An Active Learning Method for Data Streams with Concept Drift

Cheong Hee Park and Youngsoon Kang
Department of Computer Science and Engineering
Chungnam National University
Daejeon, Korea
Email: cheonghee@cnu.ac.kr, yskang@hanafos.com

***Abstract*—In analyzing streaming data in which the underlying data distribution may change or the concept of interest may drift over time, the ability of a classifier to adapt to drifted concepts is very important to maintaining the prediction performance. However, the true class labels of data samples are often available only after some period of time or they are obtained by experts' efforts. In this paper, we develop an effective method for active learning on data streams with concept drift. The proposed method combines active learning and adaptive incremental learning. For unlabeled data samples, the degree of concept drift is estimated and used for both data selection for labeling and adaptive incremental learning of the current classifier. Experimental results on five artificial data sets and two real data sets demonstrate a competent performance of the proposed method.**

***Index Terms*—Active learning; Adaptive incremental classifier; Concept drift; Streaming data;**

I. Introduction

A data stream is a sequence of data samples that is continuously generated as time goes on [1]. In dynamically changing environments, the underlying data distribution may be changed or the concept of interest can drift over time. Hence, updating a classifier appropriately so as to reflect the recent trend is very important. However, in some applications, obtaining true class labels is costly and time consuming. The true class labels of new data samples are often available only after some period of time or they are obtained by experts' efforts.

Active learning aims to select the most informative data samples to train a classifier whose class labels are acquired from an expert [2]. In classification on data streams, optimal learning with a minimal training set can be pursued through active learning. Hence, it is important to select data samples that can represent the underlying data distribution. The selected data samples should also be able to reveal a new trend when concept drift occurs.

As active learning methods in a batch mode select the most difficult data sample for the current classifier[3], [4], active learning on a data stream also utilizes uncertainty in the prediction for data samples. The method in [5] selected samples that were closest to the separating hyperplane by using a support vector machine (SVM) and in [6] data samples were chosen with a probability that is inversely proportional to the margin by a current linear hypothesis. In [7], [8], the error on a decision tree was estimated and a change was indicated by a significant change in errors, which triggered selecting data samples randomly or by the uncertainty from the recent data chunk. The method in [9] labeled the data samples that caused the classifier ensemble to have the largest ensemble variance.

When most active learning methods update a current classifier with a newly selected data sample incrementally, the updating formula is the same regardless of whether or not the selected sample indicates concept drift. In other words, the degree of concept drift indicated by the newly selected data sample is not used in updating a classifier. Therefore it may be slow to adapt the classifier to a new concept, especially at a low labeling budget. In [10], a concept drift detection method was run over labeled data samples, and a new classifier was constructed when drift was detected. However, it is difficult to detect concept drift accurately with only small number of labeled data samples.

In this paper, we propose an active learning method for an adaptive incremental classifier for data streams with concept drift. The proposed method measures the degree of concept drift indicated by an incoming unlabeled data sample and computes the selective probability based on the degree of concept drift. If the sample is selected for active labeling, an adaptive incremental updating formula of a classifier is used. Further, we present a simple approach to choosing the optimal learning rate in adaptive learning based on an ensemble method. The proposed method can effectively combine drift detection in unlabeled data samples, active leaning, and adaptive model update.

The rest of the paper is organized as follows. Section II gives a brief review of several active learning algorithms on streaming data. Section III presents the proposed active learning method. Experimental results comparing the proposed method with other methods are given in Section IV and discussions follow in Section V.

II. Related Work

A fundamental and simple way to select a data subset for manual labeling is to randomly sample a small number of data samples from an unlabeled data stream. Random sampling is very easy and fast to implement, but the data samples selected by random sampling are not guaranteed to be optimal.

While each data sample has an equal probability for selection in a simple random sampling, usually most of the methods

978-1-4673-9005-7/16/$31.00 ©2016 IEEE

for active learning give more chance to be selected to a data sample for which the current classifier has the least confidence in the prediction. The method in [5] selects samples from the recent batch that are closest to the separating hyperplane by an SVM. The selected samples are added to the training set and the oldest samples of the same class are deleted from the training set. On the other hand, in [6] any sample has a positive probability to be selected, but the sample with a large margin to the hyperplane has a small probability for selection. When $\hat{p}$ is the margin for a data sample achieved by the current linear classifier, the label is obtained with probability $b/(b+\hat{p})$, where b is a scaling factor.

