A New Approach of Prediction/Classification Model Management

Jin-Hee Song*, Joung Woo Ryu**
*Shinhan University, Korea, **SafeTia Co.,Ltd., Korea
E-mail : jhsong@shinhan.ac.kr*, ryu0914@gmail.com**

1. Introduction

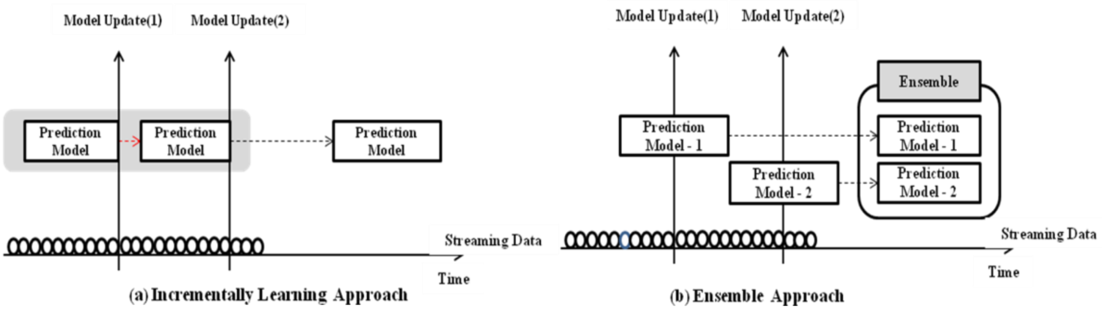
Many companies which deal with customer’s preferences or life styles provide reliable services using the prediction/classification model management method. Streaming data samples should be applied to a prediction/classification model in real time. The characteristics of streaming data is changed in data distribution according to data generation status. Usually prediction/classification models are periodically updated, and those methods has the following disadvantages[1]. First, if the accuracy of the model is high, the model need not to be update. However, even in this case, the prediction model is updated because of the updates period. Second, if it doesn’t complete the update period, the model would not be updated even though the distribution of the samples is changed within the update period. Third, human experts should labeled samples, and then evaluate or refine the current classifier using them. In a real-world application, it is very impractical process that a human expert give the classifier feedback on its decision for every single samples.

The accuracy of the prediction model will depend on the accuracy of the classification. Also, intervention of experts can reduce the efficiency of the prediction model. Thus, we propose a new active-learning modeling approach that can be used to accurately predict the safe state of human in the field of occupational safety than ensemble method in this paper. The proposed model is based on the safety status streaming data generated from wearable devices of a worker. The new modeling approach can be operate with minimal intervention modeling expert than conventional ensemble model and it also improve the accuracy of the prediction model.

2. Prediction/Classification Model Management

There is “Incrementally learning approach” or “Ensemble approach” as a general approach for updating the prediction models. In a stream classification problem, the current classifier assigns a label based on a predefined asset to each input sample to each input sample by a predefined asset. After classification, only when the correct labels of all new samples are available, the current classifier’s performance can be evaluated and the current classifier can be refined. A typical approach is to divide a data stream into subsets of streaming samples for periodically adjusting and improving a classifier using a fixed time interval. The subset of streaming samples is referred to a “chunk” or “batch”. In applications for detecting the abnormal patterns of the safety status of the site worker, the classifier decides the pattern is normal or not in real time. The classifiers are usually refined periodically using a “chunk” of consecutive sequence samples. Thus, human experts should analyze all of the samples of streaming data in the chunk in order to find the abnormal patterns.

2.1. Comparison Between the Conventional Model Management Methods



(a) Incrementally Learning Approach (b) Ensemble Approach

Figure 1. Update methods of the conventional prediction/classification model

For example, let us suppose that 10 samples are suggested to update model. The management methods of models as described in above set the maximum number of data that can be stored in the model update time interval.

- Incrementally learning approach : This method has updated parameters of the current prediction model that has with the latest 10 data as shown in Figure 1-(a).
- Ensemble approach : After generating a new prediction model-2, and is updated by adding the prediction model- 2 into the ensemble shown in Figure 1-(b).

