International Journal of Database Theory and Application
Vol.9, No.8 (2016), pp.13-20
http://dx.doi.org/10.14257/ijdta.2016.9.8.02

Novel Ensemble Tree for Fast Prediction on Data Streams

Shashi[1], Priyanka Paygude[1], Snehal Chaudhary[1], Debnath Bhattacharyya[2] and Hye-jin Kim[3]

[1, 2]Department of Information Technology, Bharati Vidyapeeth Deemed University College of Engineering, Vignan's Institute of Information Technology, [3]Business Administration Research Institute, Sungshin W. University
Pune-411043, Maharashtra, India, Visakhapatnam-530049, AP, India, 2, Bomun-ro 34da-gil, Seongbuk-gu, Seoul
19.shashisingh@gmail.com, {pspaygude,sdchaudhary}@bvucoep.edu.in, debnathb@gmail.com, hyejinaa@daum.net
(Corresponding Author)

Abstract

Data Stream is a continuous set of data records. When data arrive at a very high speed and continuously, so predicting the class in timely manner is important. Class prediction of data Stream is an important task in data mining. Nowadays Ensemble Modeling technique growing rapidly in Data Stream Classification. Ensemble learning become popular because of its advantage to handle large quantity of data stream, means it can handle the data in a bulk and also it can handle concept drifting. Earlier studies, mostly focused on accuracy of ensemble model, prediction efficiency has not considered much because existing ensemble model predicts in linear time, which is enough for general or small applications and existing models works on integrating small number of classifier. But real world application have large volume of data stream so we need more base classifier to identify different patterns and build a high grade ensemble model. To overcome these challenge we propose height balanced tree indexing structure (Ensemble tree) of base classifier for fast prediction on data streams by ensemble modeling technique. Ensemble Tree handles ensembles as spatial databases and it make use of an R-tree like structure to achieve sub linear time complexity.

Keywords*: data Stream mining, classification, machine learning, Spatial indexing, concept drifting*

1. Introduction

In today's information system data stream is presents everywhere, so it is a difficult job to handle such big data, also the storage and analysis of it. Visualization of such large volumes of data to classify it is also difficult. Utilization of conventional or traditional algorithms becomes more difficult as the data stream may often shows the concept drift [4]. So it becomes a more difficult or research area for data mining is classification of data stream. There is several possible assumptions for such applications. This kind of data stream classification is popularly used where in real-time intrusion detection is necessary, also to check the spam mails or websites. For security purpose, biosensor measurements (city) system classification and analysis is an essential rising application. To address the challenges of large volume data and concept drifting, many machine learning technique *i.e*. Ensemble concept is been proposed by researchers. This technique includes weighted classifier machine learning that is nothing but weight of node [1-2], incremental classifier ensembles, classifier and cluster ensembles [3], *etc*. Basic technique used by ensemble

ISSN: 2005-4270 IJDTA

models is to divide the data stream into a small block of data and from each block of data, it build the base classifier, and then combine these classifier into number of ways for prediction.

Earlier studies, up to date have been mostly focused on accuracy of ensemble model, prediction efficiency has not considered much because existing ensemble model predicts in linear time, which is enough for general or small applications and existing models works on integrating small number of classifier. But real world application have large volume of data stream so we need more base classifier to identify different patterns and build a high grade ensemble model. Such applications need fast sub linear prediction solution. Here basic classification used in Ensemble model is constructed using decision tree without any loss of generalization. At that point, each base classification module contained some rules to make a decision. Every decision condition covers some area in the decision space that is known as spatial object. In this method every base class is get changed to a set of spatial objects, whereas ensemble model is changed into a spatial database. Using this method, the difficulty of classification can be deducted, that is used for some patterns sharing among all spatial entities used for base classification in the spatial database [5,8] used in machine learning. Numbers of different indexing structures for spatial database are available that use common patterns for all spatial data. This is used to decrease the number of queries and updating costs in database. There are many techniques which are used to index an ensemble models like M-tree, R+ -tree. There are many advantages of using ensemble model indexing as compared to conventional method of spatial indexing: (1) Ensemble model uses decision rules for indexing, these rules have much affluent data like label of classes , probability distribution and classifier weight. Conventional method of indexing planned for Spatial data like multimedia, roadmaps, rectangles and images *etc.* (2)Aim of ensemble model is to predict faster whereas aim of conventional method is fast updating and retrieval. (3) Because of concept drifting technique, stream of data change very frequently, so decision area applied by decision rule in ensemble models may belong to different categories class or class labels.