Updating a decision tree incrementally is one of the well-known methods for classification of streaming data [11], [12]. An active learning method [7] in a decision tree stores statistical information to monitor concept drift without knowing any true labels. The error or loss of the current model on a new batch is estimated using data distribution in leaf nodes. When a significant change is suspected, leaf nodes are modified by randomly selecting data samples from the new batch. While the method in [8] also estimates the error of the current decision tree on a new batch, it uses a Bayesian classifier to select samples by uncertainty when a significant increase of the error on a new batch indicates concept drift.

The method in [10] combines active learning and explicit concept drift detection and proposes three active learning strategies for streaming data. Under the given labeling budget, the decision for active selection of a data sample is made and the selected data sample is used to update the current classifier. Meanwhile, a method for concept drift detection is run on the acquired labeled data and drift detection causes the current classifier to be replaced by a new classifier. However, at a low labeling budget, it is difficult to reliably detect concept drift with only a small number of labeled data.

III. An active learning for Data Streams with Concept Drift

A change in a prediction pattern by the current classifier can be an important indicator of concept drift. Under this supposition, we monitor a prediction pattern by the current classifier without knowing the true class labels and select data samples based on the degree of concept drift. The proposed method combines active learning and adaptive incremental learning based on the degree of concept drift estimated from unlabeled data samples.

A. Monitoring a prediction pattern for unlabeled data samples

Let $f_i(x)$, $i = 1, \cdots, r$, with $\sum_{i=1}^{r} f_i(x) = 1$ denote the probability that a data sample x would belong to the class i in the prediction by a current classifier. We define a random variable in order to describe a prediction pattern of a classifier without knowing the true class labels. Let $\hat{y}(x) = [\hat{y}_1(x), \cdots, \hat{y}_r(x)] \in R^{1\times r}$ be a predicted class label vector where the component $\hat{y}_i(x)$ corresponding to the largest value $f_i(x)$ is 1 and the other components are 0 such that

$$\hat{y}_i(x) = \begin{cases} 1 & \text{if } i = argmax_{1\le j\le r} f_j(x), \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The random variable is defined as

$$X(x) = \|f(x) - \hat{y}(x)\|_2^2, \tag{2}$$

where $f(x) = [f_1(x), \cdots, f_r(x)]$. The value $X(x)$ can be interpreted as an impurity measure for the prediction made by the current classifier. As the uncertainty about the prediction increases, the value of X increases.

The upper and lower bounds of X can be computed as follows. The lower bound for X holds when $f_i(x) = 1$ for any class i and $f_j(x) = 0$ for all $j \neq i$, and therefore $X(x) = 0$. The upper bound for X is obtained when the class conditional probability for all classes is equal. Assuming that $f_1(x)$ is the largest value among $f_1(x), \cdots, f_r(x)$, the following equation derives the upper bound.

$$\begin{aligned} X(x) &= \sum_{i\neq 1} f_i(x)^2 + (f_1(x) - 1)^2 \\ &= \sum_{i\neq 1} f_i(x)^2 + \left(\sum_{i\neq 1} f_i(x)\right)^2 \\ &\le \sum_{i\neq 1} f_i(x)^2 + (r-1)\sum_{i\neq 1} f_i(x)^2 \qquad (*) \\ &= r\sum_{i\neq 1} f_i(x)^2. \end{aligned}$$

The first inequality marked with (*) holds by the Jensen's inequality which is

$$\phi\left(\frac{\sum a_i z_i}{\sum a_i}\right) \le \frac{\sum a_i \phi(z_i)}{\sum a_i} \tag{3}$$

