IJCTA, 9(12), 2016, pp. 5623-5630

The Decision Tree and Ensemble Learning Algorithm for Mining High Speed Data Stream

***N. Sivakumar **Dr. S. Anbu**

Abstract : The ensemble learning is a process to build a forecasting model by joining the collection of simpler base models. These models can be divided into two tasks. They are

1. Increase the set of base learners from the training data and then
2. Merge them as a decision tree to form a complex predictors. In this paper, the boosting algorithm helps to go to one step additional and constructs an ensemble model by conducting ordered and supervised search in a high dimensional space of weak learners. To increase the performance of decision tree ensemble, the DTEL algorithm (Decision Tree and Ensemble Learning) is proposed in this paper.

Keywords : Ensemble learning, Decision tree, Data mining, Machine learning, High speed data streams, Data streams, DTEL algorithm.

1. INTRODUCTION

Mining data stream is a succession of data arriving at a high speed and which is ordered by date and time. Mining high speed data streams has recently come out as a growing field of study in the research.

Data streams includes some research areas such as Machine learning, Database, Artificial intelligence, Statistics, Decision making, Scientific inventions etc.

For example the data stream real world applications such as finding the credit card fraud, detecting network violation, whether forecasting, mobile applications and application related to high volume of data. Using these applications the high speed data related problems can be modeled.

Ensemble Learning : It is refers that the several methods that are combined and learns a target function by training a number of individual learners and combining their forecasting.

Some applications of ensemble learning are :

1. In election the combined voters list helps to choose a good candidate.
2. The opinion of expert committee makes a good decision.
3. The Collection of individual partial knowledge could be comes as model.

The ensemble learning research in data streams is done using MATLAB software. We believe that this software is needed to help to improve the evaluation of the method.

We present in section 3 that the block diagram of the proposed new model. Also we present the ensemble learning concepts and its various methods in this section.

* Research Scholar, Department of Computer Science and Engineering, St. Peter's University, Avadi, Chennai, Tamilnadu, India. Email: sivakumarn002@gmail.com

** Professor, Department of Computer Science and Engineering, St. Peter's College of Engineering and Technology, Avadi, Chennai, Tamilnadu, India. Email:anbuss16@gmail.com

We present in section 4 that the decision tree which belongs to the most commonly used computational intelligence model. This can be easily expressed in the form of single logical rule. Furthermore we present algorithm and output.

2. RELATED WORK

A. Single classifier

The single classifier works for stable data but cannot process for temporary data. Some of these approaches are related to VFDT approach [1]; this is used in decision tree learning and Hoeffding bounds for confirmed output methods. These are assumed that the data process is stable time and space. But in this approach we could not store the examples therefore it was improved to CVFT [2] to concentrate on concept drift.

B. Concept Drift

In the research area it is referred as temporal evaluation and temporary data. The concept drift is assumed to be unpredictable, hard and it is usually not considered as a concept drift problem.

C. Ensemble classifier

It is accurate than any other classifiers. In this the classifier ensembles are quickly increasing the accuracy. Also this provides ensemble modifiers. This uses SEA approach, which is intended to read the blocks of data and builds an ensemble by progressively. Also it uses chunk, in this the small chunk causes an error. In order to overcome these type of errors we use algorithms to reduce the error rate.

