Novel and Recurring Class Detection using Ensemble of Classifiers: A Class-based Approach

Mohammad Raihanul Islam
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh 1000

ABSTRACT

Over the recent years, concept-evolution has received a lot of attention to the research community because of its importance in the context of mining data streams. Mining data stream has become a crucial task due to its wide range of applications such as network intrusion detection, credit card fraud identification, identifying trends in the social networks etc. Concept-evolution means introduction of novel class in the data stream. Many recent works address this phenomenon. In addition, a class may appear in the stream, disappears for a while and then reemerges. This scenario is known as recurring classes and also remained unaddressed in most of the cases. As a result, generally where a novel class detection system is present, any recurring class is falsely detected as novel class. This results in unnecessary waste of human and computational resources. In this paper, we have investigated the idea of a class-based ensemble of classification model addressing the issues of recurring and novel class in the presence of concept drift. Our approach has shown impressive performance compared to the state-of-art methods in the literature.

General Terms:

Unsupervised Learning, Data Mining

Keywords:

Novel Class, Recurring Class, Concept Evolution, Stream Classification

1. INTRODUCTION

The problem of data stream classification has been studied among the research community over the recent years. One of the major characteristics of data stream mining is that, the classification is a continuous process. As a result, the size of the training data can be considered infinite. Therefore, it is almost impossible to store all the examples to train the classifiers. Some methods regarding incremental learning are proposed in [5, 16] to address this problem. Moreover, it is a common scenario that, the underlying concept may changes overtime; a characteristics known as *concept-drift*. A number of studies addressing the issue of concept-drift are presented in the literature [2, 6, 11, 14].

However, another significant phenomenon of the data stream is *concept-evolution*, which is considered as the emergence of novel classes in the stream. For example, a new topic may appear in social network or a new type of intrusion may be identified in the network. If the number of classes in the classifiers is fixed and no novel class detection system is present, then the novel class is falsely identified as existing class. Concept Evolution has become a new research direction for the researchers recently because of its practical importance. For example, if a new types of attack occurs in the network, it imperative to identify it and take actions as soon as possible. Several approaches regarding this issue have been studied in the literature [7, 9].

A special case of concept-evolution is *recurring class* where a class reemerges after its long disappearances from the stream. For example, a popular topic may appear in a social network at a particular time of the year (i.e. festivals or elections). This result in a change of topics in the discussion on the social network over the time period and then when the event ends the topic disappears again. A recurring class creates several discrepancies if not properly handled. If it is not properly identified, then it is erroneously considered as a novel class or an existing one. As a result, a significant amount of human resources is wasted to detect its reappearance. Some studies regarding the problem of recurring class are present in [1, 8].

The classification model for data stream can be constructed by ensemble of classifiers. In an ensemble approach, multiple base classifiers learn the decision boundary on the learning patterns and their decisions on test example ares fused to reach the final verdict [12]. The ensemble approach is more popular among the research community because of their higher accuracy, efficiency and flexibility [7].

In this paper, we propose a new technique to generate ensemble of classifiers to detect novel and recurring class in the data stream. For each class $\mathbb{C}$ in the stream we construct an ensemble of sub-classifiers of size $\mathbb{L}$, where each ensemble of classifiers is composed of $\mathbb{K}$ components. Initially, all the sub-classifiers are trained from the initial data chunk. We have observed the phenomenon that, if the class boundary between two classes is very close, then it is possible to get a false prediction if the instances fall closely to boundary region. In our approach, we have employed several strategies to mitigate this problem. Moreover, we have also used boundary augmentation to address the issue of noise. In addition, we have also used the falsely predicted instances to update our model. Our proposed method has outperformed the state-of-the-art techniques in the literature.

The rest of the paper is organized as follows. In Section 2, we discuss the previous works regarding data stream classification in the literature. We present our approach in Section 3. We discuss the experimental results in Section 4. We conclude in Section 5.

2. PREVIOUS WORKS

Several studies are present in the literature on data stream classification [1, 3–5, 8, 13–17]. It has been observed that, existing approaches can be divided into two categories. First one is single model approach where one classification model is used and periodically updated for new data. On the other hand, batch-incremental method constructs each model using batch learning. When older model can no longer give satisfactory results, it is replaced by newer models [6, 14, 16]. The advantage of ensemble model is that, updating the classification model is much simpler in this case. However, these techniques generally do not include novel or recurring class detection.

In [10], authors have proposed an ensemble of classification model to identify novel class in the stream. Initially a decision boundary is built during training. Any instances outside the decision boundary are labeled as *outliers*. The outliers are examined to see that, whether there is enough *cohesion* among the outliers and *separation* from the training data. If enough cohesion and separation is found then the instances is declared as novel class. In [9], authors have proposed an adaptive slack space to mitigate false alarm rate. Moreover, to distinguish between concept-evolution and drift Gini Coefficient is used. However, these works does not handle recurring classes.

An approach to identify recurring class is presented in [8]. Here, in addition to primary ensemble model, an auxiliary ensemble of classifiers is present. The auxiliary ensemble model is responsible for storing all the classes even after they disappear from the data stream. When an instance is detected as outlier in the primary ensemble, but falls within the decision boundary of the auxiliary ensemble, the instances is identified as recurrent class. Any test data outside the decision boundary of both ensembles are analyzed for novel class.

The approaches described in [8, 10] are considered as *chunk-based* method. A *class-based* ensemble approach is presented in [1]. Here an ensemble model is constructed for each class C of the data stream. Each ensemble has K micro-classifiers. Initially, micro-classifiers are trained from the data chunk. When a latest labeled chunk of data arrives, a separate micro-classifier is trained for each class. Then the newly trained micro-classifier replaces the one with highest prediction error of the respective class. An instances falls outside the decision boundary of all the micro-classifiers of all the classes is considered as an outlier and saved in a buffer. The buffer is checked periodically to detect novel class. Authors of [1] have shown experimentally and empirically that, class-based approach is better than the chunk-based technique.

In this paper, we propose a more sophisticated approach to construct a class-based ensemble of classifiers. We have also present a better way to update and maintain the ensemble model. Moreover, we propose two types of outliers to update the classifiers and novel class detection and also take the wrongly predicted data into account to modify the classifiers. Experiments show the effectiveness of our methods compared to other techniques.

3. OUR APPROACH

Here, we discuss the fundamental concept of data stream classification. Then we describe our approach for stream classification subsequently.