To overcome above mentioned challenges, we proposed a novel ensemble tree that integrates the base classifier into a height balanced tree structure for fast prediction that is to achieve the sub linear time complexity. E-tree (Ensemble tree) provides an efficient indexing structure. With the help of this approach we attain logarithmic complexity of time for predicting data streams.

Paper is assembled as follows: Section 2 consist of work related to survey. In Section 3 we focused on motivation behind this paper. Proposed system is discussed in next Section 4, it describes the architecture and basic operation of tree. Section 5 and 6 includes mathematical model and experimental results respectively and in Section 7 conclusion is presented.

2. Literature Survey

There are many algorithms and techniques which are proposed for handling large volume of data or data streams based on ensemble learning. In this paper [6] classification and novel Class detection for Concept-Drifting was proposed for data streams under time constraints, this issue has not been focused by majority of the existing classification technique. Existing techniques of classification systems expect that overall number of classes is fixed in the data stream. But like intrusion detection real world classification of data streams, a novel class like a new intrusion may become visible at any time. Conventional techniques of classification would not be able to find out the novel class up to the time models are instructed by labeled instances of missing novel class. Hence, instances owned by a novel class are misclassified by the current systems. They demonstrate to identify novel classes consequently not withstanding is required when the trained dataset with the novel class is not available for classification model. Novel class

identification turns out to be additionally major challenging part in the presence of concept-drift. In this approach they not only find out the intrusion but also tells that intrusion is of new kind. In concern to figure out if an instance of data stream adjacent to a novel class, the classification technique intermittently needs to look for additional test instances to find out the similarities between those instances. For classification of an instance a maximum permissible wait time Tc is applied as time constraints. Moreover, traditional data stream classification techniques presume that correct label of an instance data can be retrieved directly after classification of data instance. But in reality labeling process is time consuming so there is a T1 time delay in calculating the correct label. Their intention is to discover the novel class in absence of training model for that class. This is an important task as they need to make decision when to instantly do the classification and when to delay the classification and wait additional test instances to find out the similarities between those instances. The major difficulties in class detection are: 1.efficiently save training data without much memory utilization, 2. must have knowledge about when to detain the classification of stream and when to immediately classify the test instance 3. Delayed instances are classified within Tc unit of time and 4. Prediction of a novel class rapidly and accurately. They apply their technique on two distinct classifiers: first is decision tree classification and second is k-nearest neighbor classification. But this technique may lead to an inefficiency of classification model, in context of memory utilization and execution time. So as to build more effective model and improve efficiency, they used K means clustering with the training data.

In this paper [2] author proposed solution for large scale classification *i.e.* large scale or data stream classification problem is overcome, to do this they propose algorithm. They build a committee or ensemble classifier which is a combination of many classifiers, each built on a subset of the accessible data points. Lately, a lot of consideration in the machine learning group has been coordinated towards strategies like bagging and boosting. As these approaches works on resampling and reweighting technique, so they are not very useful for large scale applications. Their approach was based on the merging of classifiers as follows: Individual classifier is constructed from small data set, read data in blocks rather than entire data set at a time, then component classifiers are inserted into an ensemble with a fixed size. If ensemble is full, new classifier is inserted only if they satisfy the quality criteria so that the performance of ensemble is improved. Performance estimation are done by testing the existing ensemble and the new tree is built on the next set of data. They performed some experiment based on this framework. Their results included: (1) Better generalization will be achieve by increasing the size of the ensemble up to 20 to 25 classifier. But there is interchange between the number of classifiers and number of points per classifier with data sets of limited size, unless re sampling is performed. (2) Even if the accuracy of the trees was increased, accuracy of ensemble is decreased by pruning the individual trees. (3)Ensemble accuracy does not effect by simple little or more variations in majority. The Key performance of this technique is the strategy used to figure out which existing or outdated tree should be deleted and also whether or not a new tree should be added to the group. Their technique shows that classification diversity is also important, accuracy alone is not the best practice. Moreover to the issue of deciding the right concession between diversity and accuracy, in this technique they find out that diversity is quite hard to calculate. Ensemble is changing continuously and constantly, so collecting the calculation over time is complicated and also estimates are noisy. Another probability is that to support classifiers which do better on points that are misclassified but it leaves data noisy. So they rather support classifiers on which ensemble is not decided yet and they classify points correctly.

