IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 53-59
www.iosrjournals.org

A Decision Support System for Managing Electricity Demand Using an Incrementally Optimized Very Fast Decision Tree Ensemble

[1]Surya S, [2]Sangeeta Srinivas
Final Year MTech CSE Computer Science and Engineering Department Sree Narayana Gurukulam Engineering College Kadayiruppu, Kolenchery.
Associate Professor Computer Science and Engineering Department Sree Narayana Gurukulam Engineering College Kadayiruppu, Kolenchery

Abstract: *Electric utilities such as Smart Grids incorporate intelligent methods to efficiently manage the day to day operations. The consumption behavior of consumers can be classified from streaming consumption data and abnormal consumption behavior patterns can be obtained which can be utilized for fine-grained management of power demand. Mining streaming data is a challenging task owing to the highly resource constrained setting in terms of memory and learning time. Incrementally Optimized Very Fast Decision Tree (iOVFDT) is an efficient method for Streaming Data Classification. iOVFDT achieves a compact and reliable model by efficiently balancing the accuracy, tree size and learning time. To enhance the classification performance multiple iOVFDT classifiers can be combined into an efficient ensemble. In this paper we propose a methodology for developing an adaptive and computationally lightweight Incrementally Optimized Very Fast Decision Tree Ensemble and present its application as a decision support mechanism for managing Electricity Demand.*
Keywords: *Data mining, stream mining, iOVFDT ensemble, streaming data processing, situational awareness.*

I. INTRODUCTION

Electricity is a scarce resource which cannot be stored on a massive scale. Electric Utilities can obtain streaming consumption data of consumers in the form of smart meter data. Smart meters are devices which ensure two way communications between the consumer and Utility. By analyzing streaming smart meter data useful insights can be obtained for fine-grained management of electricity demand. Stream mining typically applies to those systems which need some actionable knowledge with minimum latency. Traditional batch oriented techniques doesn't fit well into the stream setting. Hence stream data mining requires some specialized techniques and algorithms which can handle the streaming scenario. Streaming data refers to continuously arriving, high speed and potentially infinite amount of data. The mining algorithm must be well equipped to ingest such high speed data. The traditional batch methods require the entire data to be in main memory before building the model. Also it may require several passes over the data to build the model. Large amount of continuously arriving training data thus brings considerable overhead in terms of memory and computation time for traditional algorithms. Owing to special characteristics of Data Streams incremental learning methods are required to process such Data and make predictions within limited amount of time. Very Fast Decision Tree called VFDT is an incremental tree induction algorithm specially suited for streaming data classification. VFDT uses Hoeffding Bound for split evaluation for Model growth. Incrementally Optimized Very Fast Decision Tree called iOVFDT is an improved version of VFDT which uses a reliable multi-objective incremental optimization mechanism for node splitting in addition to the Hoeffding Bound criteria. iOVFDT also uses functional tree leaf for enhanced prediction. iOVFDT achieves a compact Model by efficiently balancing accuracy, size and learning time. One of the limitations of iOVFDT is the inability to adapt its structure to changing concepts. The use of multiple iOVFDT learners can improve the reliability of the model for decision making. The proposed classifier incorporates random subspace attribute selection mechanism and dynamic adaptation capability into iOVFDT. The resulting classifier is applied as the base learner in an online bagged ensemble. Leveraging bagging is the online bagging method used. The proposed ensemble can be used to classify the power consumption behavior of individual consumers by analysing high resolution streaming smart meter data. Consumption behaviors are accumulated for very short intervals to yield the most recent behavior patterns. A rule based consumption behavior pattern analysis is performed and timely alerts for abnormal usage patterns are generated for both individual and aggregate levels. Such a system can promote conservation of Electricity thus benefitting both the consumer and the utility. These alerts help in creating a near-real time situational awareness among the consumers to achieve voluntary regulation of their usage levels thereby bringing down the stress on

the Utility. The consumer can track his usage as well as voluntarily respond to alerts when the grid goes under stress or receive incentives for reduced usage or timely response to emergency signals from the Utility. The consumption alerts at the aggregate level can be used by the Utility as an early warning signal for an emerging abnormal consumption behavior and readjust their demand supply planning in a fast and fine-grained manner.

