International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 04, 2020
ISSN: 1475-7192

LASH Tree: LASSO Regression Hoeffding for Streaming Data

[1]D.Christy Sujatha,[2] Dr.J.Gnana Jayanthi

ABSTRACT --*Streaming data is a challenging research area for the last two decades which comes in high volume and rapid speed and cannot be stored using existing memory. Dealing with model adaptability with evolving data over time and memory usage arethe major challenges in streaming data predictive models. Recently there is a rising attention in developing Regression Tree models due to it's high interpretability and accuracy. Additionally, the linear function at the leaf node evaluates the target variable more accurately by analysing the correlation between predictor variables and target variable. The proposed LASSO Regression Hoeffding Tree (LASH Tree) is a Regression Tree model which incorporates LASSO Regression with Hoeffding Tree that produces better predictions and better insights. In this paper, an exhaustive empirical testing of the proposed methodology is performed and compared with other standard model like CART, Hoeffding based Linear Regression Model (ORTO) using solar energy data set. The obtained results show that the proposed LASH Tree significantly outperforms the existing approaches and it is proved that there is boosting of accuracy and usedless memory usage when compared with other algorithms.*

Keywords --*Hoeffding Tree, LASSO Regression, Prediction accuracy, Model adaptability.*

I. INTRODUCTION

In many real world situations data flows continuously and dynamically in the form of numerical data streams at high speed and huge volume which is a challenging research area to afford an efficient solution for building predictive model with limited resources [1].Incremental learning methods or On line learning methods are one of the approaches in handling streaming data by constructing the model sequentially using either one example at a time or mini batch at a time .Recently *there* is a growing interest in developing Regression Tree models due to it's high interpretability and accuracy. Moreover the function at the leaf node evaluates the target variable by diagnosing the correlation between predictor variable and target variable [2].The proposed LASSO Regression Hoeffding **T**ree (LASH Tree) algorithm is a MultiLinear Regression Tree model which incorporates LASSO Regression with Hoeffing Tree that produces highly interpretable and better insights of both linear and non linear relationship of the data. LASSO is the abbreviation of **L**east **A**bsolute **S**hrinkage and **S**election **O**perator is formulated by Robert Tibshirani a powerful method performs Feature selection and Regularisation and used to minimize the prediction error [3]. Hoeffding Tree is the most popular Incremental algorithm introduced by Domingos et al., who used the statistical inequality sample size of Hoeffding Bound (HB) and proved the

[1] *Research Scholar, PG and Research Department of Computer Science, Rajah Serfoji Government College, Thanjavur ,Tamil Nadu, India, Christy_se@pmu.edu*

[2] *Assistant Professor, PG and Research Department of Computer Science, Rajah Serfoji Government College, Thanjavur, Tamil Nadu, India, jgnanajayanthi@gmail.com*

DOI: 10.37200/IJPR/V24I4/PR201415
Received: 22 Sep 2019 | Revised: 13Oct 2019 | Accepted: 15 Jan 2020

induced Hoeffding Tree is close to the one produced by using entire stored data [4]. The concept of the LASH Tree is presented as the first phase of our proposed ensemble approach in an International Conference [5] and published in our previous paper.

The benefits of the proposed LASH tree:

- LASH Tree finds both linear and non linear relationship between target and predictor variable.
- It reduces error rate by constructing separate regression models at each leaf node using sub set of data stream instead of using entire batch of stream.
- It is highly interpretable as it produces decision rules which can be easy comprehendible by the analyst.
- It improves prediction accuracy by producing normal fitted model by reducing both Underfitting and Overfitting.
- It occupies less memory which in turn reduces cost complexity and time complexity.
- An exhaustive empirical testing of the proposed methodology was performed and compared with other Regression Tree models ORTO [6] and CART[7] algorithms using Solar Energy data set. The obtained results show that the proposed LASH Tree significantly outperforms the existing approaches and there is enriched accuracy and occupied less memory usage when compared with other algorithms. This memory and cost efficient LASH Tree can be used as base learner in ensemble algorithms.This paper is organised as follows : Section 2 describes the related work similar to linear regression tree. In section 3 the proposed LASH Tree is elaborated with it's algorithm and flow chart. Section 4 describes empirical study with comparative study of other algorithms and the results obtained are discussed. Section 5 ends with conclusion.