3. PROPOSED NEW MODEL

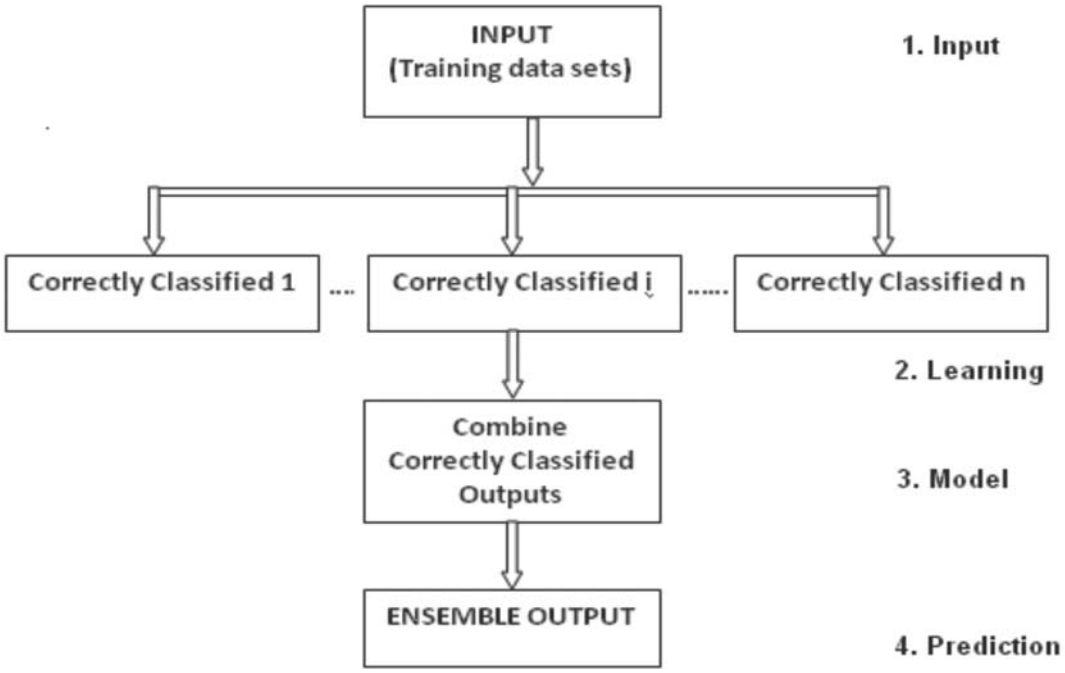

Fig. 1. Diagram for proposed new model

A. High speed Data stream

Data streams are to be represented as an input data that arrives at a very high rate. This makes stresses for transmit, compute and store the data. During the stream we need to calculate the different functions on the variable A at different time '*t*'. They are given as

Processing time per item in the stream is given as $A_i = 1,2, .. n$. The space is used to store the data on A_t time is *t*. and time needed to compute a functions on A is as follows

In figure 1, the most significant as follows

1. **Input :** Process one example at a time and inspect
2. **Learning :** Uses limited amount of memory and time
3. **Model :** Be ready to predict at any time
4. **Prediction :** The data analysis task takes numeric prediction. In this case a model will be constructed that predicts a ordered value.

These are constructs a new experimental framework for high speed data streams

B. Ensemble Machine Learning

It is refers that the several methods that are combined and learns a target function by training a number of individual learners and combining their forecasting. Consider the function of form

$$f(x) = \alpha_0 + \sum_{T_k \in T} \alpha_k T_k(x) \quad (1)$$

where T is the base function which involves thousands of trees. The above process can be divided into two stages. They are

1. A limited Tree $T_L = T_1(x), T_2(x), \ldots, T_m(x)$, of basis function is makes from the output data
2. A function $f\lambda(x)$ is used to build a path to fit the lasso rule. It is given as

$$\alpha(\lambda) = \arg\min \sum_{i=1}^{n} L[y_1, \alpha_0 + \sum_{m=1}^{M} \alpha_m T_m x(i)] + \lambda + \sum_{m=1}^{M} |\alpha_m| \quad (2)$$

Where T_L – is the limited trees which is produced by the boosting algorithm. This is used to reduce the set, and to save the storage for future forecasting

Need of Ensemble learning :

1. **Error reduction :** It is a process of reducing the similarity errors.
2. **Accuracy :** The accurate output of multiple experts can be produced
3. **Efficiency :** a complex problem can be divided into number of sub problems that are easier to understand and solve.

C. Ensemble Learning Methods

1. **Bagging :** It uses bootstrap sampling methods that are used to generate multiple input data by combining base classifier from the source learning data.
2. **Boosting :** This method is used to raise the execution of an imperfect learning algorithm which improves its behavior in reflection to the lowest error rate of a base classifier *(i.e* > 0.5 or = = 0).
3. **Random forest :** It is a technique that constructs a large number of decision trees during training time and output larger individual trees.

4. DECISION TREE BASED ENSEMBLE

It is a process of dividing the complex decision trees into several simpler decisions. In this the decision tree is in simple format which can be succinctly stored and sort out effectively as new data.

A decision tree can be built by recursively and dividing as data set into more complete. Also this reduces the complexity and provided easy understand and interpretable information.