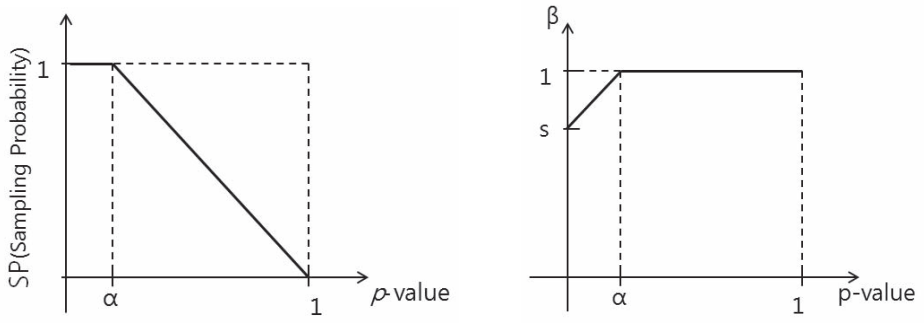
for any convex function ϕ and the equality holds if all z_i's are equal. The Eq. (3) is applied with the function $\phi(z) = z^2$ and the equality holds when $f_2(x) = \cdots = f_r(x)$. Hence, the upper bound for X

$$X(x) \le \frac{r-1}{r}$$

is obtained when $f_1(x) = f_2(x) = \cdots = f_r(x) = \frac{1}{r}$.

B. Sampling Probability for Active Learning

We compute the sampling probability based on the degree of concept drift indicated by an unlabeled data sample. Given a data stream $x_1, x_2, x_3, \cdots$, when x_{n+1} is an incoming data sample, we measure the degree of concept drift suggested by x_{n+1} and determine the sampling probability for x_{n+1} based on the degree of concept drift. The degree of concept drift can be estimated by how much $X(x_{n+1})$ deviates from the distribution of the values $X(x_{n-u+1}), \cdots, X(x_n)$ of recently received u data samples. However, it is difficult to estimate the distribution in a data streaming environment. In order to circumvent the problem, we simply assume normal distribution and estimate the mean μ and standard deviation σ. For



(a) Sampling Probability (b) weight value

Fig. 1. Two graphs for sampling probability (left graph) and the weight value of the current model in the classifier updating formula (right graph) which are mapped from the p-value by a new data sample. (α: significance level)

hypothesis testing on the occurrence of concept drift, the p-value for a new data sample x_{n+1} is computed by

$$p\text{-value} = P\left(Z > \frac{X(x_{n+1}) - \mu}{\sigma}\right) \tag{4}$$

for a standard normal distribution Z. If the p-value is smaller than the significance level, it indicates that the prediction by the current classifier is very uncertain.

Using the computed p-value, we determine the probability for the sample x_{n+1} to be selected for manual labeling. As the p-value decreases, the degree of concept drift indicated by x_{n+1} increases. Hence, two rules for the sampling probability need to be satisfied.

1) If the p-value for a data sample is smaller than the significance level α, concept drift is suspected. Therefore, it is better to select the sample for manual labeling and use it for a model update.
2) If the p-value is larger than the significance level α, the sample x_{n+1} is selected with the sampling probability $SP(x_{n+1})$ that is proportional to $1 - p\text{-value} = P\left(Z \leq \frac{X(x_{n+1})-\mu}{\sigma}\right)$.

The left graph in Figure 1 shows the function for sampling probability which was used in our experiments with the significance level α. The x-axis represents p-value, and the y-axis denotes sampling probability SP for x_{n+1}.

C. Adaptive Incremental Learning by the selected samples

The selected sample is used to update the classifier. Since a small p-value under the significance level indicates possible concept drift, we need adaptive incremental updating of the classifier where the effect of past data samples is reduced when there is a small p-value. Let β be the weight for the current model M_{cur} in an updating formula

$$M_{new} \leftarrow Update(\beta, M_{cur}, x_{n+1}).$$

The function which maps the p-value to the weight β needs to be an increasing function. In other words, when the p-value is smaller than the significance level, the weight β for the current model M_{cur} should be small. In our experiments, we used the function shown in the right graph of Figure 1 with the significance level α. When the p-value by $X(x_{n+1})$ is smaller than α, the β-value is set as a value between s and 1 by using the graph in Fig.1 (b). Otherwise, no concept drift occurs and β is set as 1 .