II. RELATED WORK

Domingos and Hulten proposed Hoeffding Tree (HT) or Very Fast Decision Tree (VFDT) using limited memory and constant time for streaming data [5]. Many researchers proposed a series of modifications to improve the predictive performance and to overcome the drawbacks of Hoeffding Tree algorithm. ORTO [6] is On line Regression tree with Options proposed by Elena Iknomovska which includes option nodes in addition to the ordinary split nodes to remove the need for selecting best attributes in the traditional Hoeffding Tree. Breiman et al proposed CART [7],a combination of classification and regression to deal with numerical and continuous values which has the problem of overfitting and instable in nature. The author used constant function in the leaves to find the prediction value.

Yi-Fei Cai proposed Tree Lasso technique to classify images where the pixels are lying on a tree to obtain stable feature sets to develop health care predictive models [8].Ricardo Pio Monti proposed a framework [9] to infer an adaptive regularization parameter to solve the problem of L1 regularization linear models using streaming data. Feihan Lua proposed the Imputed-LASSO [10] by combining Random Forest imputation and LASSO an efficient item selection approach for missing data. KaiCHEN proposed Lasso Bagging ensemble algorithm [11] to improve the learning ability, by choosing ensembles of trees based on the shrinkage estimation of lasso technology. Sanjiban Sekhar Roy proposed LASSO method based on a linear regression model[12] to predict financial market behaviour.

III. PROPOSED METHODOLOGY

A novel and memory efficient LASH Tree is proposed by incorporating Hoeffding tree and LASSO Regression to produce highly interpretable and better insights which finds both linear and non linear relationship of the data. The proposed LASH Tree is a top down regression tree, reads the batch of data at the root node and stores the necessary statistical values in the leaf node.

3.1SPLITTING CRITERIA USING LASSO

3.1.1 Feature Selection using Shrinkage and Selection operation

In the proposed approach the best predictor variable is selected using *Least Absolute Shrinkage and Selection Operator*(LASSO), instead of Information Gain or Gini Index used in the existing approaches. LASSOis an extension of multiple linear regression. If a set of N independent variables of the form[(X_1, X_2,…X_N) , Y_i] is given , where Y_i is the numeric dependent outcome , Xi is the discrete or continuous predictor variable vary from 1 to N, target variable or prediction is obtained by the following Multiple Linear Regression Equation

$$Y_i = \beta 0 + \beta 1 X_1 + \beta 2 X_2 + \cdots \beta k X_k + \Phi \qquad \text{------------- (1)}$$

Equation (1) can be simplified as

$$Y= \beta 0 + \sum \beta_i X_i (i=1 \text{ to } k) + \Phi \qquad \text{------------ (2)}$$

- Y_i is the dependent target variable
- $X_i$is the predictor variable, correlates with the target variable Y_i
- β0 is the coefficient value which represents the model intercept
- βi is the coefficient value represents the model slope, that gives the information about the positive or negative correlation with the target variable
- Φ is the error term that involves variability

LASSO operation is denoted by ,

$$\text{Minimize } (\beta_0, \beta) \quad \left\{\frac{1}{2N}\sum_{i=1}^{N}(Yi - \beta 0 - \sum_{j=1}^{p} Xij\ \beta j)^2 + \lambda\sum_{j=1}^{p}|\beta j|\right\} \quad \text{------------(3)}$$

N is the number of features ,λ is the tuning parameter to fix the penalty value . Fig 1. Shows the algorithm and flow chart for the feature selection using LASSO regression.

3.1.2 L1 Regularisation

The tuning parameter λ value in equation –(3) is used to control the regularisation term. Smaller the value of λreleases the variables from under fitting and makes the model more closely to the training data. On the contrary, larger values of λ restricts the variables from overfitting to fit the data less closely to the training data. Hence an intermediate value of λ strikes a good balance between these two extremes, that produces the most accurate model with some L1 Regularisation incoefficients equal to zero and minimize the weightage of the remaining coefficients.

3.1.3 Assessing the fitness of Regression Model

Assessing the fitness of Regression model is essential to evaluate the fitness of the model with the new data and Root Mean Square Error (RMSE) is one of the metric used to evaluate the fitness performance. RMSE is defined as the square root of the variance of the residuals.Lower the values of RMSE indicates better fit and high RMSE value indicates lower fit.