3.1 Preliminaries

Each data in the stream arrives in the following format:

$$D_1 =< x_1, ...x_S >,$$

$$D_2 =< x_{S+1}, ...x_{2S} >,$$

$$.....$$

$$D_\Gamma =< x_{(\Gamma-1)S+1}, ...x_{\Gamma S} >$$

where x_i is the i^{th} instance in the stream and S is the size of the stream. D_i is the i^{th} data chunk and D_Γ is the latest data chunk. The problem is to predict the class of each data point. Let l_i and $\hat{l_i}$ be the actual and predicted label of instance x_i. If $l_i = \hat{l_i}$ then the prediction is correct otherwise it is incorrect. The goal is to minimize the prediction error.

Stream classification can be used in various applications such as labeling message in social network or identify intrusion in the network traffic. For example, in credit card fraud detection system, each transaction can be considered as an instance or data point and can be predicted either as *authentic* or *fraud* by any classification technique. If the transaction is predicted as *fraud*, then immediate action can be taken to withhold the transaction. Sometimes, the predicted decision can be wrong (authentic transaction predicted as fraud or vice versa). This can be verified from the cardholder later. The feedback can be considered as "labeling" the instance and used to refine the classification model.

Another similar example can be drawn from network traffic where each action in the network can be considered as normal or intrusion. If an action is considered as an intrusion then the connection from the server can be switched off until further verification. If the action is verified later, then the connection may be reopened. This verification is the "labeling" of the data by human expert.

The major task in the data stream classification is to keep the classification model up-to-date by modifying it periodically with the most recent concept. The overview of our proposed approach is shown in Figure 1. The major parts of the algorithm will be described step-by-step.

3.2 Ensemble Construction and Training

In this section, we present the approach for generating the ensemble model. We will refer our model as **R**ecurring and **N**ovel **C**lass **D**etector **E**nsemble (RNCDE).

Initially, the data chunk is partitioned into $\mathbb{C}$ disjoint groups ($\mathcal{G}^1$,$\mathcal{G}^2$,...,$\mathcal{G}^\mathbb{C}$) according to the class labels, where $\mathbb{C}$ is number of classes in the chunk. Therefore, each group contains the instances of one class only. Then an ensemble of size $\mathbb{L}$ is constructed for each class i using $\mathcal{G}^i$. Each ensemble $\mathbb{E}^i_l$, $i \in \mathbb{C}$, $l \in \mathbb{L}$ is composed of a sub-classifier $\mathbb{S}^i_l$. Each sub-classifier $\mathbb{S}^i_l$ is trained on the instance of class i ($\mathcal{G}^i$). We apply K-means clustering to generate $\mathbb{K}$ clusters on the instances of each class i. For each cluster $\mathbb{H}^i_{l_j}$ of ensemble l of class i, where $j \in \mathbb{K}$ we keep a summary of the cluster [9] i.e. μ, the centroid, r, the cluster radius (distance between centroid and the farthest data point of the cluster) and η, the number of points belonging to the cluster. This way we do not need every data point of the cluster. Therefore, each sub-classifier $\mathbb{S}^i_l$ is the union of all the clusters built from the instances of class i ($\mathbb{S}^i_l = \bigcup_{j=1}^{K} \mathbb{H}^i_{l_j}$). This process for generating sub-classifiers $\mathbb{S}^i_l$ is repeated $\mathbb{L}$ times to construct the ensemble model $\mathbb{E}^i$ for class i ($\mathbb{E}^i = \bigcup_{l=1}^{\mathbb{L}} \mathbb{S}^i_l$). Finally, the overall model is the union of all the

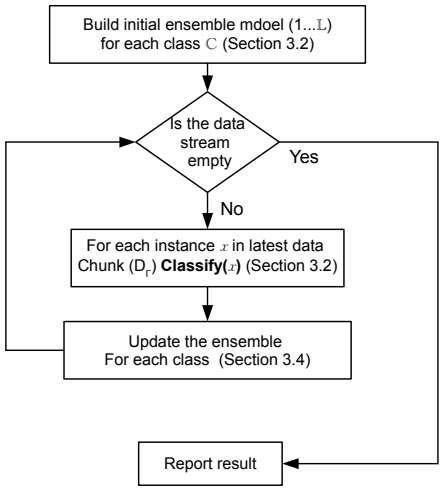


Fig. 1. Overall Aproach

ensemble built for each class i ($\mathbb{E} = \bigcup_{i=1}^{\mathbb{C}} \mathbb{E}^i$). For visual purpose, the partial structure of the ensemble model is shown in hierarchical form in Figure 2. It should be noted that, each ensemble for class i has only one sub-classifier, so the term $\mathbb{E}^i_l$, l^{th} ensemble model for class i and the l^{th} sub-classifier for class i, $\mathbb{S}^i_l$ can be used interchangeably.

