ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2015

Ensemble Models Using Logical Bayesian Decision Tree For Stream Classification And Indexing Data

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

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

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

ABSTRACT: Data Mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data Mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high performance client/server or parallel processing computers, data mining tools can analyse massive databases.

KEYWORDS: Stream classification**,** VFDT, K-d trees, SVM Models.

I. INTRODUCTION

Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. Among several types of knowledge representation present in the literature, classification normally uses prediction rules to express knowledge. Prediction rules are expressed in the form of IF-THEN rules, where the antecedent (IF part) consists of a conjunction of conditions and the rule consequent (THEN part) predicts a certain predictions attribute value for an item that satisfies the antecedent

Training set

Age	Heart Rate	Blood pressure	Hearty problem
65	78	150/70	Yes
37	83	112/76	No
71	67	108/65	No

Prediction set

Age	Heart Rate	Blood pressure	Hearty problem
43	89	147/89	?
65	57	106/60	?
84	72	158/65	?

Table 1 – Training and prediction sets for medical database

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2015

Using the example above, a rule predicting the first row in the training set may be represented as following:
IF (Age=65 AND Heart rate>70) OR (Age>60 AND Blood pressure>140/70) THEN Heart problem=yes
In most cases the prediction rule is immensely larger than the example above. Conjunction has a nice property for classification; each condition separated by OR's defines smaller rules that captures relations between attributes. The decision tree learning method is one of the methods that are used for classification or diagnosis. Decision tree learning has several advantages. One of the advantages is that it gives a graphical representation of the classifiers which makes it easier to understand.

The decision tree algorithm

There are several procedures and methods for building decision trees, such as ID3, C4.5 and CART. ID3 is a simple decision tree learning algorithm developed by Ross Quinlan (1983). The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node. Advantages of Decision Tree Classifications

- Decision Trees able to generate understandable rules,
- They are able to handle both numerical and the categorical attributes,
- They provide a clear indication of which fields are most important for prediction

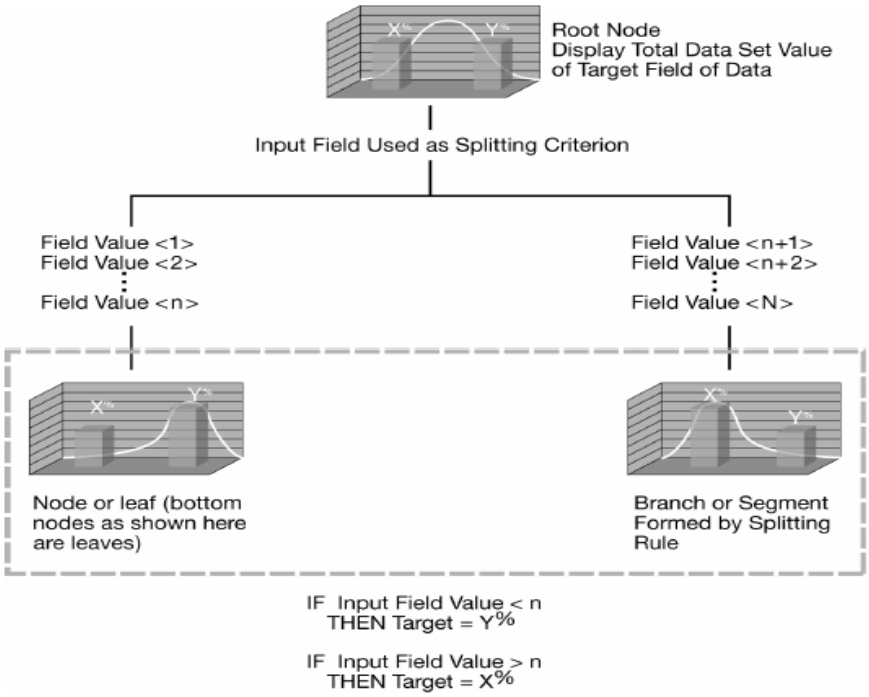


Figure 1: Illustration of the Decision Tree

II. RELATED WORK

Stream classification: Existing data stream classification models can be roughly categorized into two groups: online/incremental models and ensemble learning.

Stream indexing: A stream classification model can be considered as a special query model that queries for class labels of incoming stream records. However, it differs from traditional query models, such as aggregate query

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2015

that aims to obtain statistical results from data streams, and Boolean expression query where queries appear as DNF or CNF expressions .