Here we depict a modification in the tree ensemble method that concentrates on individual rules. Moreover to increase the large number of ensembles in the tree, we build set of rules from each tree. In the following section, we find the rules by searching the trees.

Figure 2 depicts a small tree, with numbered nodes. The following rules can be derived from this tree:

1. A1(x) = I($x1 < 2.1$)
2. A2(x) = I($x1 \geq 2.1$)
3. A3(x) = I($x1 \geq 2.1$). I($x3 \in \{S\}$)
4. A4(x) = I($x1 \geq 2.1$). I($x3 \in \{M,L\}$)
5. A5(x) = I($x1 \geq 2.1$). I($x3 \in \{S\}$). I($x7 < 4.5$)
6. A6(x) = I($x1 \geq 2.1$). I($x3 \in \{S\}$). I($x7 \geq 4.5$)

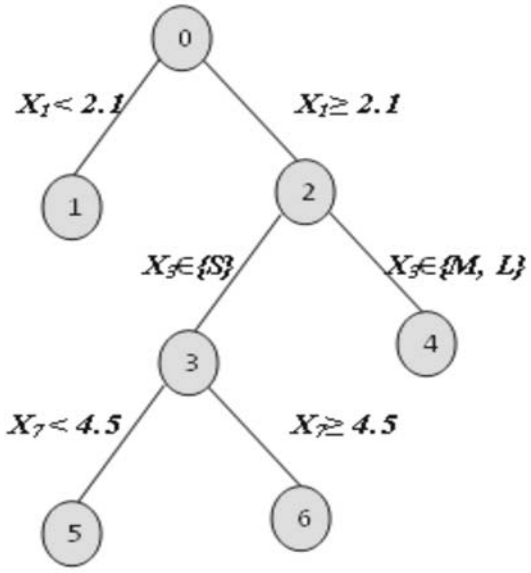

Fig. 2. Decision Tree Ensemble diagram

From the above, the rule of one dimensional expression A1, A4, A5 and A6 is equivalent to the tree itself; hence it is an over-complete basis for the tree.

For every tree T_m in an ensemble T, we can make its mini-ensemble of rules T^{M}_{Rule} and then aggregate them to form a larger ensemble, the formula is given below

$$T_{Rule} = \bigcup_{m=1}^{M} T^{m}_{Rule}$$

This is then treated like any other ensemble.

Algorithm

Step 1: DT(example, attributes, default)
Step 2: if example → null then return default
Step 3: else if attributes → null then return mode
Step 4 : else
Best ← choose-attribute
Tree ← decision tree with root
Step 5 : For each value vi of best do
Examples$_i$ ← { elements of best v_i}
Subtree ← DT {example, attributes-best
MODE(example)}
Step 6: Add tree + subtree
Step 7 : return

5. ERROR REDUCTION PROCESS

The fit ensembles are used to calculate and store the error in the ensemble objects which can be given as MSE below

$$\text{MSE} = \sum_{n=1}^{N} \overline{d}_n^{(t)} y_n - h_t(x_n))^n \quad (4)$$

Where

- $d_n^{(t)}$ are reflection angle at step t (the angle is added up to 1)
- $ht(x_n)$are forecasting of the model h_t which is burst to response value of y_n

This makes the intensity of individual learner for the weighted mean-squared errors.

6. DECISION TREE AND ENSEMBLE LEARNING (DTEL) ALGORITHM

The decision tree constructs the training data sets and the DTEL algorithm helps to construct the ensemble.

Algorithm :

Input

Load the data

Output

Step 1: Obtain the data

x is predictor data.

y is response data.

Step 2: Test the quality of ensemble as

$$y = 1 \text{ if } x_1 + x_{2+} + \ldots + x_n > 2.5 \text{ else } y = c$$

Step 3: Partition the data for independent test as

$$x = 30\% \; y = 30\% \text{ Test } x,y$$

Step 4: Create the ensemble for 200 trees as

$$\text{Bag} = \text{ensemble Fit}(x, y, \text{Bag}, 200, \text{'Tree'}, \text{'classification'})$$

Step 5: Error classification

Apply mean squared error formula

Step 6: Compact the ensemble as

tl: Remove the tree learners y: Remove the data

size: Reduce the memory

$$C = \text{loss}(tl, y(\text{test}), \text{size})$$

Step 7 : End

7. EXPERIMENTAL RESULTS

The algorithm is written in MATLAB– R2014B software with higher configuration system. The training data sets are used as bagged ensemble problems which can be generated an artificial dataset with 20 predictors. Also various transactions such as independent test, loss test, validation tests and error reports are generated and their outputs are stored.