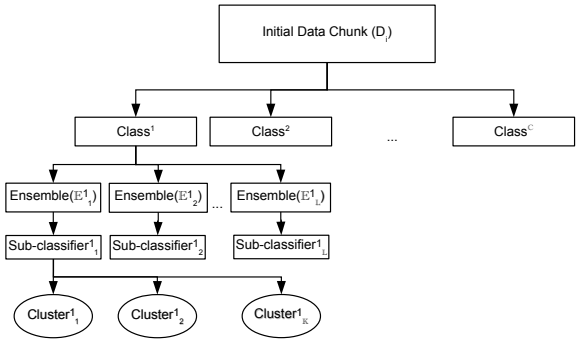


Fig. 2. Partial Structure of the Ensemble Model

Note that, each sub-classifier $\mathbb{S}^i_l$ of an ensemble $\mathbb{E}^i$ is trained on the same data $\mathcal{G}^i$. We vary the seed parameters ($\gamma_1, \gamma_2, ...\gamma_{\mathbb{L}}$) of K-means clustering to diversify the sub-classifier. We have shown our method using a hypothetical example in Figure 3. In Figure 3(a), the instances of the same class are shown. The K-means clustering is applied to construct sub-classifier 1 using seed parameter γ_1 (Figure 3(b)), where $\mathbb{K} = 3$. Then again sub-classifier 2 is constructed by K-means clustering initialized by the seed parameter γ_2 shown in Figure 3(c). We can see that, identical instances belong to different clusters at each sub-classifier. This process is repeated $\mathbb{L}$ times to construct $\mathbb{L}$ alternating sub-classifiers $\mathbb{S}^i_1 ....\mathbb{S}^i_{\mathbb{L}}$ for class i. In other words, each ensemble (sub-classifier) is a similar depiction of another where identical instances belong to different clusters. The superposition of both sub-classifiers is shown in Figure 3(d). Diversity is ensured by constructing the ensemble this way.
Each sub-classifier is the union of $\mathbb{K}$ clusters. The union of the clusters of sub-classifiers represents the decision boundary of of the corresponding class. The decision boundary of an ensemble model $\mathbb{E}^i$ for class i is the combination of all the sub-classifiers of the of class i. An instance is outside a cluster if the distance from the centroid of the cluster to that instance is greater than the radius of that cluster. An instance is outside a class if it is outside of all the clusters of all the sub-classifiers. The advantages of K-means clustering is that, its lower time complexity will allow to built classifiers in reduced time which is a critical requirements for data stream mining. Another benefit is that, after construction of the clusters, it is easy to modify them compared to other types of classifiers.

3.3 Classification

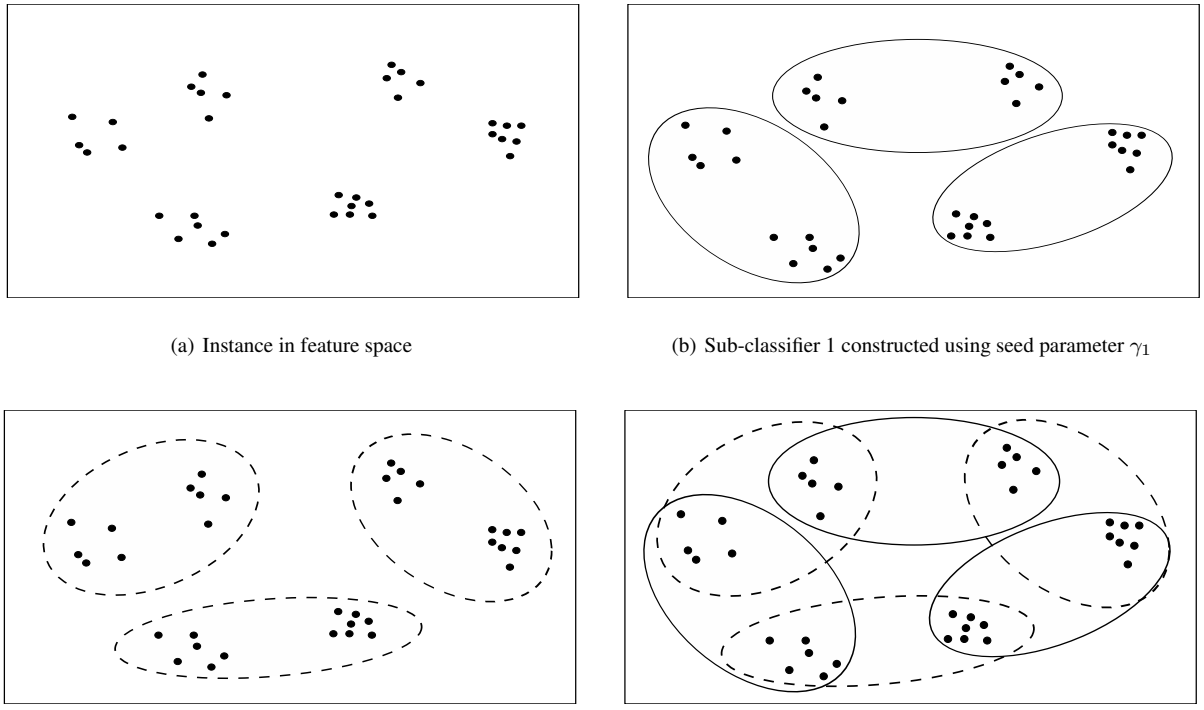
Here we describe our classification procedure and outlier detection. Each data point in the most recently arrived chunk is first checked for whether it is an outlier. In our approach, we have maintained two types of outlier i.e. class-outlier ($\mathbb{C}$-outlier) and universal-outlier ($\mathbb{U}$-outlier). If any instance is outside the decision-boundary of all the sub-classifiers of all the ensembles $\mathbb{E}^i$, then it considered as a $\mathbb{U}$-outlier. If a data point is a $\mathbb{U}$-outlier, then it is saved in *buffer* to analyze it further. If an instance x_i is not a $\mathbb{U}$-outlier then, it is inside the decision boundary of any class. It is possible that, x_i may be inside of more than one class due to noise and the curse of dimensionality. Let $\mathcal{E}_{x_i}$ be the set of such classes. We decide which class x_i belongs to by computing a coefficient (m-value). We called this coefficient *membership coefficient*. The m-value ($\tau^i_{l_j}$) for cluster $\mathbb{H}^i_{l_j}$, where $i \in \mathcal{E}_{x_i}$, $l \in \mathbb{L}$ and $j \in \mathbb{K}$ can be computed using the equation below,

$$\tau^i_{l_j} = \left(\frac{\eta^i_{l_j}}{\max\limits_{m \in \mathcal{E}_{x_i}, n \in \mathbb{L}, o \in \mathbb{K}} \eta^m_{n_o}} \right) / \left(\frac{d^i_{l_j}}{\max\limits_{m \in \mathcal{E}_{x_i}, n \in \mathbb{L}, o \in \mathbb{K}} d^m_{n_o}} \right)^{\beta}, \quad (1)$$

where $d^i_{l_j}$ is the Euclidean distance between the instance x_i and the centroid of cluster $\mathbb{H}^i_{l_j}$ where $\eta^i_{l_j}$ is the size of the cluster. Here β is the relative importance of the inverse of distance over the size of the classifier. We refer this constant as ξ-coefficient. The max size and max distance is used for normalization. After computing m-value for each cluster of all the sub-classifiers, the class label for instance x_i is computed using the equation below,

$$c = \arg \max_{i \in \mathcal{E}_{x_i}, l \in \mathbb{L}, j \in \mathbb{K}} \tau^i_{l_j} \quad (2)$$