II. RELATED WORK.

Domingos and Hulten[1] proposed Hoeffding Tree algorithm for streaming data classification. The classifier can handle huge amount of training data which would otherwise bring memory and computational overhead for traditional batch oriented classifiers. Hoeffding Tree never stores examples in main memory for processing. The classifier keeps the sufficient statistics in its leaves. The model is incrementally grown unlike the batch versions. Every instance is routed to leaf by following a specific root to leaf path. When the count of instances accumulating at a leaf reaches a threshold an evaluation for growing the model is performed. VFDT periodically evaluates the need for growing the current model. This involves the computation of information gain for all the attributes for an example. The attribute which obtains the highest Information gain is selected as the best candidate. Hoeffding Bound (HB) is used to decide when a split evaluation should be performed for Model growth. When the difference in Information Gain of the best attribute and second best attribute exceeds the Hoeffding Bound a split is enforced. Hoeffding Bound states with a probability 1-δ that if n observations of a random variable is made and their computed mean is m then the true mean is at least m-ε.Hoeffding Bound is given by

$$\varepsilon = \sqrt{\frac{R^2 \ln\left(\frac{1}{\delta}\right)}{2n}}$$

where R is the range of the random variable. Here the random variable is the difference in information gain of the best attribute and second best attribute and the range R is calculated as the base 2 logarithm of the number of class labels possible.The number of examples or observations is given by n.The model thus grows incrementally and evolves over time until the entire training examples have arrived.Ties occur when the difference in Information gain between the best attribute and second best attribute is smaller than Hoeffding Bound.As a result a split cannot be enforced. If such conditions persist ties may take long time to resolve. Such evaluations are performed periodically regardless of split enforcement. Whenever a tie occurs and the longer it takes to resolve the tie unnecessary overhead in terms of periodic split evaluations will be incurred. Very Fast Decision Tree uses a user defined threshold to resolve ties between the best attribute and the second best attribute. This was proposed as a remedial mechanism to break ties where only minor difference exist between the Information Gain of the two attributes which may be smaller than the computed Hoeffding Bound measure.

Hang Yang and Simon Fong [2] proposed Incrementally Optimized Very Fast Decision Tree (iOVFDT) as an improved version of existing VFDT Classifier. iOVFDT overcomes the limitations of the VFDT for tie-breakingVery Fast Decision Tree uses a user defined threshold to resolve ties between the best attribute and the second best attribute.This was proposed as a remedial measure to break ties where only minor difference exist between the Information Gain of the two attributes which may be smaller than the computed Hoeffding Bound measure and enhances the predictive performance at the leaves.In iOVFDT an adaptive tie breaking Mechanism is used instead of a fixed user defined Threshold. A multi-objective optimization is used to attain a balance between accuracy, tree size and learning time by maintaining the Model Cost within an Optimum Range. The Model Cost is computed as a Function of Error, Tree Size and Learning Time every time a predefined number of examples arrive at the leaf. This information is used for Split Evaluation in addition to the Hoeffding Bound Criteria used by VFDT.The Model is grown either on meeting the Hoeffding Bound Criteria or when the Model Cost computed as the multi-objective function of Error, Tree Size and learning time goes out of the optimum Range. For enhancing the predictive performance at the leaves weighted Naive Bayes Classifier and an Error Adaptive mechanism which are embedded at the leaves of the Classifier for enhancing the predictive accuracy. iOVFDT cannot work well with data whose distribution changes with time. Albert Bifet and Richard Gavalda [3] proposed an extended version of VFDT called Hoeffding Adaptive Tree(HAT) for detecting the change in data distribution and dynamically adapt to these changed data. HAT-ADWIN is a variant which uses ADWIN as change detector and estimator. ADWIN raises an alarm when a change is detected. This indicates a possible drift in concept.

A computationally light weight learner is highly desirable for an ensemble in the streaming scenario. Diversity among base learners is another requirement to offer a good classification performance. Breieman [4] proposed Random Forests a batch ensemble which uses Random subspace method for attribute selection for

constructing each base learner and resampling with replacement for input space of each base learner. Abdulsalam et al. [5] proposed a streaming variant of random forests. Albert Bifet et al.[6] proposed Leveraging Bagging as an improved version of Online Bagging with three variants. Leveraging Bagging increases the diversity of the weights by using a Poisson distribution with a mean greater than 1.This increases the diversity of input space for the learners in the ensemble thereby improving the performance.