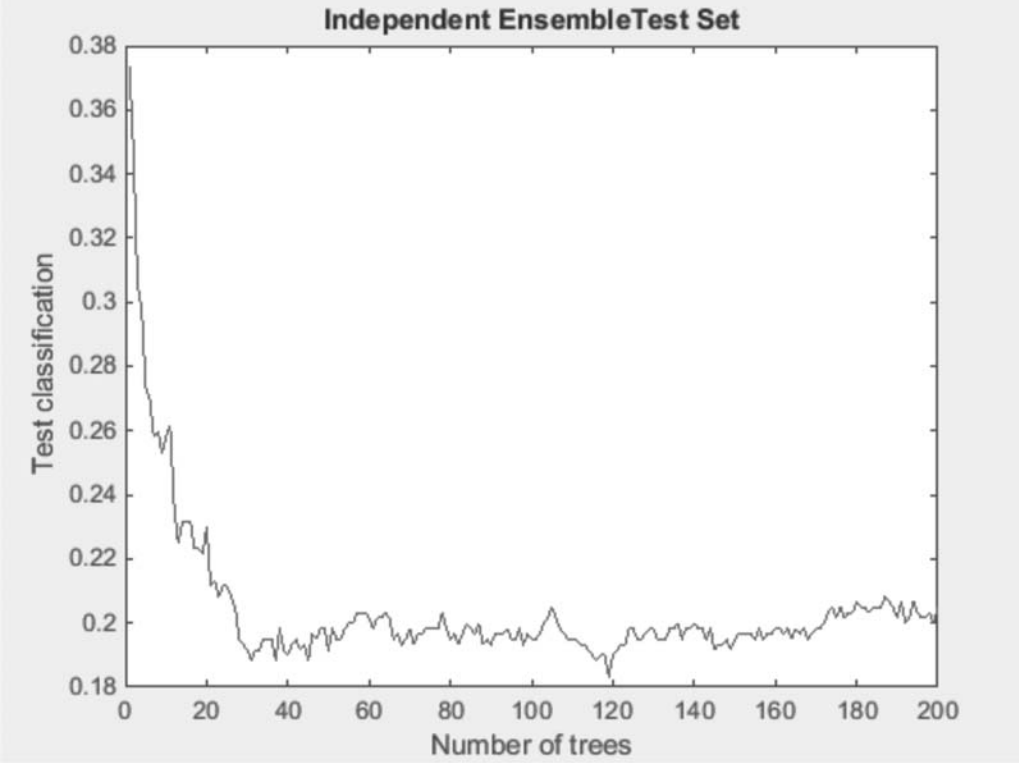

Fig. 3. From the above figure the different trained tree values are generated and their results are shown. Mainly it was generated from 200 trees of bagged ensemble classification.