The reason behind introducing the cluster size in the classification process is depicted in Figure 4. Here a hypothetical scenario is shown where two different clusters of different classes are present. Boundary of one of the clusters (cluster 1) is shown in continuous line (Class 1) and the other (cluster 2) is in dashed line (Class 2). We have also shown the data points of the clusters (i.e. dots and crosses). Now consider an instance shown by "O" in the figure. It is inside the boundary of both class. If we only consider only the Euclidean distance then it belongs to Class 2. However, from the figure it is evident that, it is more prone to the centroid of cluster 1 than cluster 2. Since size of cluster for Class 1 is larger, the decision boundary of cluster 1 is more expanded. Considering only the nearest neighbor to label the instance may result in erroneous prediction. However, if we make the assumption that, all the data points of a cluster are uniformly distributed, then the number of points in the overlapped region (common region between two clus-



(a) Instance in feature space

(b) Sub-classifier 1 constructed using seed parameter γ_1

(c) Sub-classifier 2 constructed using seed parameter γ_2

(d) Sub-classifier 1 and 2 are superimposed depicting the instances belong alternate clusters

Fig. 3. A hypothetical example of layer for 2-dimensional search space

ters) will be greater for cluster 1 than cluster 2. In this case, the test instance will be labeled as Class 1. Therefore, a more sophisticated measurement can be possible if we take account the size of the cluster in the classification process. That is why we propose Equation 1 and 2 for classification. The overall process is shown in Algorithm 1.

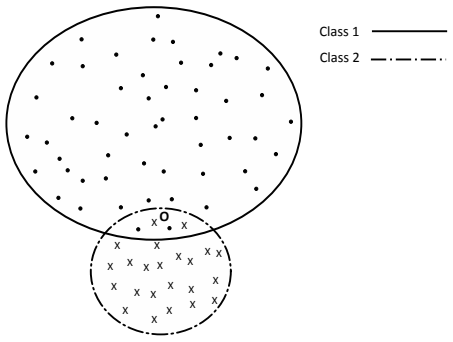


Fig. 4. A Hypothetical example of decision boundary of 2 clusters of two different classes in the Euclidean space

Algorithm 1 Classify

Input: x_i, the latest test instance
$buffer$ to store $\mathbb{U}$-outlier
Ensemble model for each class $\mathbb{C}$ $(\mathbb{E}^1, \mathbb{E}^2, ...\mathbb{E}^{\mathbb{C}})$
Output: $\hat{l}_i$, the predicted class label by the ensemble model

if x is a $\mathbb{U}$-outlier **then**
  $buffer \leftarrow x_i$
else
  $\mathcal{E}_{x_i} \leftarrow \{\mathbb{E}^i — x_i$ is inside the decision boundary of $\mathbb{E}^i\}$
  for all $\acute{\mathcal{E}} \in \mathcal{E}_{x_i}$ **do**
    for all sub-classifier $\mathbb{S}_n^{\acute{\mathcal{E}}} \in \acute{\mathcal{E}}$ **do**
      for all cluster $\mathbb{H}_{n_o}^{\acute{\mathcal{E}}} \in \mathbb{S}_n^{\acute{\mathcal{E}}}$ **do**
        compute $\tau_{n_o}^m$ by Equation 1
      end for
    end for
  end for
  Calculate the class label from Equation 2 for the instance x_i
end if
if size of the *buffer* exceeds the threshold **then**
  detect novel class
end if

3.4 Ensemble Update

We update the ensemble in two ways. One is single-update and the other is batch-update. First we discuss the single update.

When the labels for data points of a chunk are available (labeled by human expert), the incorrectly predicted data ($\mathcal{W}$) by the ensemble model is identified. Then the wrongly predicted data are separated according to their correct label. As a result, the all the inaccurately predicted data are partitioned into disjoint sets ($\mathcal{W}^1, \mathcal{W}^2, ...\mathcal{W}^{\mathbb{C}}$). Each instance of $\mathcal{W}^i$ is then analyzed. First the nearest cluster $\mathbb{H}^i_{m_n}$ is identified for each instance $x_n \in \mathcal{W}^i$ where i is the true label for x_n. Suppose, the distance between x_n and the centroid of $\mathbb{H}^i_{m_n}$ is $d^i_{m_n}$. Now if

$$\frac{d^i_{m_n} \cdot \left(\eta^i_{m_n} + 1\right)}{\eta^i_{m_n}} \geq r^i_{m_n} \tag{3}$$

then the radius and the size of the cluster $\mathbb{H}^i_m$ is updated using the following equations:

$$r^i_{m_n} = \frac{r^i_{m_n} \cdot \left(\eta^i_{m_n} + 1\right)}{\eta^i_{m_n}}, \tag{4}$$

$$\eta^i_{m_n} = \eta^i_{m_n} + 1 \tag{5}$$

Recall that, $r^i_{l_j}$ represents the radius of j^{th} cluster of l^{th} sub-classifier belong to class i. The reason behind this radius augmentation is, due to noise some data points may fall outside the decision boundary of the actual class. In addition, concept-drift may responsible for this kind of phenomena. To avoid the these errors on the future prediction this mechanism is implemented.
After single-update method the instances satisfied the Equation 3 are removed from each set $\mathcal{W}^i$. Then the remaining data in $\mathcal{W}^i$ are clustered using K-means clustering. The number of clusters $\mathcal{K}$ for K-means clustering is computed using the following equation:

$$\mathcal{K} = \frac{|D_\Gamma|}{ChunkSize} \cdot \mathbb{K} \tag{6}$$

Here $ChunkSize$ is a constant which can be initialized manually. These newly formed clusters can be called $\mathcal{C}_i$-outlier clusters where $i \in \mathbb{C}$. The union of $\mathcal{C}_i$-outlier is called $\mathbb{C}$-outlier. After the formation of $\mathcal{C}_i$-outlier clusters, the Euclidean distance from each $\mathcal{C}_i$-outlier clusters to each $\mathbb{H}^i_{l_j}$ is computed. Now based on the distance among the clusters we make two types of modifications. One is cluster merge and the other is cluster replacement.
If the distance between a $\mathcal{C}_i$-outlier clusters and one of the clusters ($\mathbb{H}^i_{l_j}$) in the ensemble is less than the radius of $\mathbb{H}^i_{l_j}$ ($r^i_{l_j}$), then the two clusters are merged. Recall that, the data points of $\mathcal{C}_i$-outlier are actually the wrongly predicted instances clustered according to the actual class label i. So it is normal that, any cluster from $\mathcal{C}_i$-outlier will tend to very remain very close to the $\mathbb{H}^i_{l_j}$ in the ensemble model. A possible scenario depicting the condition for merging the clusters is shown in Figure 5. Here the distance between $\mathbb{C}_i$-oulier cluster and the centroid of $\mathbb{H}^i_{l_j}$ is less than the radius of $\mathbb{H}^i_{l_j}$ ($r^i_{l_j}$).
Now to merge the cluster, we have to calculate the new centroid, the cluster size and the radius. To calculate the position of new centroid we have used the the equation below:

$$\mu^i_{l_j} = \frac{\eta^i_{l_j} \cdot \mu^i_{l_j} + \eta_{\mathcal{C}_i-\text{outlier}} \cdot \mu_{\mathcal{C}_i-\text{outlier}}}{\eta^i_{l_j} + \mu_{\mathcal{C}_i-\text{outlier}}}, \tag{7}$$

where $\eta_{\mathcal{C}-\text{outlier}}$ and $\mu_{\mathcal{C}-\text{outlier}}$ are the size and centroid of the $\mathcal{C}_i$-outlier. Since two clusters are merged, size is addition of the size of two clusters. The radius is computed by combining the radii of two clusters with the distance between the centroids. Here it should be noted that, some information is lost when the clusters are formed. This is performed to save space. As a result, radius calculation is not so accurate in this case. The position of the farthest data points for each cluster can be stored additionally for more accurate calculation of the radius of a cluster while merging.

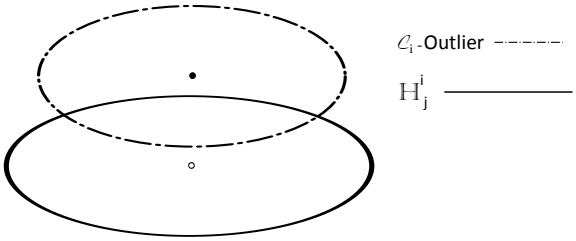


Fig. 5. A hypothetical scenario for cluster merging

After the merging of clusters the remaining $\mathcal{C}_i$-outlier clusters are replaced with the clusters from the sub-classifier. The replacement policy is as follows. We keep a count of error $\varepsilon^i_{l_j}$ for each cluster $\mathbb{H}^i_{l_j}$ for each ensemble model. Recall that, classification is computed by the m-value of the cluster. If prediction is wrong then count of error is increased by 1 for the cluster with max $\tau^i_{l_j}$, because it falsely identified the class as i. Now we replace the remaining un-merged clusters with clusters with highest $\varepsilon^i_{l_j}$ values accordingly. This way, the sub-classier can get rid of the obsolete clusters and the issue of concept-drift is resolved. Since we replace the older clusters with the cluster constructed with the most recent data points, the ensemble model remains up-to-date with the latest concept.

3.5 Novel Class Detection

We have extended and generalized the idea of novel class detection in [1]. The primary assumption behind the novel class detection in [1] was, data points of the same class should be closer to each other (*cohesion*) and farther apart from the other classes (*separation*). However, first assumption (i.e. cohesion) may prove different in some complex cases. It may be possible that, data points of the same class may be clustered together in various groups where these groups may be scattered through the feature space.
If the data points of a novel class emerge in the stream, we can assume that, the instances belonging to novel class will be far from the decision boundary of existing classes. Since data points of $\mathbb{U}$-outlier are outside the decision boundary of all the existing classes, these data are analyzed for novel classes. Recall that, the $\mathbb{U}$-outliers are stored in a buffer, if the size of the buffer reaches a threshold then they are analyzed for novel class. We have used the metric called q-NSC used in [1] for detecting novel class instance. We have modified the definition of and called it ***q*-mNSC**. In this method, another metric called *q,c-neighborhood* is used. We modify the definition of q,c-neighborhood also, which we called q,h-neighborhood. We define it as follows:

q,h-neighborhood: The q,h-neighborhood (q,h(x) in short) of an $\mathbb{U}$-outlier x is the set of q clusters that are nearest to x. (q-nearest cluster h neighbor of instance x).

Here q is a user defined parameter which can be initialized at the beginning. In summary, we compute the nearest q number of clusters from instance x regardless of the class the clusters belong to.

Now suppose, $\bar{D}_{h_{out},q}(x)$ be the mean distance of a $\mathbb{U}$-outlier instance x to its q nearest $\mathbb{U}$-outlier neighbors. Moreover, let $\bar{D}_{h,q}(x)$ be the mean distance from x to its q, $h(x)$ and $\bar{D}_{h_{min},q}(x)$ be the minimum value among all $\bar{D}_{h,q}(x)$. Here, h is the set of clusters from the existing classes. Then the q-mNSC of x can be computed according our definition:

$$q-\text{mNSC(x)} = \frac{\bar{D}_{h_{min},q}(x) - \bar{D}_{h_{out},q}(x)}{\max(\bar{D}_{h_{min},q}(x), \bar{D}_{h_{out},q}(x))} \quad (8)$$

The value of q-mNSC(x) ranges between -1 to +1. When the value is positive x is closer to $\mathbb{U}$-outlier instances and away from the existing classes resulting more cohesion and vice versa.
Now we explain how we can utilize the metric to detect novel class. First, we apply K-means clustering on $\mathbb{U}$-outliers to partition the data to $\mathcal{K}_0$ number of clusters, where $\mathcal{K}_0 = \mathbb{K} \cdot \frac{|\text{buffer}|}{ChunkSize}$. The reason for applying clustering is to reduce time complexity(reduces from $O(n^2)$ to $O(\mathcal{K}_0^2)$, where n is the total number of data points in $\mathbb{U}$-outlier). For each $\mathbb{U}$-cluster we compute q nearest cluster h_n for all the sub-classifiers of all the class. After that, for each $\mathbb{U}$-cluster we compute q-nearest neighbor cluster of that $\mathbb{U}$-cluster. Then we apply the Equation 8 to compute the q-mNSC for each $\mathbb{U}$-cluster. This way we get a q-mNSC value for each $\mathbb{U}$-cluster in the ensemble. From the value of q-mNSC, we can decide whether a novel class arrives in the stream.

4. EXPERIMENTAL FINDINGS

We present our experimental results in this section. First we discuss about the data set and then the parameter settings. Later, we describe the results and our remarks.

