ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

An Efficient Model Ensemble Logical Bayesian Decision Trees for Ensemble Models On Data Streams

S.Mohanapriya[1], Dr.M.Rathamani[2]

Research Scholar, Department of Computer Science, NGM College, Pollachi, India[1]

Assistant Professor, PG Department of computer Application, NGM College, Pollachi, India[2]

ABSTRACT: Many real world applications, such as Web traffic stream monitoring, spam detection, and intrusion detection, and web click stream, generate continuously arriving data, known as data streams. To aid Ensemble learning (EL) based decision making, to correctly classify an incoming data stream based on the model learnt from past labeled data. Existing studies, to date, have been mainly focused on building particular ensemble models from stream data. However, a Ensemble-tree (E-tree) indexing structure focusing the prediction incurs spatial or temporal data analysis in response time problem for where data nodes have arbitrary extents, which is a legitimate research problem well motivated by increasing real-time applications. To address this problem, the proposed system takes a optimal Ensemble-tree spatial indexing structure based on logical Bayesian decision tree on to compute all base classifiers in an ensemble for express prediction. Experimental studies on both the synthetic and real-world data streams show the performance of our proposed approach.

KEYWORDS: C4.5, ripper, naive Bayesian, ensemble classifier.

I. INTRODUCTION

Data stream classification represents one of the most important tasks in data stream mining [1], [2], which has been popularly used in real-time intrusion detection, spam filtering, and malicious website monitoring. In the applications, data arrive continuously in a stream fashion, timely predictions in identifying malicious records are of essential importance. Compared to traditional classification, data stream classification is facing two extra challenges: large/increasing data volumes and drifting/evolving concepts [3], [4]. To address these challenges, many ensemble-based models have been proposed recently, including weighted classifier ensembles [5], [6], [7], incremental classifier ensembles [10], classifier and cluster ensembles [11], to name a few. While these models vary from one to another, they share striking similarity in their design: using divide-and-conquer techniques to handle large volumes of stream data with concept drifting. Specifically, these ensemble models partition continuous stream data into small data chunks, build one or multiple light-weight base classifier(s) from each chunk, and combine base classifiers in different ways for prediction. Such an ensemble learning design enjoys a number of advantages such as scaling well, adapting quickly to new concepts, low variance errors, and ease of parallelization. As a result, ensemble has become one of the most popular techniques in data stream classification.

In the past couple of decades, extensive research has been carried out on multidimensional indexing structures, to enable efficient range queries and nearest neighbour searches. However, most of the recent studies have focused on high-dimensional feature-based similarity searches into a relatively small number of point data items. Many scientific instruments, ranging from sensors on Earth orbiting satellites to light microscopes, can produce hundreds of gigabytes of spatio-temporal daily, consisting of billions of individual data elements. Storing each data element in a huge scientic dataset into a multidimensional indexing tree is impractical, because the size of the index could be even larger than the raw dataset, and the performance of queries would be poor due to size of the index.

The uncertainty of classification outcomes is of crucial importance for many safety critical applications such as spam detection and prediction of survival of network traffic monitoring. In such applications Bayesian model averaging

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

provide reliable estimates of the classification uncertainty. The use of Decision Tree (DT) classification models within a Bayesian averaging framework gives experts additional information by making the classification scheme observable [8, 9].

Technically, an E-tree based logical decision tree has three key operations: (1) Explore: traverse a decision tree to classify an incoming stream record x; (2) Insertion and Deletion: Join together new classifiers and pruning outdated classifiers; (3) Logical Expression: Constraints based logical decision. As a result, the logical Bayesian decision approach not only guarantees a time complexity for prediction, but is also able to adapt to new conditions and patterns in stream data.