Any classifier with adaptive incremental learning formula can be used. We used adaptive incremental linear discriminant analysis modified from the method of [13]. The updating formula $M_{new} \leftarrow Update(\beta, M_{cur}, x_{n+1})$ is given as follows. Assuming that conditional density $p(x|i)$ has normal distribution, Bayes classifier gives a discriminant function

$$\begin{aligned} g_i(x) &= ln(P(i)p(x|i)) \\ &= -\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) - \frac{d}{2}ln(2\pi) - \frac{d}{2}ln|\Sigma_i| + lnP(i). \end{aligned} \tag{5}$$

$P(i)$ is the prior probability of class i, d is a data dimension, and μ_i and Σ_i are the mean and covariance of class i respectively [14]. When the covariance matrices for all of the classes are assumed to be identical to the within-class scatter matrix Σ, the discriminant function in Eq. (5) is simplified to

$$g_i(x) = \mu_i^T\Sigma^{-1}x - \frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i + lnP(i). \tag{6}$$

The updating formulas of the means, prior probability and covariance matrix for a new data sample x_{n+1} belonging to the class i are given as

$$\begin{aligned} \mu_i &\leftarrow (\beta n_i\mu_i + x_{n+1})/(\beta n_i + 1) \\ P(i) &\leftarrow (\beta n_i + 1)/(\beta n + 1) \\ P(j) &\leftarrow (\beta n_j)/(\beta n + 1), \quad j \neq i \\ \Sigma^{-1} &\leftarrow \frac{\beta n + 1}{\beta n}\left(\Sigma^{-1} - \frac{\Sigma^{-1}zz^T\Sigma^{-1}}{\beta n + z^T\Sigma^{-1}z}\right), \end{aligned} \tag{7}$$

where $z = x_{n+1} - \mu_i$ and n_j denotes the number of data samples in the class j. As data samples arrive continuously, the sample size n gets bigger and bigger and the weight multiplied by n does not make a role in reducing the effect of outdated data. In order to maintain effective sample size, we update the sample size n and n_i recursively such that

$$\begin{aligned} &n_i \leftarrow \beta n_i + 1 \text{ and } n_j \leftarrow \beta n_j, j \neq i \\ &n \leftarrow \beta n + 1, \end{aligned} \tag{8}$$

when a data sample x belongs to the class i.

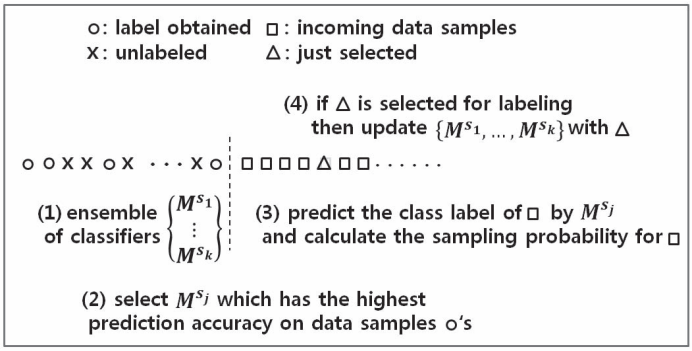


Fig. 2. The process to select a model with the optimal parameter value s by using classifier ensemble

D. Determining the value of a parameter s by Classifier Ensemble

The parameter s in the right graph of Figure 1 should be set differently according to the types or speed of the concept drift. With abrupt concept drift, the weight for the old model should be lower. On the other hand, no concept drift or a slow drift need a high s value. However, it is difficult to estimate exactly the type of concept drift in a data stream. In order to overcome this problem, we propose an ensemble approach. We keep an ensemble of classifier models by various s values and the classifier with the highest prediction accuracy on labeled data samples is used to predict the labels of samples in a data stream and to determine the selection of the data samples for active learning.

Figure 2 shows the proposed ensemble approach. The mark o denotes data samples labeled by selection and x represents data samples unlabeled because they were not selected. □ describes incoming unlabeled data samples. Let $\{M^{s_1}, \cdots, M^{s_k}\}$ be an ensemble of classifiers equipped with various values of the parameter s. On incoming streaming data, the classifier M^{s_j} which has the highest prediction accuracy on data samples o's is chosen from the ensemble and it is used to predict the class label of an unlabeled data sample □ in the data stream. In addition, the p-value and sampling probability for the unlabeled data sample □ are computed as explained in Sec. III-B. If the sample △ is chosen for active labeling, it is used to update the accuracy of the members of the ensemble.