DOI: 10.37200/IJPR/V24I4/PR201415
Received: 22 Sep 2019 | Revised: 13Oct 2019 | Accepted: 15 Jan 2020

RMSE = $\frac{1}{n}\sum_{i=1}^{n}(\overline{X}i - Xi)^2$ --------------- (4)

n – Number of observations $\overline{X}i$ - Predicted value X_i -- Observed value

3.2 Tree Growing and Assigning LASSO Regression function at the leaf node

3.2.1 Sample Size

After selecting the significant correlated variables using LASSO, LASH Tree is constructed through binary recursive partitioning, that splits the data set into partitions and continues to split each partition into smaller groups. More samples are observed at the root node until the difference between the $(i)^{th}$ best predictor variable and the $(i+1)^{th}$ best predictor variable is greater than Hoeffding inequality Bound ε , here $\varepsilon = \sqrt{\frac{R^{2}\ln(1/\delta)}{2n}}$ ---------

Algorithm 1: Pseudo code for Splitting criteria using LASSO

Inputs
X_i : Independent Predictor variables
Yi : Dependent Target variable
Λ : Tuning parameter or Penalty Value
Output : Set of Best Correlated or Predictor Variables

1. Let $(X_1, X_2, X_3 \ldots X_i)$ be the Independent Predictor Variables, where i= 1..k and let Y_i be the Dependent Target Variable in the data stream which is arriving at the root node of the Hoeffding Tree.
2. Update the required statistics at leaf node. (Initially root node)
3. Adjust the penalty value (Λ) from 0 to ∞ in the LASSO Regression equation
4. Compute RMSE = $\frac{1}{n}\sum_{i=1}^{n}(\overline{X}i - Xi)^2$ for each penalty value
5. Compute the minimised coefficient value for each non zero variable
6. Eliminate the variables which are shrunk to zero
7. Identify the optimal Λ value which has minimum RMSE using k fold – cross validation
8. Fix the highly correlated variable for the root node splitting and the remaining variables for the inner nodes
9. Update the required statistics at the leaf node.

----- (5)

R is the range of n independent variable ,

1-δ is the confidence level , n is the number of observations.

3.2.2 Splitting Value

DOI: 10.37200/IJPR/V24I4/PR201415
Received: 22 Sep 2019 | Revised: 13Oct 2019 | Accepted: 15 Jan 2020

Splitting value or cut point value is a threshold value with minimum and maximum to split the data set . It is measured by evaluating mean value of the attribute (Xi) and split the data sub set into Left and Right sub node based on CutPoint value. After splitting, the current node is denoted as $(X_i)^P$ and it's left descendent node is denoted as $(X_{i+1})^L$ and the right descendant node by $(X_{i+1})^R$ based on the result of the splitting criteria.