II. RELATED WORK

In [2] authors address this problem of Ensemble-tree (E-tree) indexing structure to classify all base classifiers in an ensemble for quick prediction. On one hand, E-trees treat ensembles as spatial databases and employ an R-tree like height-balanced structure to reduce the expected prediction time from linear to sub-linear complexity. In [3] authors address this issue of a data stream classification technique that integrates a novel class detection mechanism into traditional classifiers, enabling automatic detection of novel classes before the true labels of the novel class instances arrive. Novel class detection problem becomes more challenging in the presence of concept-drift, when the underlying data distributions evolve in streams. In order to determine whether an instance belongs to a novel class, the classification model sometimes needs to wait for more test instances to discover similarities among those instances. In [4] authors illustrated about the general framework for assessing predictive stream learning algorithms. It defends the use of Predictive Sequential methods for error estimate - the prequential error. The prequential error allows us to monitor the evolution of the performance of models that evolve over time. Nevertheless, it is known to be a pessimistic estimator in comparison to holdout estimates. To obtain more reliable estimators we need some forgetting mechanism. Two viable alternatives are: sliding windows and fading factors. Its observe that the prequential error converges to a holdout estimator when estimated over a sliding window or using fading factors. In [5] authors proposed a general framework for mining concept-drifting data streams using weighted ensemble classifiers. We train an ensemble of classification models, such as C4.5, RIPPER, naive Bayesian, etc., from sequential chunks of the data stream. The classifiers in the ensemble are judiciously weighted based on their expected classification accuracy on the test data under the time-evolving environment. Thus, the ensemble approach improves both the efficiency in learning the model and the accuracy in performing classification. In [6] Authors had proposed the Ensemble methods have recently garnered a great deal of attention in the machine learning community. Techniques such as Boosting and Bagging have proven to be highly effective but require repeated re-sampling of the training data, making them inappropriate in a data mining context. The methods presented in this paper take advantage of plentiful data, building separate classifiers on sequential chunks of training points. These classifiers are combined into a fixed-size ensemble using a heuristic replacement strategy. In [7] authors illustrated the problem of learning from concept drifting data streams with noise, where samples in a data stream may be mislabeled or contain erroneous values. The essential goal is to build a robust prediction model from noisy stream data to accurately predict future samples. For noisy data sources, most existing works rely on data preprocessing techniques to cleanse noisy samples before the training of decision models. In data stream environments, these data preprocessing techniques are, unfortunately, hard to apply, mainly because the concept drifting in a data stream may make it very difficult to differentiate noise from samples of changing concepts.

III. PROPOSED ALGORITHM

A. *Bayesian Decision Trees (BDT):*

Bayesian Decision trees are prevailing and popular tools for classification and prediction. A BDT tree is a flow chart like ensemble tree structure, where each interior node denotes a test on an attribute, each division represents an outcome of the test, and child nodes represent classes or class distribution. The main idea of BDT classification models is to recursively partition data nodes in an analogous approach. Such that arbitrary approach provides a natural attribute selection and uncover the attributes which make the important part to the classification.

A BDT tree is an analytical machine-learning approach that decides the final value (dependent variable) of a new sample based (Back propagation) on various attribute values of the available data. The internal nodes of a decision tree

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

denote the different variables; the branches between the nodes tell us the promising values that these variables can have in the observed samples, while the deadly nodes tell us the final value (classification) of the dependent variable. The attribute that is to be predicted is known as the dependent variable, since its value depends upon, or is decided by, the values of all the other attributes.

B. *Ensemble-tree spatial indexing structure*

The Ensemble-tree spatial indexing structure consists in training the different classifiers with bootstrapped duplications of the original training data-set. That is, a new data-set is shaped to train each classifier by randomly drawing (with replacement) samples from the original data-set (usually, maintaining the original data-set size).

An indexing algorithm does not require re-computing any kind of weights; therefore, neither is necessary to adjust the weight update formula nor to change computations in the algorithm. In these methods, the key factor is the way to collect each spatial indexing order Logical decision (Algorithm 1), that is, how the class indexing problem is dealt to obtain a useful classifier in each iteration without forgetting the importance of the diversity.