The proposed active learning method is denoted as AAIL(Active and Adaptive Incremental Learning method) and is summarized in Table I. The random variable X defined by the classifier M^{s_i} in the ensemble is denoted as X_i. The labeling budget B means the upper limit about the proportion of data samples selected for active labeling. We used the formula in [10] to estimate the labeling cost. The labeling cost $\hat{b}_t$ at time t was approximated using $w = 100$ such that

$$\hat{b}_t = \frac{w-1}{w} * \hat{b}_{t-1} + \frac{1}{w} * \begin{cases} 1 & \text{if the label of } x_t \text{ is asked,} \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

$\hat{b}_t$ estimates how many true labels were queried within the last w data samples[10].

IV. Experimental Results

For the comparative evaluation of the proposed method, we used five artificial data sets and two real data sets. The proposed method was compared with the sampling strategies in [10].

1) Random : A true label is requested with a probability B and a classifier is updated incrementally.
2) Random-cdd[10] : A true label is requested with a probability B and a classifier is updated incrementally. If concept drift is detected with the Drift Detection Method (DDM) [15], a new classifier is built.
3) Var-Uncer.[10] : An instance with uncertainty below the threshold is selected while the threshold is adjusted according to the classifier stability. Incremental updating with the DDM is also used.
4) SPLIT[10] : Random sampling and Var-Uncertainty strategies are combined. Incremental updating with the DDM is also used.

We used incremental linear discriminant analysis as a base classifier for all compared sampling strategies, where the incremental updating formula $M_{new} \leftarrow Update(\beta, M_{cur}, x_{n+1})$ is given by setting $\beta = 1$. The test was performed under the budget limit B = 0.1, 0.3, 0.5. In the proposed method AAIL of Table I, the classifier ensemble $\{M^{s_i}\}$ was composed of five classifiers each of which was modeled using s =1, 0.9, 0.7, 0.5, 0.3 respectively. The initial classifier was constructed with the first chunk of data samples. (The size of the data chunk is explained later.) $\{X(x_i)\}$ was constructed from the next $u = 100$ data samples while updating incrementally the initial classifier. The significance level α was set as 0.05.

The evaluation method used in our experiments is the Interleaved Test-and-Train approach [16]. Each data sample is first used for testing, and then it is used to train the model. The accuracy was measured as the final percentage of data samples that were correctly classified over the interleaved evaluation.

A. Artificial Dataset

Five artificial data sets were generated using MOA data stream software [16]. MOA is an open source software framework in Java designed for online settings as data streams. Table II shows the details for the artificial data sets including

TABLE I
THE PROPOSED ACTIVE LEARNING METHOD AAIL

Input:
$\{M^{s_i}\}$: initial classification models equipped with various values of the parameter s
$\{\mu_i, \sigma_i\}$: the mean and standard deviation from $F_i = \{X_i(x_{n-u+1}), \cdots, X_i(x_n)\}$ constructed by M^{s_i}
B : the labeling budget
$x_{n+1}, x_{n+2}, x_{n+3}, \cdots$: an incoming data stream

Let $t = n + 1$.
While (an incoming data sample x_t is available)
  for each model M^{s_i}
    Compute the p-value$_i = P\left(Z > \frac{X_i(x_t)-\mu_i}{\sigma_i}\right)$.
    Remove the oldest element from F_i and add $X_i(x_t)$ to F_i.
  end for
  Select M^{s_j} which has the highest prediction accuracy.
  Predict the class label for x_t by using M^{s_j}.
  Compute the sampling probability $SP(x_t)$ based on p-value$_j$.
  if ($\hat{b_t} < B$ and (Bernoulli($SP(x_t)$)==success))
    Acquire the true class label for x_t.
    Update the prediction accuracy of each model M^{s_i}.
    Update each model M^{s_i} based on p-value$_i$.
  end if
  update the labeling cost $\hat{b_t}$
  $t = t + 1$.
end of while

TABLE II
THE DETAILED DESCRIPTION FOR ARTIFICIAL DATA SETS

Name	Data generator	classes	attributes
SEA	SEA generator	2	3
LED	LED generator with drift	10	24
HYPER	Hyperplane generator	2	10
STAGGER	Stagger generator	2	3
RBF	Random RBF generator with drift	10	10

TABLE III
THE ACCURACY(%) BY THE COMPARED METHODS ON ARTIFICIAL DATA SETS.