Ensemble classifiers have many advantages when we are dealing with the concept drift as compared to single Classification technique, like easily scalable and easy to parallelize, ensemble classifier get used to changes more quickly than single classifier and they are more accurate. In next paper [7] author discussed about ADWIN Bagging and Adaptive-

size Hoeffding Tree (ASHT). These are two new approaches *i.e*. ADWIN Bagging and ASHT proposed for concept drift study. Firstly, they proposed a new method of bagging using Hoeffding Trees of various sizes. It is an incremental approach for constructing tree. Decision tree induction algorithm is used in this method. This algorithm is capable for learning from large volume of data streams, with assumption that the distribution generating examples does not change very frequently. With some differences from Hoeffding Tree algorithm the Adaptive-Size Hoeffding Tree (ASHT) is derived. Listed are as: (1) ASHT has a maximal size or number of split nodes. (2) To reduce its size it deletes some nodes, approach used is as, once node splits, and number of split nodes is greater than the maximum value. The perception of using this method is: small size trees accommodate to changes more quickly whereas large size trees build on large data so they do better in one condition when there is small changes or no changes. Trees with size s are going to reorganize twice as compared to 2s size trees. There are two different options of deletion when the size of tree is greater than its maximum value. First is delete outdated node that is delete root and all children of the oldest node except the node where split operation has been done, then new root will be the root of the child which is not deleted. Second method is to start with a new root that is delete operation is performed on all nodes. Secondly, they proposed new method of bagging using ADWIN. ADWIN solves the problem of tracking the average stream of bits or real-valued numbers in a well specified way. It automatically detects and adapts with the current rate of changing which is both parameter-as well as assumption-free. It does not look after any program explicitly, but handles it by exponential histogram technique. There is only one online bagging method named as ADWIN Bagging which is used along with the ADWIN algorithm for detecting a change and also estimating the weights of the boosting method. When a change is detected, the most outdated classifier is deleted and a new classifier is added to the tree.

In 2010, Classifier and Cluster ensembles [3] was proposed for mining concept drifting data streams. This paper, overcomes the two main challenges of the existing system of ensemble classifiers which are as follows: (1) obtaining an authentic label of class for every unlabeled cluster? Here clustering models used allocate cluster ID to every cluster rather than using genuine class label, so in tree there is challenge in finding the correlation between the cluster IDs and its class labels. (2) in tree as node represent the class need to decide how to allocate weights to all base classifiers and clusters correctly, so that the concept drifting problem handle by ensemble predictor. To handle these challenges weighted tree classifiers and clusters model is used for concept drifting. To overcome the difficulties (1) first construct a graph which constitute of all classifiers and clusters. This graph is used to presents a new label mechanics to newly introduced class, it firstly promote this label data from all classifiers to the clusters, and then at each iteration of insert operation modifies the outcomes by reproducing resemblance between all clusters. (2) A consistency-based weighing technique is also proposed by author in which it uses assignment of a weighted value to each base class of decision tree depending on their frequencies with reference to the updated base model. After following this method we get a combined result of both classifiers and clusters for accurate prediction through a weighted averaging schema.

Traditional concepts are facing two main problems of knowledge discovery that is large volume of incoming data streams and other one is concept drifting. To overcome these challenges, in this paper they propose a weighted classifier [1] for mining the data streams with concept drifting. For timely prediction of data streams it is important that data streams should take new and updated patterns. There are some challenges like Accuracy, Efficiency and ease of use for maintaining an up to date and accurate classifier works on large volume of data streams (infinite) with concept drifts. (1) Accuracy: It is hard to decide whether the concepts are outdated or not and hence their effects should be cut off from the current model. The most common technique used to do it is to remove the

old concepts at some constant rate. Nevertheless, High rate will affect the accuracy of the current model *i.e*. less accurate, low rate will lead to less delicate model. (2) Efficiency: Another challenge is efficiency, Decision trees are quite unstable as they are developed in a divide and conquer approach. Even a slight use of the base of them may activate a huge modification in the tree, and may drastically result into a undermine efficiency. (3) Ease of Use: Along with the existing implementation methods, classification methods should also be adapted which consist of decision trees to handle data streams with drifting concepts in an increased manner. Reusability of the approach cannot be used directly as it is limited. In consideration of the above challenges, in this paper they propose an ensemble of weighted classifier for mining the input data streams with concept drift. As an alternative of regularly reexamine a single classifier, they train multiple classifier through ensemble approach from consecutive data streams. We should not expect to use the most qualified classifier, because there are chances of valuable information wastage by discarding outcome of classifiers which are less accurate and trained before. In order to avoid over fitting and the problems of conflicting concepts, the deletion of old data must be based on data distribution instead of based on their arrival time. Ensemble approach achieve this competence by assigning all classifier a weight, which is based on the expected accuracy of prediction on current data of the model. Other advantages of this approach is its efficiency and easy to use.