Algorithm 1: Spatial Indexing order Logical Decision
Input: *S*: Data Stream Training set; *T*: Number of iterations;
N: Indexing size; *I*: Logical learner
Output: Logical Ensemble classifier: H(x) = $\sum_{t=1}^{T} h_t(x)$sign where $h_t \in$ [-1, 1] are the induced classifiers
For t = 1 to T do
$S_t \leftarrow$ RandomSampleReplacement (n, S)
$h_t \leftarrow$I (S_t)
End for

C. *Logical Decision Trees*

The logical decision trees applied within the framework of the Bayesian decision theory form the basic framework of ensemble for fast prediction. This may be realized by classification of the fact that data stream analysis serves the purpose of decision-making. Either the indexing structure analysis shows that the structures are acceptable and one does nothing, otherwise it is found that the decision are not acceptable and one has to do something. The logical decision analysis is the proposed framework for the evaluation of the spatial arbitrary as well as for the evaluation of how to reduce the complexity most efficiently in time manner.

IV. CONCLUSION AND FUTURE WORK

In this paper proposed the E-tree spatial indexing structure based on logical Bayesian decision tree algorithm for the address the prediction efficiency for ensemble models on data streams problem. The proposed technique using an extensive approach outperforms the logical Bayesian DT technique in terms of classification uncertainty. The suggested technique also provides shortest DTs which can be easily interpreted by all domain experts.

In future work, we intend to enhance the logical decision algorithm to develop the experimental methods for non-linear optimization to control the growth of tree attributes of the result data.

REFERENCES

1. C. Aggarwal, Data Streams: Models and Algorithms. Springer, 2006.
2. P. Zhang, J. Li, P. Wang, B. Gao, X. Zhu, and L. Guo, "Enabling Fast Prediction for Ensemble Models on Data Streams," Proc. 17th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), 2011.
3. 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 Trans. Knowledge and Data Eng., vol. 23, no. 6, pp. 859-874, June 2011.
4. J. Gao, R. Sebastiao, and P. Rodrigues, "Issues in Evaluation of Stream Learning Algorithms," Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), 2009.
5. H. Wang, W. Fan, P. Yu, and J. Han, "Mining Concept-Drifting Data Streams Using Ensemble Classifiers," Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), 2003.
6. W. Street and Y. Kim, "A Streaming Ensemble Algorithm (SEA) for Large-Scale Classification," Proc. Seventh ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), 2001.
7. P. Zhang, X. Zhu, Y. Shi, L. Guo, and X. Wu, "Robust Ensemble Learning for Mining Noisy Data Streams," Decision Support Systems, vol. 50, no. 2, pp. 469-479, 2011.
8. L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees. Belmont, CA: Wadsworth, 1984.
9. W. Buntine, "Learning classification trees", Statistics and Computing, 2, pp. 63-73, 1992.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

10. A. Bifet, G. Holmes, B. Pfahringer, R. Kirkby, and R. Gavalda, "New Ensemble Methods for Evolving Data Streams," Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), 2009.
11. P. Zhang, X. Zhu, J. Tan, and L. Guo, "Classifier and Cluster Ensembles for Mining Concept Drifting Data Streams," Proc. IEEE 10th Int'l Conf. Data Mining (ICDM), 2010.

BIOGRAPHY

Dr.M.Rathamani received M.Phil degree at Manonmaniam Sundaranar University. She completed Ph.D. from Mother Teresa Women's University, Kodaikanal. She is an assistant professor at P.G.Department of Computer Applications, Nallamuthu Gounder Mahalingam College, Pollachi. She has over 18 years of teaching experience and 10 years of research experience. She attended International/National conferences and published more than 25 papers. She is the author/Co author of more than ten publications. Her area of research is Cloud Computing and Image Processing.

S.Mohanapriya is a Research Scholar in Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi. She received her master of science (M.Sc) in 2014 from Nallamuthu Gounder Mahalingam College, Pollachi under Bharathiar University, Coimbatore. She has presented papers in International/National conferences and attended workshop, seminars. Her research focuses on Data mining.