III. PROPOSED METHODOLOGY.

In this paper we propose a methodology for developing a lightweight and adaptive Incrementally Optimized Vey Fast Decision Tree Ensemble for streaming data classification. We attempt to combine the multi-objective node splitting mechanism and functional tree leaf for enhanced prediction in iOVFDT with dynamic adaptation ability and random subspace feature selection to develop the base learner of the proposed ensemble. We used an improved version of Online Bagging called Leveraging Bagging for building the ensemble. Leveraging Bagging is used to improve the diversity of input space for the proposed Ensemble. We call the proposed Ensemble as Adaptive Random iOVFDT Ensemble. We apply the proposed Ensemble as a decision support mechanism for managing electricity demand by analysing streaming smart meter data.

Assumptions
1 N- No of examples seen at the leaf
2. N_{min} –Minimum no of new examples to be seen at the leaf to perform a split evaluation
3.G_a-Information Gain of best attribute a
4.G_b-Information Gain of second best attribute b
5.$G_{diff}=G_a-G_b$
6. ε =Hoeffding Bound
7.$Cost_{current}$-Current Cost of the Model computed as a Cost function of current Error Size and Learning Time
8.$Cost_{min}$ –Minimun acceptable Cost of the Model.
9.$Cost_{max}$-Maximum acceptable cost of the Model
10.$Cost_{min}$ and $Cost_{max}$ are initialized to predefined ideal values
11.Split Node is a non leaf node.
12.MC-Majority Class
13.NB –Naive Bayes method.
14.WNB-Weighted Naive Bayes method

A. iOVFDT Algorithm
1. Sort the example to a leaf.
2. Update the statistics at the leaf.
3. Compare the predicted class to actual class
4. Update the Error
5. If N mod N_{min}=0 and instances seen so far are not of same class then
6. Update the learning time
7. Compute the information gain of all the attributes.
8. $G_{diff}=G_a-G_b$
9. Compute $Cost_{current}$
10 . If $G_{diff}>\varepsilon$ or $Cost_{current} > Cost_{max}$ or $Cost_{current} < Cost_{min}$
11. Enforce a split on the best Attribute
12. Grow the tree with 2 leaves
13. Update the tree size
14 . If $Cost_{current} > Cost_{max}$ then $Cost_{max} = Cost_{current}$
15. If $Cost_{current} < Cost_{min}$ then $Cost_{min} = Cost_{current}$

B. Update Statistics at leaf
1. Classify the example using MC,NB and WNB methods
2. Find the Classification error of MC, NB and WNB 3.Assign the class label predicted by the method with least classification error.
4.Update the Error and attribute class statistics at the leaf.

C. Incorporating Adaptive learning in iOVFDT
Incrementally Optimized Very Fast Decision Tree (iOVFDT) lacks the ability to dynamically adapt its structure with changing concepts. To develop an adaptive variant of iOVFDT the features of iOVFDT with the adaptation

ability of the extended version of VFDT called Adaptive Hoeffding Tree are combined. The dynamic adaptation capability is incorporated into iOVFDT by embedding instances of ADWIN change detectors in every node. These change detectors are made to continuously monitor any change in data distribution when each training example pass down the tree. For each example ADWIN monitors the classification error rate of the sub trees rooted at every split node in the path taken by the example. When there is significant increase in the error rate at a node a change is signalled which indicates a change in data distribution. The steps for adapting the iOVFDT model on change detection at a split node are given below.

1. Grow an alternate sub tree
2. Monitor the average error for the sub tree rooted at that node and the new alternate sub tree
3. If the error rate of the sub tree exceeds that of alternate sub tree by a bound
4. Replace the sub tree with the alternate sub tree.

D. Using Random Subspace Feature Selection for Split Evaluation in iOVFDT

The formation of new nodes or branches during adaptation invokes the split evaluation process at a leaf. Incrementally Optimized Very Fast Decision Tree considers all attributes for split evaluation. To reduce the computation for each base learner in the proposed ensemble Random subspace method is used for attribute selection at every split evaluation in the base learner during learning phase. A subset of attributes is randomly chosen for finding the best attribute. The number of attributes in the subset is equal to the square root of the original number of attributes. The steps for incorporating multi-objective node split evaluation using a random subset of features in iOVFDT are as follows.

