http://jmlr.org/papers/v19/18-251.html