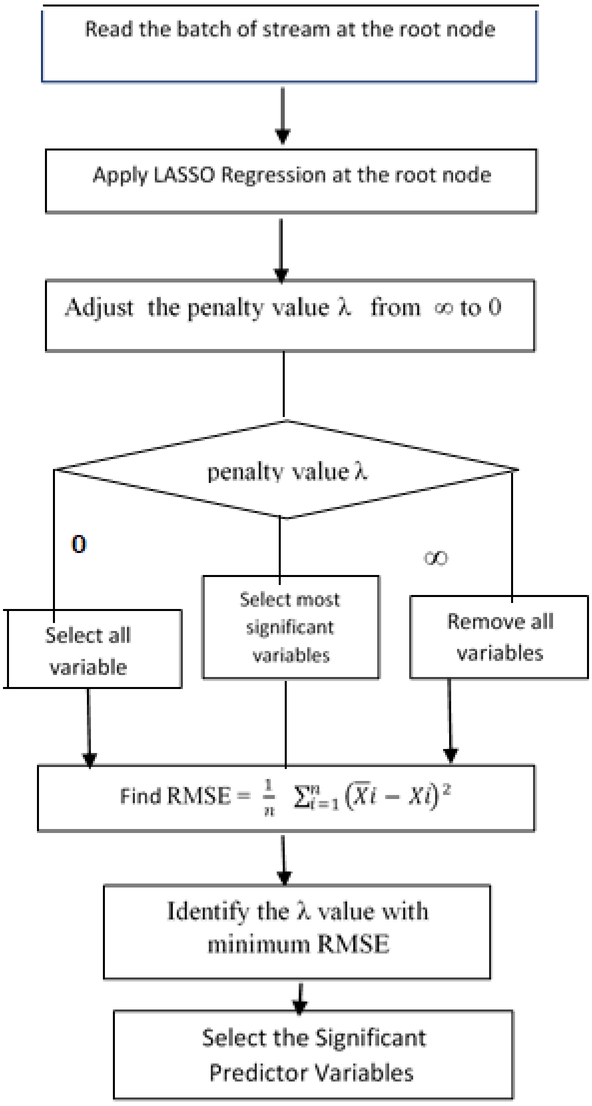


Figure 1: Feature Selection using LASSO Regression

3.2.3 Prediction Strategy

Each leaf node of the LASH Tree is assigned with LASSO Regression Equation using equation (1) in order to predict the new instance. When a new sample is passed down from the root to a leaf based on it's attribute value criteria through every internal node and it's predicted value is evaluated based on the LASSO Regression equation constructed at the leaf node. The proposed LASH Tree construction is depicted in Fig.2 and its corresponding algorithm and flow chart is shown in Fig 3.

DOI: 10.37200/IJPR/V24I4/PR201415
Received: 22 Sep 2019 | Revised: 13Oct 2019 | Accepted: 15 Jan 2020

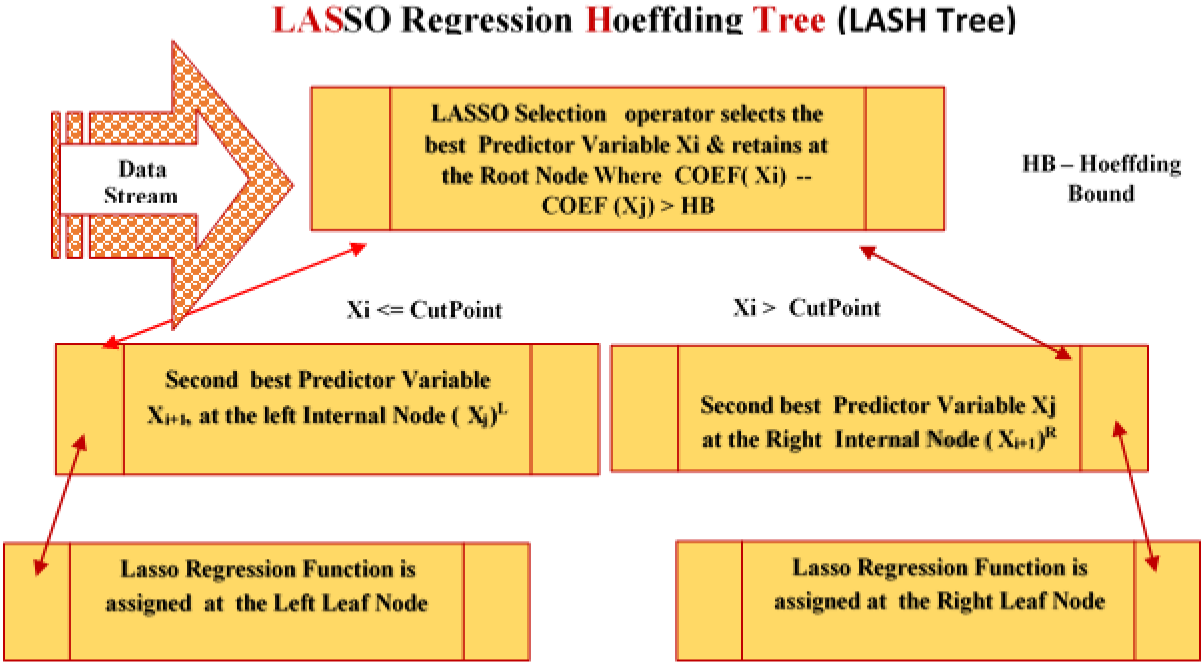


Figure 2: LASSO Regression Hoeffding Tree (LASH Tree) construction

IV. EMPIRICAL EVALUATION

4.1 Experimental Set up

The proposed work is implemented using R analytics package with the environment of Windows 10 PC , Intel Quad 2.8GHz CPU and 8G RAM. The available memory is set to evaluate 50000 instances of RAM to scan the batch data once.