3. Motivation

Previous work mostly focused on accuracy of the ensemble model. Efficiency of prediction has not been considered much because typically prediction consumes a time that it makes linear time, which is enough for small or general application without limiting prediction efficiency. Real world application have large volume of data stream which may change continuously so we need more base classifier to identify different patterns and build a high grade ensemble model. Such application calls for sub linear prediction solution. The indexing objects are decision rules for ensemble models with much affluent data that includes class labels, class relative frequency of distribution, and weights of decision tree node *i.e*. classifiers. Ensemble model indexing mostly preferred for quick prediction or classification of incoming data stream. Data depiction Changes: Data stream changes consistently because of concept drifting, and hidden patterns. Therefore a decision region used for decision tree in machine learning method may be of distinct class labels. To overcome the challenges of existing systems, in this paper we propose a novel ensemble (machine learning) tree that integrates the base classifier into a height balanced tree structure for fast prediction that is to achieve the sub linear time complexity. E-tree (Ensemble tree) provides an efficient indexing structure.

4. Proposed System

Structure of E-Tree consist of mainly two parts: first is a tree structure like R-Tree [5] and second is a table where tree stores decision rules and table stores information about the classifier like ID and classifier weight. Both structures are coupled by linking every base classifier of the table to its related decision rules. E-Tree consist of three basic operations as follows: Search Operation: it is traversing operation that is traveling of tree to classify input data. Insertion Operation: it is used for integrating new base classifier into a tree. Deletion Operation: When E-tree is full, delete operation is called and deletes outdated (not used anymore) classifier.

There are two main modules of the system: Training module and Prediction Module shown in Figure 1. Training Module: For each input data stream (unlabeled stream is coming), in training module this stream data record is stored in buffer until buffer is full for labeling. This labeled data is used for creating or introducing a new class to label data stream which is inserted in an E-Tree using Insertion operation. When E-Tree is full then

classifiers which are outdated will be deleted using Deletion operation. Prediction Module: Prediction modules also maintain the similar copy of E-Tree generated in training module, so when unlabeled data stream comes, Prediction module call search operation for predicting the unlabeled data. The updated tree from Training module will be integrated or synchronized with the tree in prediction module every time when new classifier is added in E-Tree. Buffer: It is used to store data records until it is full. In this module stream will be labeled by experts. Classifier: This module is used for building a new base classifier from all the labeled data records. We use data streams from KDD data sets which are listed in Table 1.

Operations on E-Tree: Ensemble Tree (E-Tree) is similar to R-Tree and consist of three basic operations which are:

4.1. Search Operation

Whenever a new record comes, search operation is called to predict its class label. In this operation we first travel the tree to find the suitable decision rule which covers the record in the leaf node. We perform the depth first search on tree.

4.2. Insert Operation

Insertion operation on Ensemble based Tree are same as insertion operation on other trees that is, It is used for integrating new base class into E-Tree, this helps to ensemble model to adapt new trends or class of data streams. Insertion operation on Ensemble based Tree are same as insertion operation on other trees that is, It is used for integrating new base class into E-Tree, this helps to ensemble model to adapt new trends or class of data streams.

4.3. Delete Operation

This operation remove or delete the old classifiers which are not in use from long time when the E-Tree reaches its capacity. There are basically two different type of methods for deletion which can be preferred. First method is similar to the method used in B tree. Merging operation is performed to all the full nodes to any one siblings that results from the least area increase. The second one appears deletions in R-trees and performs in sequence of delete-then-insert. Firstly it goes to deletion of the under-full node, then only add using insert operation on the remaining entries into the tree. This method has more advantageous as: (1)Implementation is easy; and (2)Re-insertion operation will automatically or reproduce the spatial structure of the tree as new class found or some new features are added to existing class training dataset.