	B	Random	Random-cdd	VAR-Uncer.	SPLIT	AAIL
SEA	0.1	89.67	89.60	89.59	89.63	89.78
	0.3	89.65	89.63	89.25	89.72	89.84
	0.5	89.62	89.57	89.25	89.73	89.86
LED	0.1	74.00	73.95	73.93	73.97	74.00
	0.3	74.00	73.98	72.70	74.00	74.00
	0.5	74.00	73.99	73.89	73.99	74.01
HYPER	0.1	75.7	80.05	73.28	85.91	85.36
	0.3	76.83	77.72	68.06	87.85	88.01
	0.5	77.08	77.33	66.02	88.43	88.49
STAGGER	0.1	97.68	97.56	98.92	98.63	99.97
	0.3	99.22	98.28	99.26	99.47	99.98
	0.5	99.53	98.89	99.54	99.52	99.99
RBF	0.1	31.27	31.31	31.68	31.78	53.59
	0.3	31.22	31.36	32.06	31.98	73.32
	0.5	31.22	31.41	31.90	32.01	78.81

data generators. For all the parameters, default values were used except that in Random RBF generator and Hyperplane generator. In Random RBF generator the speed of change was set as 0.01, and in Hyperplane generator the magnitude of the change was 0.5. The initial classifier was constructed with the first 5000 data samples and the initial mean and standard deviation was computed from the next $u = 100$ data samples. For each data set, one million data samples were generated and the experiments were repeated 50 times.

The average accuracies over the repeated runs achieved by the tested algorithms are presented in Table III. The results by the proposed method AAIL are shown in the last column. For SEA and LED data, the compared methods gave very similar accuracies. In the proposed method, it was observed that for the prediction and selection of most of the data samples, the member with s_i= 0.9 or 1 was used. Furthermore, the accuracy improvement by large labeling budget B is not noticeable. From those observations, we conjecture that the concept drift was not abrupt. On the other hand, in the data sets by the RBF generator, the proposed method gave remarkably high accuracies compared with other methods. In STAGGER and HYPERPLANE data, the proposed method and SPLIT method

TABLE IV
THE ACCURACY(%) BY THE COMPARED METHODS ON TWO REAL DATA SETS.

	B	Random	Random-cdd	VAR-UNCER	SPLIT	AAIL
Elec.	0.1	64.42	64.04	63.00	65.05	72.50
	0.3	64.48	65.39	64.67	66.46	77.74
	0.5	64.50	67.47	70.07	66.47	79.59
Cov.	0.1	66.31	73.67	73.64	72.31	82.14
	0.3	67.29	77.78	77.25	74.65	85.63
	0.5	67.48	82.78	82.22	75.13	88.69

showed the competent performance.

B. Real Dataset

The two real data sets used in our experiments were Electricity [17] and Forest Covertype [18]. Electricity data describes changes in electricity prices based on the electricity market in the Australian state of New South Wales. It consists of 45,312 instances with 8 attributes and the class label "UP" and "DOWN" identify the change of the price related to a moving average of the last 24 hours. We used the data Elec2-3 which has five attributes: Day of week, Time, NSW electricity demand, Victorian electricity demand, and Scheduled interstate electricity transfer. After excluding data instances with missing values, 27,549 instances were left. Forest covertype data contains the forest cover type for 30×30 meter cells obtained from the US Forest Service (USFS) Region 2 Resource Information System (RIS) data. The goal is to predict the forest cover type from cartographic variables. It contains 581,012 instances and 54 attributes (10 numeric attributes + 44 nominal attributes) and we used 10 numeric attributes. The initial classifier was constructed with the first 100 data samples for Electricity and 5000 data samples for Covertype. The initial mean and standard deviation was computed from the next 100 data samples for both data sets.