4.2 Data Set used

In this paper we have conducted the experiments with Solar energy data set of 3,13,914 instances contains 13 attributes with no missing and repetition values. The data set has been collected from Desert Knowledge Australia Solar Centre (DKASC) during the period of 2016 to 2018 and preprocessed. Desert Knowledge Australia (DKA) is a National Organization committed to building harmony ,sustainability and prosperity for all Australian desert people [13] The proposed algorithm predicts the Active Energy Delivered-Received (kWh) by learning with the existing data samples.

4.3 Experimental Results:

4.3.1 Feature selection and shrinkage operation of LASSO Regression

The initial 50000 samples of the solar energy data set is applied with LASSO Regression in order to select the best significant variables. Fig 4shows LASSO implementation in which Log (Lamda) is in X axis and Coefficient of the variables at the Y axis. The variables above 0 coefficient value are positively correlated with the target variable and the variables below 0 coefficient are negatively correlated with the target variable. The graph shows that larger the penalty value of λ, restricts the variable in fitting the data closely, leads to under fitting and smaller the value of λ leads to overfitting

DOI: 10.37200/IJPR/V24I4/PR201415
Received: 22 Sep 2019 | Revised: 13Oct 2019 | Accepted: 15 Jan 2020

Algorithm 2: LASH Tree Construction

Inputs

X_i : ith best correlated variable
X_{i+1} : i+1 th best correlated variable
Yi : Dependent Numeric Target variable
ε : Hoeffding Bound where = $\sqrt{\frac{R^{2}\ln(1/\delta)}{2n}}$
CutPoint : Mean(Xi)

Output :Lasso Regression Hoeffding Tree (LASH Tree)

1. Get the i^{th} and $i+1^{th}$ best correlated variables and it's coefficient using Algorithm 1
2. Let |Δ COEF| = COEF(X_i) - COEF(X_{i+1})`
3. Increase the number of examples until (ΔCOEF > ε)
4. If (Δ COEF > ε) Split the dataset using CutPoint(Xi) and replace the root node by Regression Node. Let it be $(X_i)^P$
5. Let the two descendant sub nodes $(X_{i+1})^L$ and $(X_{i+1})^R$ be the left child of $(X_i)^P$ and right child of $(X_i)^P$
6. Get the minimized coefficients of the non zero predictor variables and eliminate all zero coefficient variables
7. Assign LASSO Regression Function ( ) at the leaf node using multi linear regression equation, $Y = \beta 0 + \beta 1 X_1 + \beta 2 X_2 + \cdots \beta k X_k + \Phi$ Stop splitting when there is no node for splitting

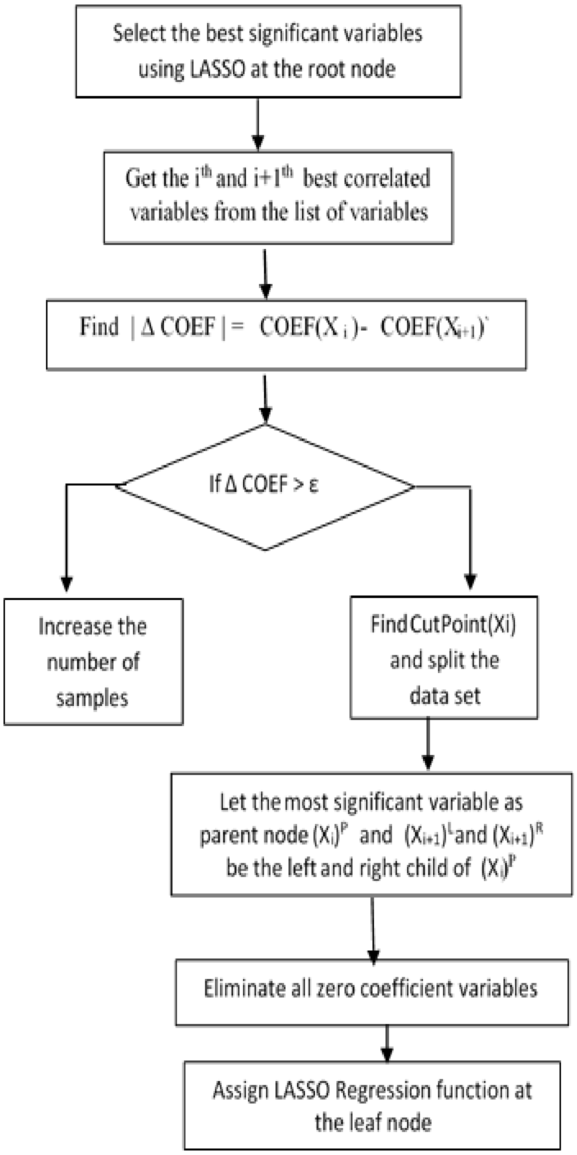
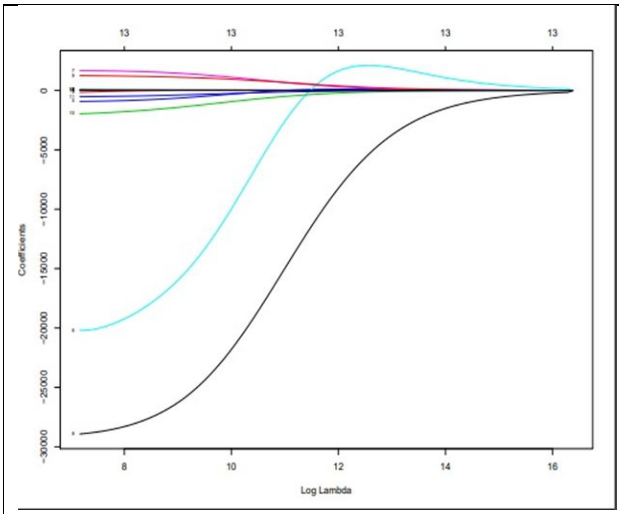