1. Sort the example to a leaf.
2. Update the statistics at the leaf.
3. Compare predicted Class to Actual Class
4. Update the Error
5. If N mod N_{min}=0 and instances seen so far are not of same class then
5.1. Update the learning time
5.2. Compute R as the square root of no of attributes
5.3. Randomly choose R number of attributes
5.4. Compute the information gain of R attributes
5.5 $G_{diff}=G_a-G_b$
5.5. Compute $Cost_{current}$
5.6. If $G_{diff} > \varepsilon$ or $Cost_{current} > Cost_{max}$ or $Cost_{current} < Cost_{min}$
5.7. Enforce a split on the best Attribute
5.8. Grow the tree with 2 leaves
5.9. Update the tree size
5.10. If $Cost_{current} > Cost_{max}$ then $Cost_{max} = Cost_{current}$
5.11. If $Cost_{current} < Cost_{min}$ then $Cost_{min} = Cost_{current}$

E. Building the proposed Ensemble using Leveraging Bagging

The variant of iOVFDT incorporated with adaptive learning ability and random subspace feature selection is used as the base learner for the proposed ensemble. An improved version of online bagging called Leveraging Bagging is used to build the ensemble which adds more randomization on the weights of the input of the classifiers. This variant uses a Poisson distribution with mean greater than 1 to assign different random weights for each of the examples fed to different base learners. The weighting is done using a Poisson random variable corresponding to a randomly chosen probability in the Poisson distribution for each base learner in the ensemble. A larger mean of Poisson distribution ensures a greater diversity of weights for the base learners and enhances the performance of the ensemble by improving the diversity of the input space. ADWIN is used to monitor the error of each base learner in the ensemble for each training example. The worst leaner is removed from the ensemble and replaced by a new learner whenever the error rate of the learner exceeds a bound.

1. Set $\lambda>1$
2. For each example
3. For each base learner in the ensemble
3.1. Randomly Weight the instance with Poisson (λ)
3.2. Input the weighted instance to the learner
4. Estimate the error of the base learner

5. If a change is detected
6. Remove the worst learner from the ensemble 9 7. Replace with a new learner
8. Compute the sum of weighted Votes of all learners. 9. Assign the class which receives the highest Votes

F. Applying the proposed Ensemble for Decision support for managing Electricity demand

The proposed Adaptive Random iOVFDT Ensemble is used for classifying the consumption behavior of each consumer in real time from respective smart meter data streams. The classified behaviour of each consumer is accumulated over a short interval in near real time basis to obtain a consumption pattern for that interval. A rule based consumption pattern analysis is done to detect abnormal usage patterns at both the individual and aggregate level and appropriate alerts are generated. The consumption behavior changes with different seasons, holidays as well as some unanticipated events. The proposed ensemble being adaptive can be retrained while in operation with newly labelled data to accommodate the changing conditions.

IV. EXPERIMENTAL RESULTS

The single iOVFDT classifier and the proposed Adaptive Random iOVFDT Ensemble is evaluated for various real world and synthetic datasets which includes the Smart Meter data used for our work. Due to privacy concerns of Smart Grid and its consumers the Smart Meter data required for our work has been generated using Java code. The performance of both classifiers is compared using an interleaved test-then train approach. The ensemble size of the proposed classifier is set as 10. A Poisson (6) distribution is used in Leveraging Bagging method for building the proposed Ensemble for all the datasets. The Accuracy comparison between iOVFDT and the proposed Adaptive Random iOVFDT ensemble is given in Table 2 and Figure 1 and learning Time comparison is given in Table 3 .

The proposed Adaptive Random iOVFDT Ensemble has produced improved Classification accuracy than a single iOVFDT Classifier. The random subspace attribute selection process takes less computation for each base learner in the proposed ensemble. The use of Random subspace method reduced the learning time for the proposed ensemble. To test the adaptation ability the proposed ensemble is initially trained with smart meter consumption data for moderate season .The classifier generated a high alert for high consumption during moderate season. The classifier is then retrained on the fly with consumption data for summer. The consumption data which generated high alert for moderate serason is again fed to the classifier after retraining with summer data. No alerts for high consumption were generated this time which shows the adaptation ability of the proposed ensemble.