4.1 Data Sets

4.1.0.1 Synthetic Data. We apply the procedure described in [7] to generate synthetic datasets with concept evolution and drift. We generate three types of datasets as described in [7]. Each dataset contain 2.5×10^5 instances with 40 real value attribute. We refer each set as SynNCX having X classes (i.e. SynNC10 where total 10 classes are present).

4.1.0.2 Real Data. We have also used the data from UCI machine learning repository and KDD CUP 1999 intrusion detection challenge. We have taken the dataset *Forest* from UCI database and the *10 percent* version of KDD CUP. First dataset contains 581000 instances with 7 classes and 54 numeric attributes while the second datasets have 490000 instances having 23 classes and 34 numeric attributes. We randomly permutate the instances and construct 10 sequences and report the average results. We have made adjustments to have novel instances in the sequences.

4.2 Comparison with other methods

We have compared our approach (RNCDE) with class-based approach (CL) [1], ECSMiner (EM) [7], the clustered-based method presented in [13] (OW) and chunk-based approach (SC) described in [8]. Here CL, EM, OW have novel class detection mechanism while EM does not support recurring class detection.

4.3 Parameter Settings

We have set the size of the ensemble $\mathbb{L} = 3$, number of clusters per sub-classifier $\mathbb{K} = 20$. The minimum number of instances to detect novel class $q = 20$. Moreover, ξ is varied between 3 to 8 and size of the *buffer* is set to the 20% of the size of the chunk. These parameters are set either according to the parameters of the previous works or by running preliminary experiments.

4.4 Evaluation

We have used the following evaluation criteria for performance measurements. M_{new} = % of novel class instances misclassified as existing class, F_{new} = % of existing class instances misclassified as novel class, OTH = % of existing class instances misclassified as existing class and ERR = average misclassfication error (average of three types of error).
Initially, we construct the ensemble model from first three data chunks. Then we begin our performance evaluation from the chunk four. Table 1 summarizes the results from all the methods. We have taken the summary results on other methods from [1] and compared with our approach. OTH can be calculated from the other errors, so we do not show it. From the table, we can see that, OW has the highest error rate, because it can not detect majority of the novel class instances. Therefore, the F_{new} rate is also high in case of OW.
EM can identify novel class but it can not detect recurring class. As a result, recurring classes are detected as novel class and it has a high F_{new} rate also. SC maintains an auxiliary ensemble model which contains classifiers for all the class including recurring class. Therefore, it has comparatively lower F_{new} rate than EM. CL uses class- based ensemble to detect novel and recurring class and it has a lower error rate than the approaches above. Our proposed method RNCDE also have shown comparatively lower rate than other method. In *KDD* dataset, the ERR is slightly higher than CL, but in other case RNCDE shows better performance than other approaches.
In Figure 6, ERR rates for both Synthetic and Real Data are shown. In each case X axis represents number of data points and Y axis represents the ERR. For example from the Figure 6(a) and 6(b), we can see that, ERR rates after 300000 data points are 20% for *forest*, 10% in *KDD*. For synthetic data ERR remains almost constant. In case of *KDD*, we can see at the beginning ERR fluctuates, but the ERR decreases afterwards. This occurs because the at first the class boundary among classes are not accurately drawn so misclassification among existing classes (OTH) raises ERR. When the concept is learned comprehensively then ERR decreases. On the other hand, in forest ERR rises gradually. This is because M_{new} increases continuously when more data points arrive. (M_{new} is 4.4 in *forest*, see Table 1).

4.5 Parameter sensitiveness

We have also changed the value of ξ-coefficient. The effect of ξ on ERR for *forest* and *KDD* are shown in Figure 7. From the figure we can see that, both higher and lower value of ξ have a negative impact on ERR.
We have varied the number of clusters per sub-classifier $\mathbb{K}$. The $\mathbb{K}$ is varied between 10 to 50. The impact of varying $\mathbb{K}$ for synthetic dataset is shown in Figure 8. We can see from the figure that, ERR decreases, if the number of cluster $\mathbb{K}$ increases. The reason behind this is when the number of clusters increases more accurate decision boundary can be drawn among the classes. When the value of $\mathbb{K}$ is increased, more clusters will be formed on the same instances. Therefore, the size of the clusters will be comparatively lower and each cluster will learn the small portion of the total concept. If the boundary between two classes is noisy then more and smaller clusters will perform better than fewer and larger clusters. In other words, the boundary of the class will be more accurate constructed if an increased number clusters is formed. That is why

Table 1. Summary results on all the datasets

Performance Criteria	Methods	SynNC10	SynNC20	SynNC40	Forest	KDD
F_{new}	OW	0.9	1.0	1.4	0.0	0.0
	EM	24.0	23.0	20.9	5.8	16.4
	SC	14.6	13.6	11.8	3.1	12.6
	CL	0.01	0.05	0.13	2.3	5.0
	RNCDE	0.01	0.04	0.03	5.8	4.8
M_{new}	OW	3.3	5.0	7.1	89.5	100
	EM	0.0	0.0	0.0	34.5	63.0
	SC	0.0	0.0	0.0	30.1	61.4
	CL	0.0	0.0	0.0	17.5	59.5
	RNCDE	0.0	0.0	0.0	14.4	60.1
ERR	OW	7.5	7.7	8.0	30.3	37.6
	EM	8.2	7.9	7.2	13.7	28
	SC	5.1	4.8	4.3	11.5	26.7
	CL	0.01	0.02	0.05	7.3	26.0
	RNCDE	**0.019**	**0.02**	**0.02**	**10.57**	**24.76**

ERR deceases if $\mathbb{K}$ is increased. However, it should be noted that, if the value of $\mathbb{K}$ is high, then it would result in high space requirements and increased time complexity, which has a detrimental effect on the performance of the model. So the value of $\mathbb{K}$ should be adjusted to balance between the performance and accuracy.

5. CONCLUSION

In this paper, we have proposed a new ensemble model for detecting novel and recurring class in continuous data stream (RNCDE) which can be considered as a class-based approach as opposed to the chunk-based approach. Our algorithm have shown good performance against state-of-the-art methods in the literature. We have built our initial ensemble model for each class and updated and modified it periodically to learn the most recent concept. Each ensemble model has a sub-classifier which is composed of a number of clusters. The union of the cluster constitutes the concept of class. Our method has been proven very effective in data stream mining. Inspired by the promising results, we will concentrate on more efficient techniques for data stream classification. We are also planning to experiment our method on other real life data.