Figure 3: Algorithm and Flow chart for LASH Tree construction



No	Predictor Variable Name	Coefficient Value
1	Current Phase Average (A)	1.891133e+09
2	Active Power (kW)	- 2.901335e+03
3	Power Factor Signed (%)	- 1.352862e+02
4	Average Voltage Line-to-Neutral (V)	- 7.233713e+02

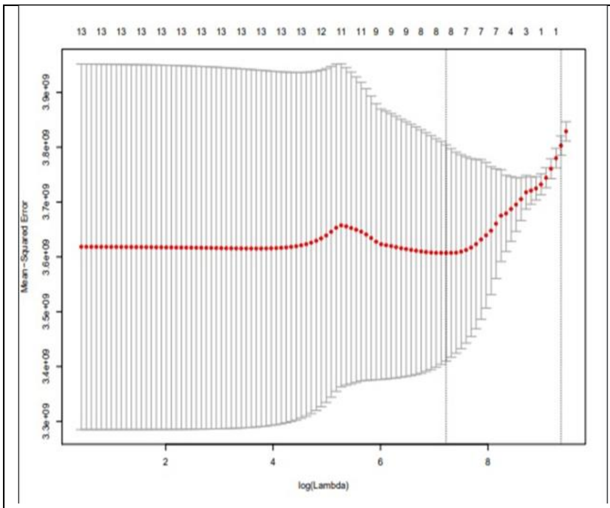
DOI: 10.37200/IJPR/V24I4/PR201415
Received: 22 Sep 2019 | Revised: 13Oct 2019 | Accepted: 15 Jan 2020

5	Frequency (Hz)	7.941301e+04
6	THD Current Average (%)	1.267598e+03
7	THD Voltage Average (%)	- 2.743622e+04
8	Wind Speed (m/s)	9.836408e+2
9	Weather Temperature Celsius (Â°C)	- 2.247667e+03
10	Weather Relative Humidity (%)	- 5.508218e+02
11	Global Horizontal Radiation (W/mÂ^2)	7.301958e+1
12	Diffuse Horizontal Radiation (W/mÂ^2)	- 9.214214e+01
13	Wind Direction (Degrees)	1.168052e+01

Figure 4: Co efficient of the Predictor variable

4.3.2 L1 Regularisation using cross validation

L1 Regularisation is implemented in the next stage to chooses the right choice of λ. value to improve prediction accuracy which is implemented using k fold - cross validation.