2.2. Management of Ensemble Classifier

The conventional ensemble approach assumes that all streaming data have correct labels. The conventional ensemble approach assigns a new weight to each classifier of an ensemble or builds new classifiers for an ensemble using cross-validation whenever a new chunk is coming. Several researchers used the weighted samples when a new classifier for an ensemble are built from a chunk[2-3]. Online learning approaches can refine the current classifier whenever the correct labels of streaming data are obtained. Such an approach is able to quickly detect points in time where changes in streaming data distribution happen.

3. Active Learning Ensemble for Streaming Data

The proposed ensemble method is able to dynamically generate new classifiers for an ensemble on streaming unlabeled data without using "chunk". The new model is based on the safety status streaming data transmitted from the wearable device of the workers such as pulse rate, body temperature and movement etc. The approach of the proposed model is as follows.

[New Active Learning Ensemble Model]

Step 1. A problem area is divied into several sub-problem area, and then an ensemble is built by prediction/classification models generated from each sub-problem area.

- Classification Area : sub-problem area(It is defined by using the training data that was used to generate the model)

Step 2. When an input sample is given, the label of the sample is predicted by k-nearest prediction/classification models of the sample.

- The next a new model of ensemble is built using samples which do not belong to the current distribution of streaming data i.e. it generates an active learning data set.
- This study identifies and selects samples which do not belong to the current distribution of streaming data, and the selected samples are used for generating a training data set of a new model.

Step 3. When the accuracy of the ensemble is reduced, a new model would be added and one of existing models in the ensemble is deleted automatically .

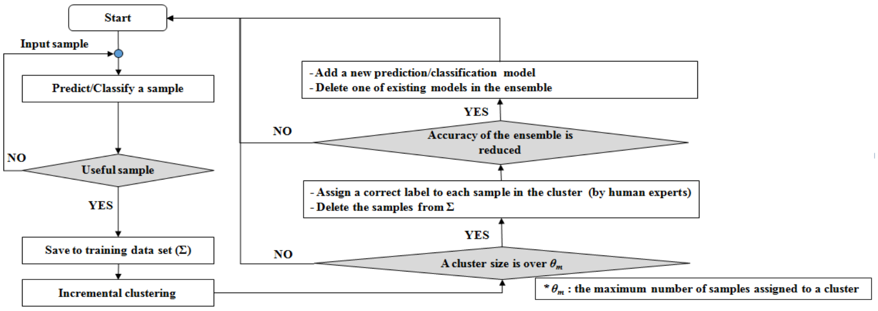

Start
Input sample
Predict/Classify a sample
NO
Useful sample
YES
Save to training data set (Σ)
Incremental clustering
- Add a new prediction/classification model
- Delete one of existing models in the ensemble
YES
NO
Accuracy of the ensemble is reduced
- Assign a correct label to each sample in the cluster (by human experts)
- Delete the samples from Σ
YES
NO
A cluster size is over θ_m
* θ_m : the maximum number of samples assigned to a cluster

Figure 2. New approach of the prediction/classification ensemble model

4. Conclusions

This paper presents a new approach in building an ensemble of prediction/classification model for streaming data. The method proposed in this paper is able to build new model for an ensemble when the new model is necessary, not systematically in time intervals for a fixed number of streaming samples. Furthermore, it is possible to build the active learning model of the ensemble using the "Training Data Set" for the safety status streaming data of the site workers, and minimize human experts intervention.

5. References

[1] Haixun Wang, Wei Fan, Philip S. Yu, "Mining concept-drifting data streams using ensemble classifiers", Proceeding of the 9th ACM SIGKDD international conference on knowledge discovery and data mining, pp.228-235, 2003.

[2] Dariusz Brezeziński, Jerzy Stefanowski, "Accuracy Updated Ensemble for Data Streams with Concept Drift", Hybrid Aritificial Intelligent Systems, Lecture Notes in Computer Science, Vol. 6679, pp155-163, 2011.

[3] Dhouha Mejri, Riadh Khanchel and Mohamed Limam, "An ensemble method for concept drift in nonstationary environment", Journal of Statistical Computation and Simulation, Vol.83, No.6, pp.1115-1128, 2013.