K-d trees and M-trees: K-d trees, M-trees and R-trees are popularly used space-partitioning structures for query high dimensional data. The major difference is that K-d trees and M-trees are designed for organizing points, while R-tree is designed for organizing rectangles.

Ensemble pruning: For the purpose of reducing computational and memory costs, ensemble pruning searches for a good subset of all ensemble members that perform as good as the original ensemble.

Random Subspace sampling method

The Random Sampling is a popular methodology to counter the problem of class imbalance. However sampling methods, especially SMOTE (Synthetic Minority Oversampling Technique), work with the entire set of features or dimensions and may not be as effective in reducing the scarcity of the data.

Decision tree

The decision tree induction presented in Algorithm 1 results in decision-Tree. Consider a data set with features $\{X'_i\}^j_{i=1}$ and an imbalanced binary target class Y . To keep the notation simple, to transform the features $\{X'_i\}^j_{i=1}$into binary features$\{X_i\}^n_{i=1}$, i.e., $\{X_i\}^n_{i=1}$Binarize($\{X'_i\}^j_{i=1}$). The inputs to the **α** -Tree algorithm are the binarized features $\{X_i\}^n_{i=1,}$ the corresponding class labels Y , and a pre-specified α.

Algorithm 1. *Grow a single Decision -Tree.*

Input: *training_data (features $X_1, X_2, \ldots X_n$ and label Y), α*

Select the best feature X, which gives the maximum* ***α -divergence criterion***

if *(number_of_data_points < cut_off_size) or*

(best_feature = None) ***then***

Stop growing

else

partition the training data into two subsets,

*based on the value of X**

left child ← Grow a single α -Tree

(data with (X = 1), α)*

right child ← Grow a single α -Tree

(data with (X = 0), α)*

end if

III.PROPOSED ALGORITHM

Data Pre-processing Methods The data pre-processing techniques can be easily embedded in ensemble learning algorithms. A re-sampling techniques that study the effect of changing class distribution to deal with imbalanced data-sets, where it has been empirically proved that the application of a pre-processing step in order to balance the class distribution is usually a positive solution. **Random under sampling:** It is a non-heuristic method that aims to balance class distribution through the random elimination of majority class examples. **Random oversampling:** In the same way as random under sampling, it tries to balance class distribution, but in this case, randomly replicating minority class instances. **Selective pre-processing of imbalanced data (SPIDER):** It combines local oversampling of

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2015

the minority class with filtering difficult examples from the majority class. It consists in two phases, identification and pre-processing.

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. 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.

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. 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.

Algorithm 3: procedure Logical classify (s : sample) returns class:
C := true
N := root
while *N ≠ leaf(c)* ***do***
let *N =* ***inode****(conj, left, right)*
if *C ^ conj succeeds in LDT ^ s*
then *C := C ^ conj*
N := left
else *N := right*
return *c*

IV. PERFORMANCE ANALYSIS

The research study performs three real-world streams [25] and one synthetic steam from the UCI repository was used [25]. Table 1 lists these streams after using the F-Score feature selection [26]. The synthetic stream is similar to Example 1. It was generated as follows. First, we randomly generated a collection of stream data {…, (x_i, y_i),…}, where $x_i \in R^d$ is a d-dimensional vector with the j^{th} element $x_{ij} \in [0, 1]$, and $y_i \in$ {normal, abnormal} is the class label. Thus the decision space is a d-dimensional hyper-cube. The class of each stream record is defined by the rule that if $a \leq x_i \leq b$, where $a \in [0, 1]^d$, $b \in [0, 1]^d$, $a_i < b_i$ for each i $\in [1, d]$, then "abnormal"; otherwise, "normal".

Measures: Three measures are used for online query evaluation. 1) Time cost.2)memory cost, 3)Accuracy. E-trees are expected to consume a larger but affordable size of memory space.Proposed systems are expected to achieve the same prediction accuracy as original ensemble models.