Table IV shows prediction accuracies by the compared sampling strategies. As the labeling budget B increases, accuracies increase in almost all the methods except random sampling strategy. In particular, the proposed method AAIL shows very good performance in Electricity and Covertype data.

C. Sensitivity of the parameters α and u

In the proposed method AAIL of Table I, the parameters for the significance level α and a window size u of $\{X_i(x_{n-u+1}), \cdots, X_i(x_n)\}$ should be set. In our experiments, they were set as $\alpha = 0.05$ and $u = 100$. In order to observe the effects of the parameters α and u on the performance, the test was performed with the various values of the parameters. The graphs in Figure 3 (a) compare the performance by the proposed method on the values $u = 100, 500, 1000$ using LED, RBF, and real data sets. The graph shows that the values of a window size do not make a big difference on the performance. But, when the labeling budget B is low, the small window size gives better accuracy. It is conjectured that using a small window helps capture a changing trend. The graphs in Figure 3 (b) compare the performance for various values of $\alpha = 0.05, 0.03, 0.01$. While there is no difference in LED, too low significance level made the method passive in detecting drift in RBF and two real data sets.

V. DISCUSSIONS

We proposed a method for active learning on data streams with concept drift. In order to estimate the degree of concept drift indicated by unlabeled data samples, the prediction pattern is monitored using an impurity measure in class-conditioned prediction probabilities. A change in the prediction pattern is described by p-value which suggests a deviation from the prediction pattern in the old data samples. Based on the estimated p-value, the sampling probability for active labeling and a weight value in the classifier updating formula are computed. The proposed active learning method effectively combines the following aspects:

- estimation about the degree of the concept drift on unlabeled data samples,
- sampling probability which reflects the degree of the concept drift by an unlabeled data sample,
- classifier updating formula which reflects the degree of the concept drift by an unlabeled data sample.

The method can also be used together with adaptive incremental learning algorithms of various classifiers. Experimental results demonstrated a competent performance of the proposed method. In particular, the ensemble approach enables the proposed method to be applied successfully under the various types of concept drift.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2015R1D1A1A01056622).

REFERENCES

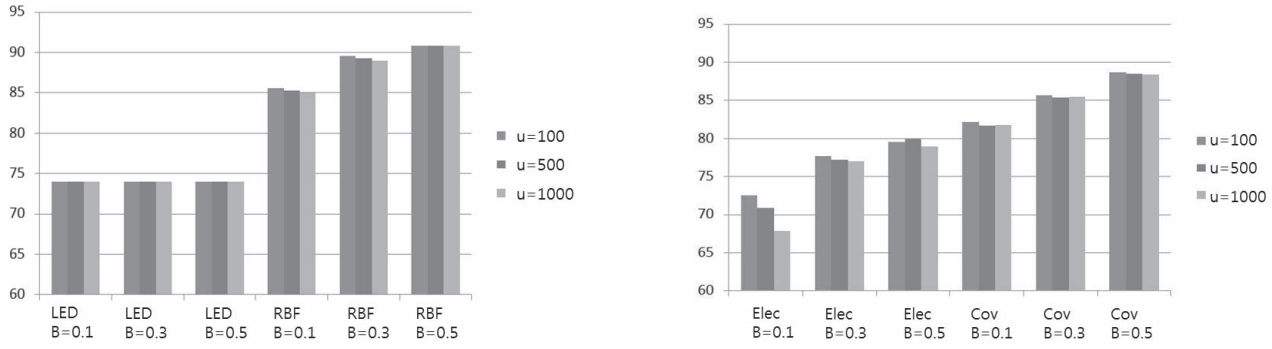
[1] J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, vol. 1(1), pp. 1:1–1:35, 2013.

[2] B. Settles, "Active learning literature survey," 2009, computer sciences technical report 1648, University of Wisconsin-Madison.