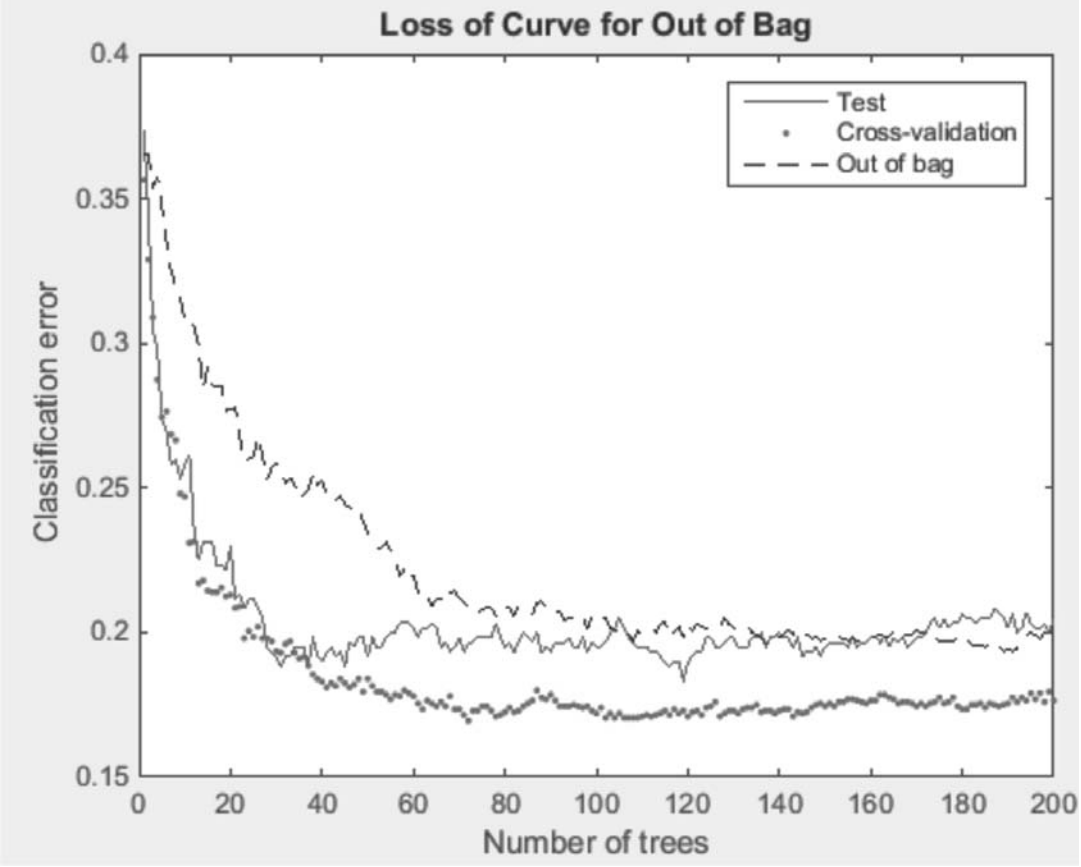

Fig. 4. From the above figure the analysis of the truth functions of the different number of trees in the ensemble are generated. This also produces the result of independent data

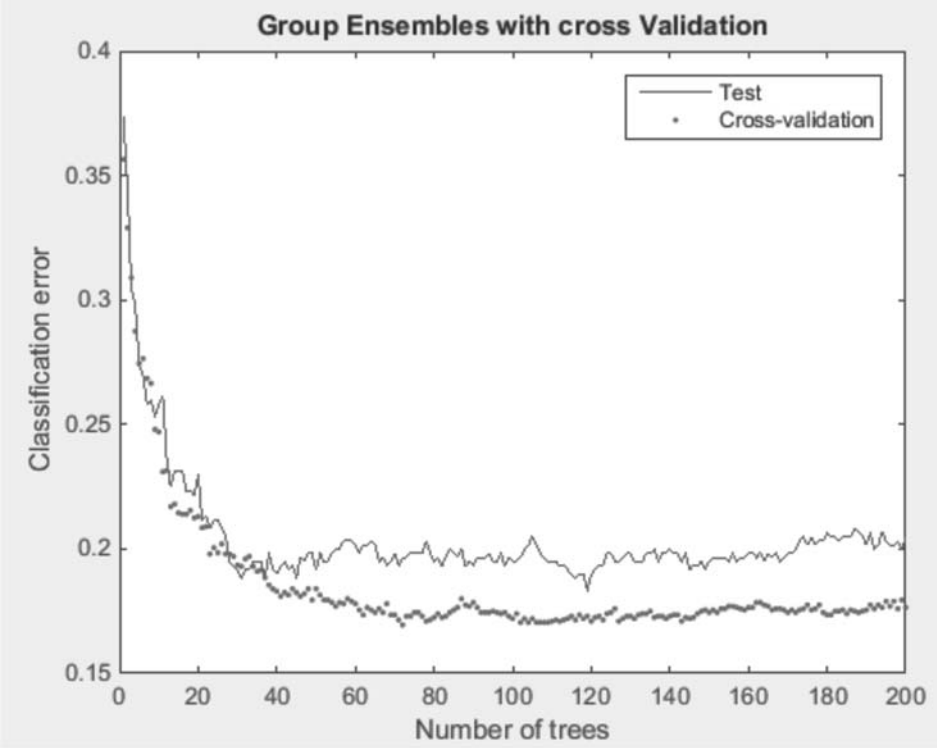

Fig. 5. From the above figure the loss curve is generated by an ensemble for out of bag approximation and plot them by other curves.