TABLE 1 Data used for evaluating the classifiers

Data	Instances	Attributes	Classes
Person Activity	164000	6	8
Electricity	45000	9	2
Random RBF Generator	500000	10	5
Cover Type	500000	8	7
Waveform Generator Drift	500000	21	3
Smart Meter	240000	8	3
Network Attack	25000	41	21
Airlines	500000	7	2

TABLE 2 Accuracy Comparisons

Data	iOVFDT	Adaptive Random iOVFDT Ensemble
Person Activity	80.25	93.59
Electricity	82.49	89.18
Random RBF Generator	91.51	94.10
Cover Type	90.14	92.06
Waveform Generator Drift	82.58	83.74
Smart Meter	94.14	98.19
Network Attack	97.39	98.38

Data	iOVFDT	Adaptive Random iOVFDT Ensemble
Airlines	65.09	67.04

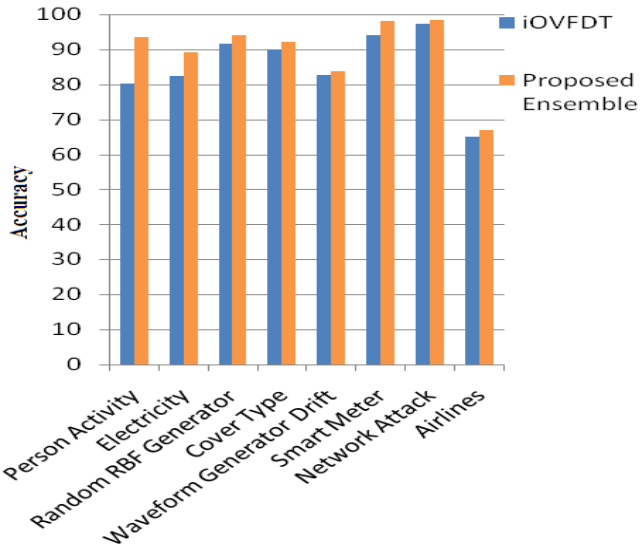

Fig 1 Comparison of iOVFDT and Proposed Adaptive Random iOVFDT Ensemble

TABLE 3 Learning Time Comparison

Data	iOVFDT	Adaptive Random iOVFDT Ensemble
Person Activity	11.53	14.22
Electricity	1.77	8.19
Random RBF Generator	62.92	173.64
Cover Type	138.92	191.80
Waveform Generator Drift	103.22	237.06
Smart Meter	3.22	26.70
Network Attack	1.70	6.66
Airlines	734.8	91.38

V. CONCLUSION

The proposed iOVFDT Ensemble achieves improved classification accuracy with a lightweight computational approach for multiclass classification problems on streaming data when compared to a single iOVFDT classifier. The classifier can be retrained online through dynamic adaptation ability. Random subspace method increases the diversity of the ensemble by introducing randomness at every node in the base learners which enhances the accuracy of the ensemble .The improved diversity of input space provided by Leveraging Bagging improves the classification performance of the proposed ensemble. Random subspace attribute selection reduces the computational cost of the proposed ensemble. This facilitates faster training on large amount of training data and faster adaptation of the ensemble to changing concepts. The use of the proposed iOVFDT Ensemble is successfully applied as a decision support mechanism for managing Electricity demand. As future work we plan for a large scale deployment of our work in a resource rich Cloud Computing environment.

REFERENCES

[1] Domingos, P., &Hulten, G. (2000). Mining high-speed data streams. In Proceedings ofthe sixth ACM SIGKDD international conference on Knowledge discovery and datamining (pp. 71–80).
[2] Hang Yang, Simon Fong, Yain-Whar Si, "Multi-objective Optimization for Incremental Decision Tree Learning" 14th International Conference on Data Warehousing and Knowledge Discovery, Springer LNCS, Vienna (Austria) September 3 - 6, 2012
[3] Bifet, A., Gavald`a, R.: Adaptive learning from evolving data streams. In: IDA (2009).
[4] L. Breiman. "Random forests". Machine Learning, 45(1):5–32, 2001
[5] H. Abdulsalam, D. B. Skillicorn, and P. Martin. Streaming random forests. In IDEAS '07: Proceedings of the 11th International Database Engineering and Applications Symposium, pages 225{232, Washington, DC, USA, 2007. IEEE Computer Society.
[6] Albert Bifet, Geoffrey Holmes, Bernhard Pfahringer. Leveraging Bagging for Evolving Data Streams Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD, 2010