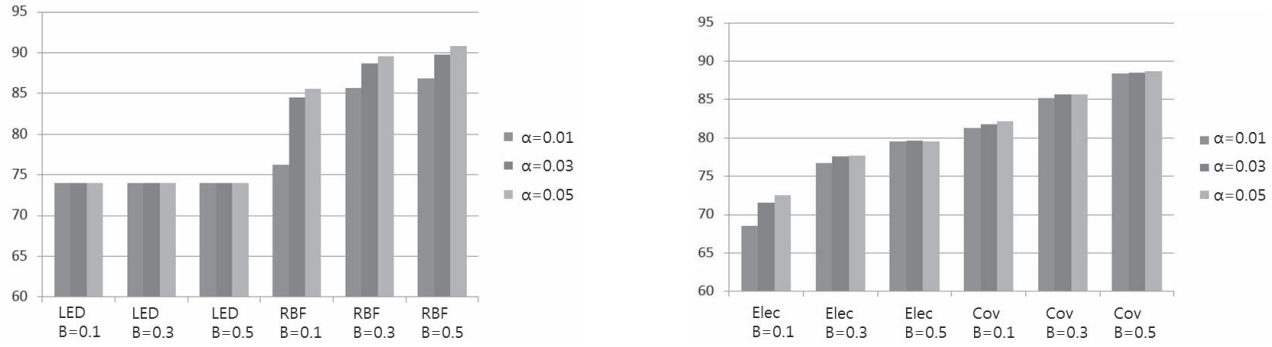
[3] Y. Freund, H. Seung, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm," *Machine learning*, vol. 28(2-3), pp. 133–168, 1997.

[4] S. Tong and D. Keller, "Support vector machine active learning with applications to text classifications," *Journal of MAchine learning research*, vol. 2, pp. 45–66, 2002.

[5] P. Lindstorm, S. Delany, and B. Namee, "Handling concept drift in text data stream constrained by high labelling cost," in *Proceedings of Florida artificial intellignece research society conference*, 2010.



(a) changing a window size u when $\alpha = 0.05$



(b) changing significance level α when $u = 100$

Fig. 3. Performance comparison with respect to a window size u and significance level α. The accuracy(%) is shown on the y-axis.

[6] N. Cesa-Bianchi, C. Gentile, and L. Zaniboni, "Worst-case analysis of selecive sampling for linear classification," *Journal of machine learning research*, vol. 7, pp. 1205–1230, 2006.

[7] W. Fan, Y. Huang, H. Wang, and P. Yu, "Active mining of data streams," in *Proceedings of Fourth SIAM international conference on data mining*, 2004.

[8] S. Huang and Y. Dong, "An active learning system for mining time-changing data streams," *Intelligent data analysis*, vol. 11, pp. 401–419, 2007.

[9] X. Zhu, P. Zhang, X. Lin, and Y. Shi, "Active learning from data streams," in *Proceedings of Seventh IEEE international conference on data mining*, 2007.

[10] I. Zliobatie, A. Bifet, B. Pfahringer, and G. Holmes, "Active learning with drifting streaming data," *IEEE transactions on neural networks and learning systems*, vol. 25(1), pp. 27–39, 2014.

[11] P. Domingos and G. Hulten, "Mining high-speed data streams," in *Proceedings of KDD*, 2000.

[12] L. Rutkowski, M. Jaworski, L. Pietruczuk, and P. Duda, "Decision trees for mining data streams based on the gaussian approxiamtion," *IEEE transactions on Knowledge and Data Engineering*, vol. 26, pp. 108–119, 2014.

[13] L. I. Kuncheva and C. O. Plumpton, "Adaptive learning rate for online linear discriminant classifiers," *LNCS*, vol. 5342, pp. 510–519, 2008.

[14] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. Wiley-Interscience, New York, 2001.

[15] J. Gama, P. Medas, G. Castillo, and P. Rodrigues, "Learning with drift detection," in *Proceedings of SBIA Brazilian Symposium on Artificial Intelligence*, 2004.

[16] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, "Moa: Massive online analysis," *Journal of Machine Learning Research*, vol. 11, pp. 1601–1604, 2010.

[17] M. Harries, "Splice-2 comparative evaluation: Electricity pricing," 1999, technical report UNSW-CSE-TR-9905 of The University of New South Wales.

[18] J. Blackard and D. Dean, "Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables," *Computers and Electronics in Agriculture*, vol. 24(3), pp. 131–151, 1999.