Table 1: Real-World Data Streams

Name	Instances	Attributes	Areas
Intrusion Detection	4000000	11	Security
Spam Detection	460001	15	Security
Malicious URL Detection	488420	14	Security

Table 1 show the real-world streams and one synthetic steam from the UCI repository were used in this research.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2015

Table 2: Comparison Methods of imbalance classification of Accuracy values

Methods	GE-tree	G-Ensemble	LE-tree	L-Ensemble	LB-tree	LBET
Intrusion Detection	0.91	0.915	0.939	0.934	0.93	**0.9412**
Spam Detection	0.6	0.68	0.84	0.89	0.91	**0.95**
Malicious Detection	0.85	0.80	0.71	0.65	0.80	**0.89**

In Table 2 describes the classification accuracy of proposed *LBET* method is produce better accuracy with existing ensemble methods.

Comparison Methods of imbalance classification of Accuracy values

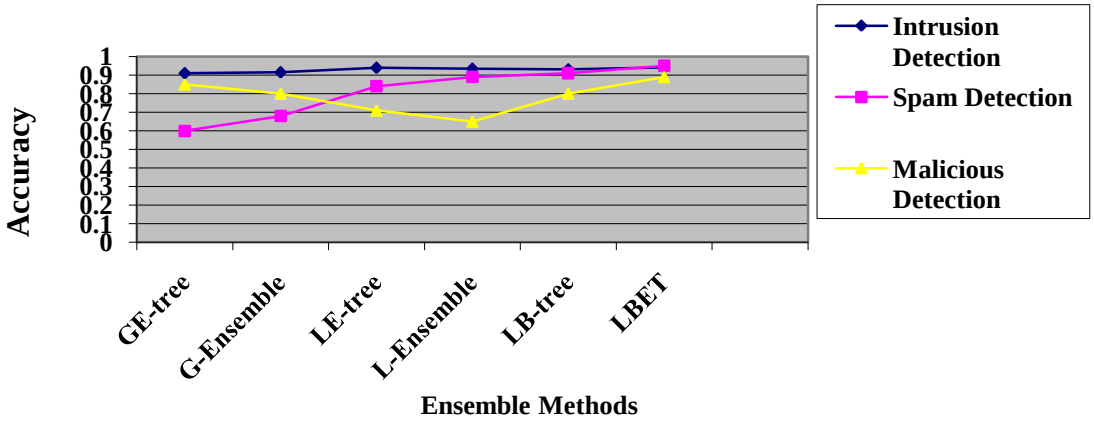


Table 3: Time Cost Comparisons with Benchmark Methods

Methods	GE-tree	G-Ensemble	LE-tree	L-Ensemble	LB-tree	LBET
Intrusion Detection	444.5	169	62	337	61.5	**58**
Spam Detection	5933	368969	3484	47406	3207	**3155**
Malicious Detection	1153	133600	782	12348	765	**68**

In Table 3 describes the Time cost comparison measures of proposed *LBET* method is produce better time cost with existing ensemble methods.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2015

Time Cost Comparisons with Benchmark Methods

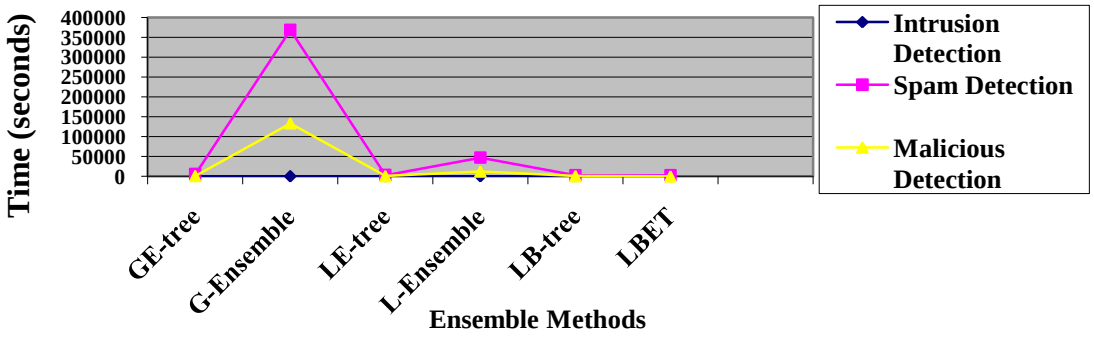
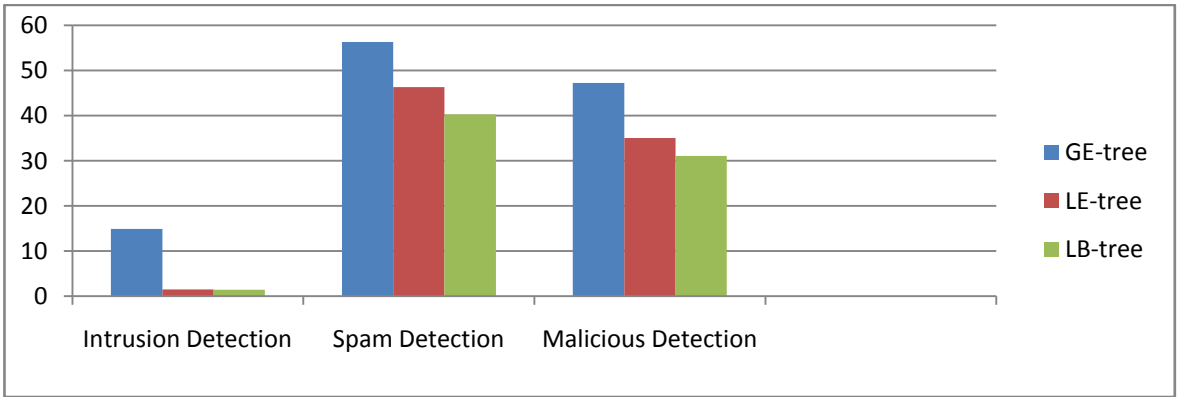