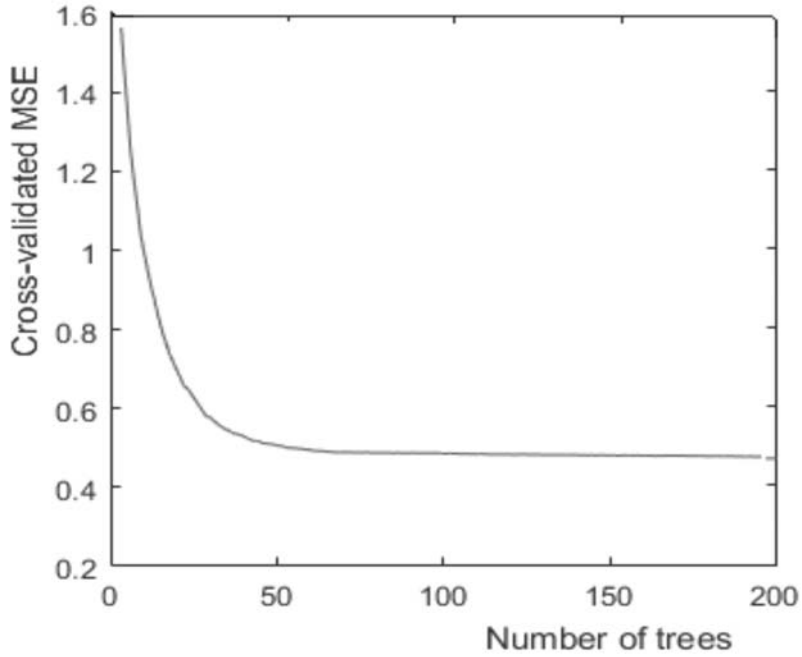

Fig. 6. From the above figure shown the smaller ensembles containing 50 to 200 trees have generated outputs. This produces the satisfactory performances. At the end we get a better accuracy and lesser learners.

7. CONCLUSION

To increase the effectiveness of ensemble learning, the decision tree algorithm and DTEL algorithms are suggested in this paper. Comparison is made between the suggested DTEL algorithm and Bagging algorithm in which the DTEL algorithm yields much better performance than the bagging algorithm.

8. REFERENCES

1. Pedro Domingos, Geoff Hulten, " Mining High Speed Data Streams",KDD-00 in proceeding of sixth ACM SIGKDD international conference on knowledge discovery and data mining, USA, 2000, pp 71-80.
2. W. Nick Street, Yong Seog Kim, "A streaming ensemble algorithm for large-scale classification", In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, 2001, pp 377 -382.
3. Rokach, L. 2010. "Ensemble-based classifiers". Artificial Intelligence Review 33 (1-2):
4. T.Dietterich. Ensemble methods in machine learning. In: First international workshop on multiple classifier systems, Lecture notes in computer science, 2000: 1-15.
5. Kuncheva, L. and Whitaker, C., Measures of diversity in classifier ensembles, Machine Learning, 51, pp. 181-207, 2003
6. Wolpert, D.H., and Macready, W.G., An Efficient Method to Estimate Bagging's Generalization Error, Machine Learning Journal, 35, 41-55, 1999.
7. Muhlbaier M., Topalis A., Polikar R., 2005, "Ensemble confidence estimates posterior probability," 6th Int. Workshop on Multiple Classifier Systems.
8. M. Muhlbaier, A. Topalis, R. Polikar, 2008, "Learn++.NC: Combining Ensemble of Classifiers with Dynamically Weighted Consult-and-Vote for Efficient Incremental Learning of New Classes," IEEE Transactions on Neural Networks.
9. R. Polikar, 2006, "Ensemble based systems in decision making," IEEE Circuits and Systems Magazine, vol. 6, no.3, pp. 21-45.
10. Roli, F., Kittler, J., Windridge, D., Oza, N., Polikar, R., Haindl, M., et al. (Eds.). Proceedings of the international workshop on multiple classifier systems 2000-2009 Lecture notes in computer science.