6. REFERENCES

[1] Tahseen Al-Khateeb, Mohammad M. Masud, Latifur Khan, Charu Aggarwal, Jiawei Han, and Bhavani Thuraisingham. Stream classification with recurring and novel class detection using class-based ensemble. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, ICDM '12, pages 31–40. IEEE Computer Society, 2012.

[2] Manuel Baena-García, Raul Fidalgo Josédel Campo-Ávila and, Albert Bifet, Ricard Gavaldà, and Rafael Morales-Bueno. Early drift detection method. In *ECML PKDD 2006 Workshop on Knowledge Discovery from Data Streams*, 2006.

[3] Jing Gao, Wei Fan, and Jiawei Han. On appropriate assumptions to mine data streams: Analysis and practice. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, pages 143–152, Washington, DC, USA, 2007. IEEE Computer Society.

[4] S. Hashemi, Ying Yang, Z. Mirzamomen, and M. Kangavari. Adapted one-versus-all decision trees for data stream classification. *IEEE Transactions on Knowledge and Data Engineering*, 21(5):624–637, 2009.

[5] Geoff Hulten, Laurie Spencer, and Pedro Domingos. Mining time-changing data streams. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, pages 97–106, New York, NY, USA, 2001. ACM.

[6] Jeremy Z. Kolter and Marcus A. Maloof. Using additive expert ensembles to cope with concept drift. In *Proceedings of the 22nd international conference on Machine learning*, ICML '05, pages 449–456, New York, NY, USA, 2005. ACM.

[7] M.M. Masud, Jing Gao, Latifur Khan, Jiawei Han, and Bhavani Thuraisingham. Classification and novel class detection in concept-drifting data streams under time constraints. *IEEE Transactions on Knowledge and Data Engineering*, 23(6):859–874, 2011.

[8] Mohammad M. Masud, Tahseen M. Al-Khateeb, Latifur Khan, Charu Aggarwal, Jing Gao, Jiawei Han, and Bhavani Thuraisingham. Detecting recurring and novel classes in concept-drifting data streams. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*, ICDM '11, pages 1176–1181, Washington, DC, USA, 2011. IEEE Computer Society.

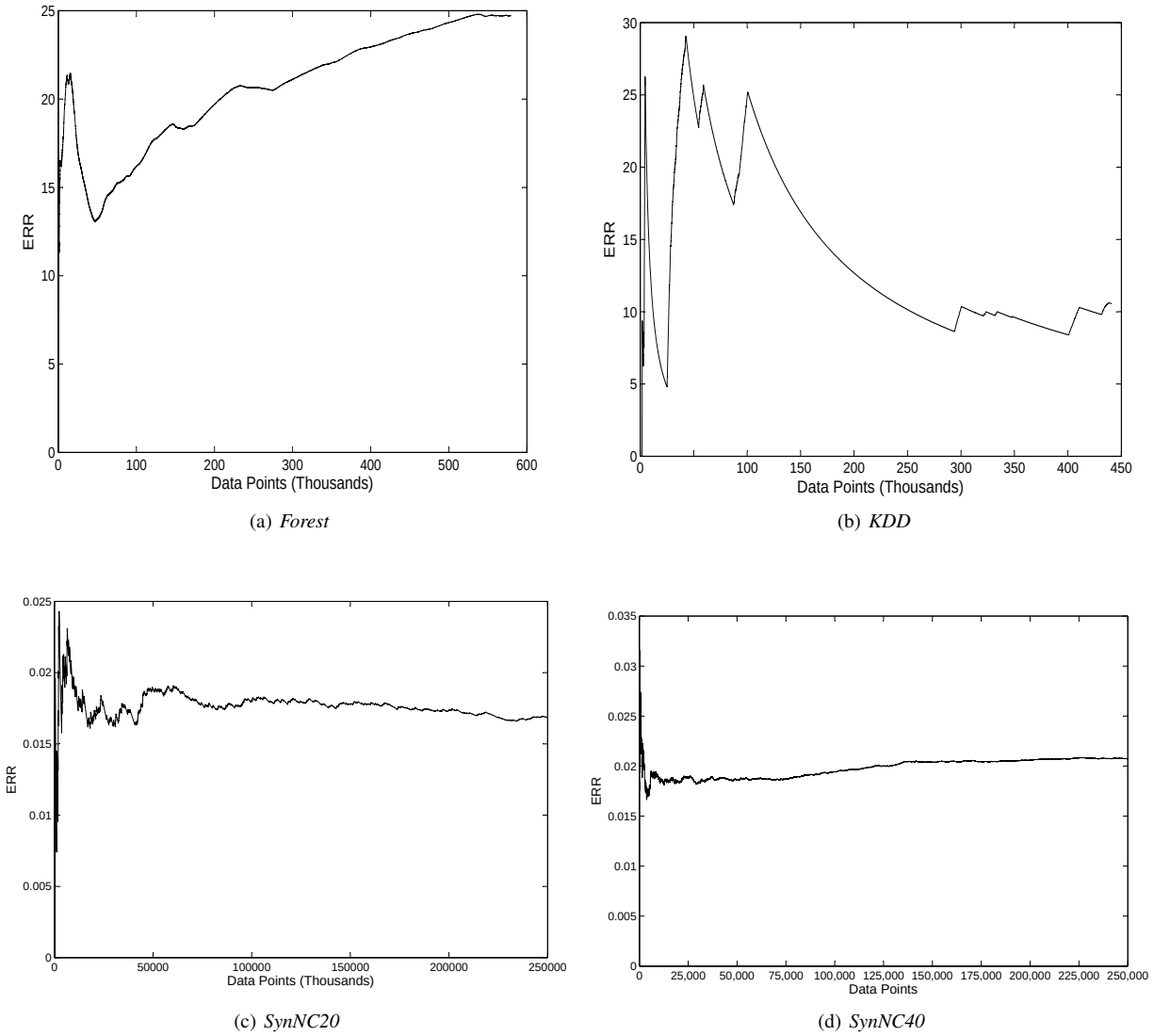
[9] Mohammad M. Masud, Qing Chen, Latifur Khan, Charu Aggarwal, Jing Gao, Jiawei Han, and Bhavani Thuraisingham. Addressing concept-evolution in concept-drifting data streams. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, ICDM '10, pages 929–934, Washington, DC, USA, 2010. IEEE Computer Society.