Table 4: Memory size Comparisons with Benchmark Methods

	GE-tree	LE-tree	LB-tree
Intrusion Detection	14.90	1.50	**1.45**
Spam Detection	56.29	46.29	**40.27**
Malicious Detection	47.26	35.04	**31.04**

In Table 4 describes the memory size comparisons of proposed *LBET* method is produce better memory consumption over the existing ensemble methods.



V.CONCLUSION

The research presents a new approach which combines techniques from various fields and adapts to solve the problem of data streaming, decision tree and indexing. The results show that the proposals generated the extremely relevant. It has been observed that as the Area under curve (i.e., Accuracy) and Time cost prevents the quality of proposed system. The classification technique is developed on the basis of logical Bayesian DT methodology has revealed promising results in our experiments on predicting the occurrences of uncertainty. Both the standard and suggested Bayesian DT techniques provide a high performance in terms of predictive accuracy

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2015

REFERENCES

[1]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.

[2] 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.

[3] 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.

[4] 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.

[5] X. Zhu, P. Zhang, X. Lin, and Y. Shi, *"Active Learning from Stream Data Using Optimal Weight Classifier Ensemble,"* IEEE Trans. System, Man, Cybernetics, Part B: Cybernetics, vol. 40, no. 4, pp. 1-15, Dec. 2010.

[6] J. Quinlan, C4.5:*" Programs for Machine Learning"*. Morgan Kaufmann Publishers, 1993.

[7] J. Ma, L. Saul, S. Savage, and G. Voelker, *"Identifying Suspicious URLs: An Application of Large-Scale Online Learning,"* Proc. 26th Ann. Int'l Conf. Machine Learning (ICML), 2009.

[8] R. Guting, *"An Introduction to Spatial Database Systems,"* VLDB J., vol. 3, no. 4, pp. 357-399, 1994.

[9] A. Guttman, *"R-Trees: A Dynamic Index Structure for Spatial Searching,"* Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 1984.

[10] T. Sellis, N. Roussopoulos, and C. Faloutsos, *"The R-Tree: An Efficient and Robust Access Method for Points and Rectangles,"* Proc. ACMSIGMOD Int'l Conf. Management of Data (SIGMOD), 1990.

[11] T. Sellis, N. Roussopoulos, and C. Faloutsos, *"The Rþ-tree: A Dynamic Index for Multi-Dimensional Objects,"* Proc. 13th Int'l Conf. Very Large Data Bases (VLDB), 1987.

[12] P. Ciaccia, M. Patella, and P. Zezula, *"M-Tree: An Efficient Access Method for Similarity Search in Metric Spaces,"* Proc. 23rd Int'l Conf. Very Large Data Bases (VLDB), 1997.

[13] T. Sellis, N. Roussopoulos, and C. Faloutsos, *"Multi-Dimensional Access Methods: Trees have Grown Everywhere,"* Proc. 23rd Int'l Conf. Very Large Data Bases (VLDB), 1997.