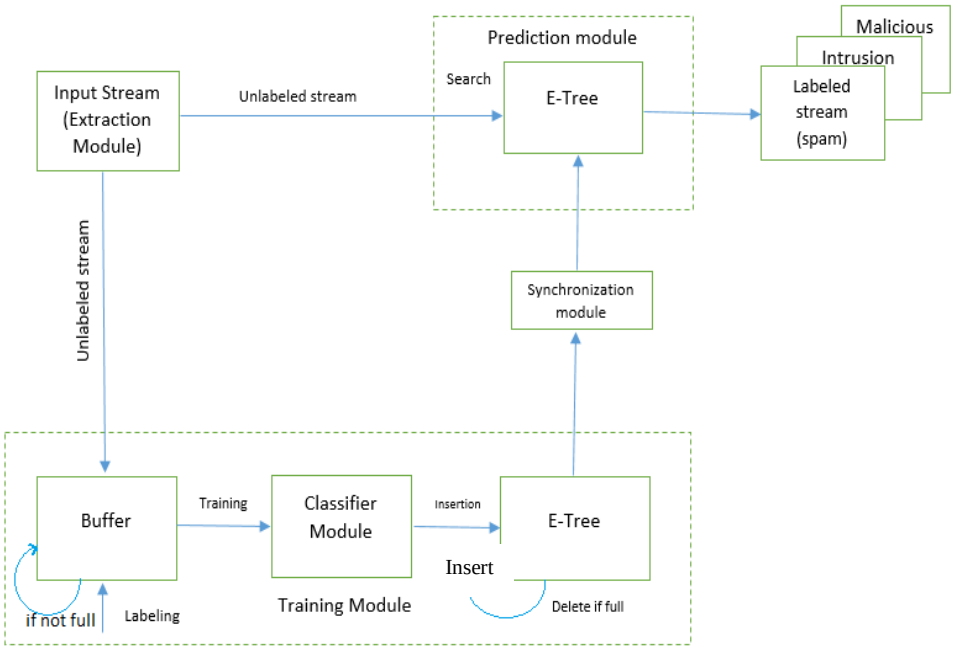

Figure 1. Architecture of Proposed System

Table 1. Data Streams

Name	Attributes	Areas
Intrusion detection	22	Security
Spam detection	35	Security
Malicious URL	12	Security

Table 2 shows the comparison results between J48 and Random Forest Algorithm which are used as a classification algorithm in previous paper and in this paper respectively. Comparison is done on basis of time and accuracy.

Table 2. Comparison

Data Streams	Instances	Classes	J48	Random Forest
			Old Accuracy	Expected Accuracy
Intrusion Detection	28119	Normal	99.91	99.98
		DOS	100	100
		Probe	100	100
		R2L	98	99.64
		UR2	99	100
Spam detection	4101	Spam	98.92	99.21
		Not spam	97.57	99.56
Malicious URL	11055	Malicious	71.16	73.08
		Benign	84.94	85.94

5. Conclusion

In this paper, we propose a novel Ensemble tree indexing structure for classifying high speed data streams to reduce the expected time complexity for prediction. We are able to reduce the linear time complexity of system to sub linear time complexity of system. Classification in this project is based on the ensemble models. In future we can extend novel Ensemble tree to other classification models of data stream.

References

[1] H. Whang, W. Fan and P. S. Yu, "Mining Concept-Drifting Data Streams using Ensemble Classifier", ACM SIGKDD International Conference on. Knowledge Discovery and Data Mining (KDD), Washington DC, **(2003)**, pp. 226-235.
[2] W. Street and Y. Kim, "A Streaming Ensemble Algorithm (SEA) for Large-Scale Classification", ACM SIGKDD international conference on Knowledge discovery and data mining, San Francisco, CA, **(2001)**, pp. 377-382.
[3] P. Zhang, X. Zhu, J. Tan and L. Guo, "Classifier and Cluster Ensembles for Mining Concept Drifting Data Streams", 2010 IEEE International Conference on Data Mining (ICDM), Sydney, NSW, **(2010)**, pp. 1175-1180.
[4] M. Kelly, D. Hand and N. Adams, "The impact of changing populations on classifier performance", ACM SIGKDD International Conference on. Knowledge Discovery and Data Mining (KDD), San Diego, CA, US, **(1999)**, pp. 367-371.
[5] A. Guttman, "R-Trees: A Dynamic Index Structure for Spatial Searching", ACM SIGMOD International Conference on Management of Data, New York, US, **(1984)**, pp. 47-57.
[6] M. Masud, J. Gao, L. Khan, J. Han and B. Thuraisingham, "Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints", IEEE Transactions on Knowledge and Data Engineering, vol. 23, no.6, **(2011)**, pp. 859-874.
[7] A. Bifet, G. Holmes, B. Pfahringer, R. Kirkby and R. Gavalda, "New Ensemble Methods for Evolving Data Streams", 15th ACM SIGKDD International Conference Knowledge Discovery and Data Mining (KDD), Paris, France, **(2009)**, pp. 139-148.
[8] R. Guting, "An Introduction to Spatial Database Systems", VLDB Journal, vol. 3, no. 4, **(1994)**, pp. 357-399.