[10] Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, and Bhavani Thuraisingham. Integrating novel class detection with classification for concept-drifting data streams. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II*, ECML PKDD '09, pages 79–94, Berlin, Heidelberg, 2009. Springer-Verlag.

[11] L.L. Minku and Xin Yao. Ddd: A new ensemble approach for dealing with concept drift. *Knowledge and Data Engineering, IEEE Transactions on*, 24(4):619–633, 2012.

[12] Robi Polikar. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3):21–45, 2006.

[13] Eduardo J. Spinosa, André Ponce de Leon F., and Jo ao Gama. Cluster-based novel concept detection in data streams applied



(a) *Forest* (b) *KDD*

(c) *SynNC20* (d) *SynNC40*

Fig. 6. ERR for Datasets

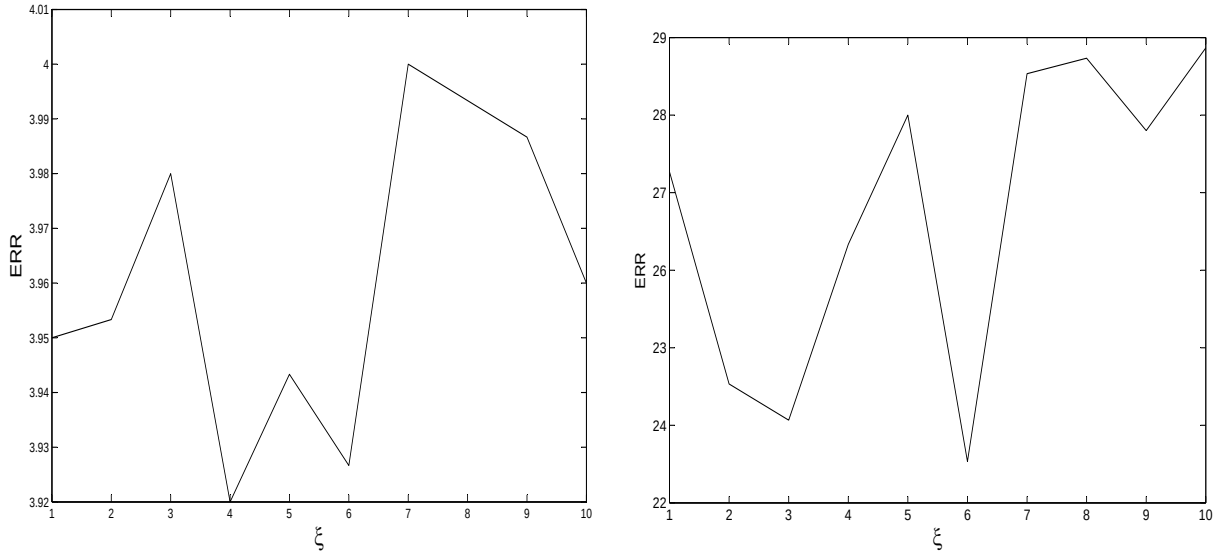
to intrusion detection in computer networks. In *Proceedings of the 2008 ACM symposium on Applied computing*, SAC '08, pages 976–980, New York, NY, USA, 2008. ACM.

[14] Haixun Wang, Wei Fan, Philip S. Yu, and Jiawei Han. Mining concept-drifting data streams using ensemble classifiers. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pages 226–235, New York, NY, USA, 2003. ACM.

[15] Peng Wang, Haixun Wang, Xiaochen Wu, Wei Wang, and Baile Shi. A low-granularity classifier for data streams with concept drifts and biased class distribution. *IEEE Transactions on Knowledge and Data Engineering*, 19(9):1202–1213, 2007.

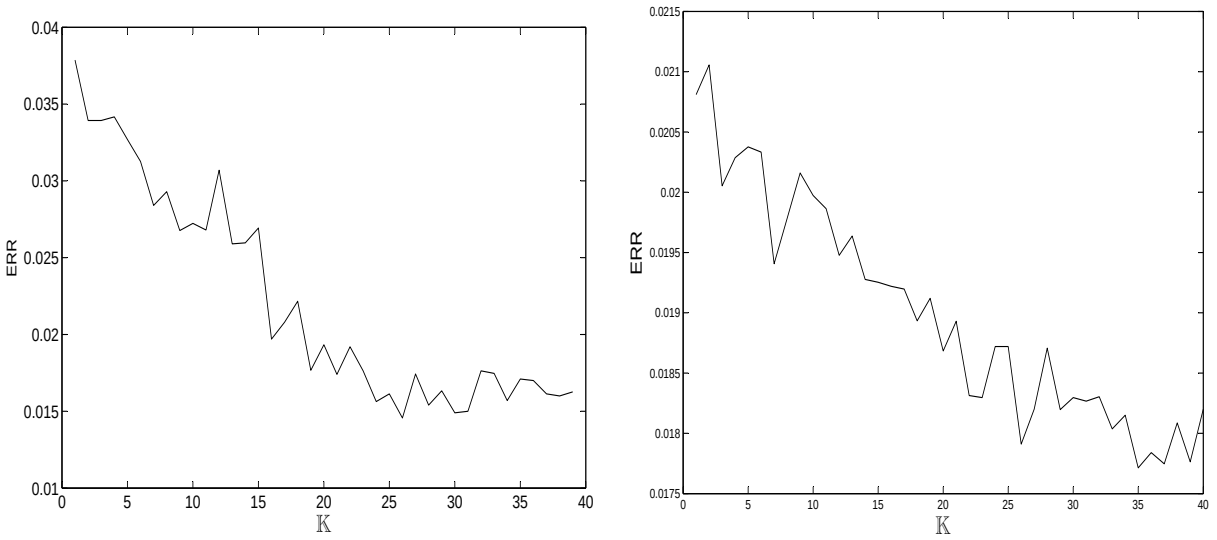
[16] Ying Yang, Xindong Wu, and Xingquan Zhu. Combining proactive and reactive predictions for data streams. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 710–715, New York, NY, USA, 2005. ACM.

[17] Peng Zhang, Xingquan Zhu, and Li Guo. Mining data streams with labeled and unlabeled training examples. In *Ninth IEEE International Conference on Data Mining*, ICDM '09, pages 627–636, 2009.



(a) *KDD* (b) *forest*

Fig. 7. ξ vs ERR



(a) *SynNC20* (b) *SynNC40*

Fig. 8. $\mathbb{K}$ vs ERR