No	Predictor Variable Name	Coefficient Value
1	Current Phase Average (A)	0
2	Active Power (kW)	0
3	Power Factor Signed (%)	0
4	Average Voltage Line-to-Neutral (V)	0
5	Frequency (Hz)	2.942801e+04

DOI: 10.37200/IJPR/V24I4/PR201415
Received: 22 Sep 2019 | Revised: 13Oct 2019 | Accepted: 15 Jan 2020

6	THD Current Average (%)	0.168598e+03
7	THD Voltage Average (%)	- 1.123622e+04
8	Wind Speed (m/s)	4.8231408e+2
9	Weather Temperature Celsius (Â°C)	- 1.297667e+03
10	Weather Relative Humidity (%)	- 1503218e+02
11	Global Horizontal Radiation (W/mÂ²)	4.201758e+1
12	Diffuse Horizontal Radiation (W/mÂ²)	0
13	Wind Direction (Degrees)	1.168052e+01

Figure 5: Cross Validation error curve

Fig 5. shows cross validation error curve obtained using L1 Regularisation. From the graph it is found that the model that minimises error value with 8 significant variables and other variables are eliminated from the model.

4.3.3 Building LASH Tree using the most significant variables obtained from LASSO

LASH tree is constructed using eight significant attributes selected from LASSO Regression. The most significant attribute is selected as the root node which splits the data sets into subsets and recursively replaces leaf node by test nodes. The proposed LASH Tree will be efficient, only if it is able to build more accurate trees and should have the ability beyond the conventional system. The performance of LASH Tree is tested against CART and ORTO.

4.3.4 Accuracy of LASH Tree

Fig 6. shows the comparison between the accuracy of the proposed LASH Tree on the solar energy data set and the other Regression Tree algorithms CART and ORTO. From the graph it is found that RMSE value for the proposed algorithm is less compared to others which shows it has more accuracy than others. It is found the

DOI: 10.37200/IJPR/V24I4/PR201415
Received: 22 Sep 2019 | Revised: 13Oct 2019 | Accepted: 15 Jan 2020

reason behind more accuracy is due to the reduction of overfitting and under fitting issue in LASH Tree makes it more accurate whereas the traditional OTTO and CART algorithm produces less accurate results due to overfitting issue.

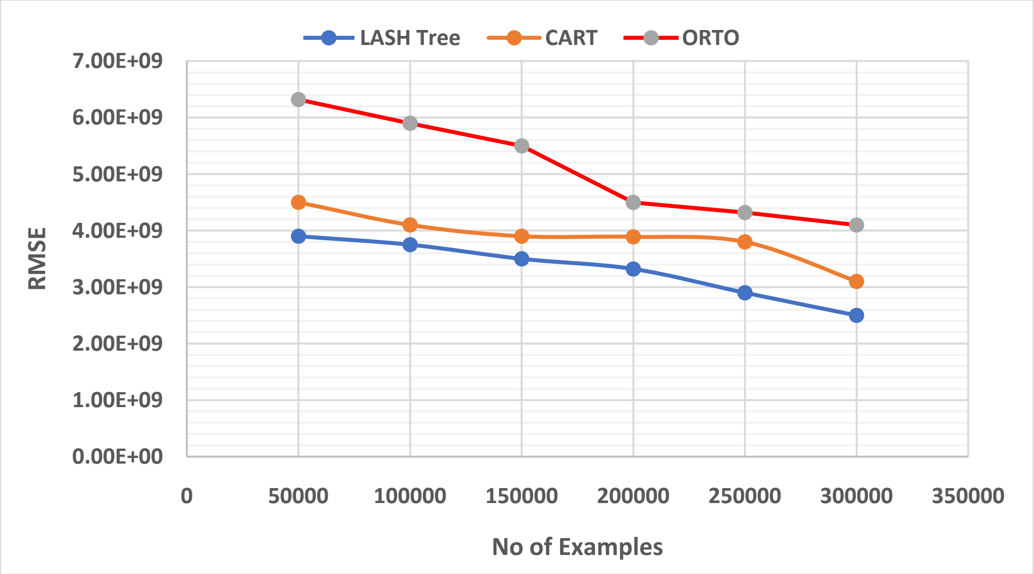


Figure 6: Comparison of Model Accuracy

4.3.5 Number of nodes& Learning Time

Fig 7 shows number of nodes induced at each induced learner . From the graph it is known that the number of nodes induced by LASH tree is lesser than the remaining two learners. As the number of leaf nodes in the tree is directionally proportional to the size of the tree, it is proved that the proposed LASH Tree consumes less memory.

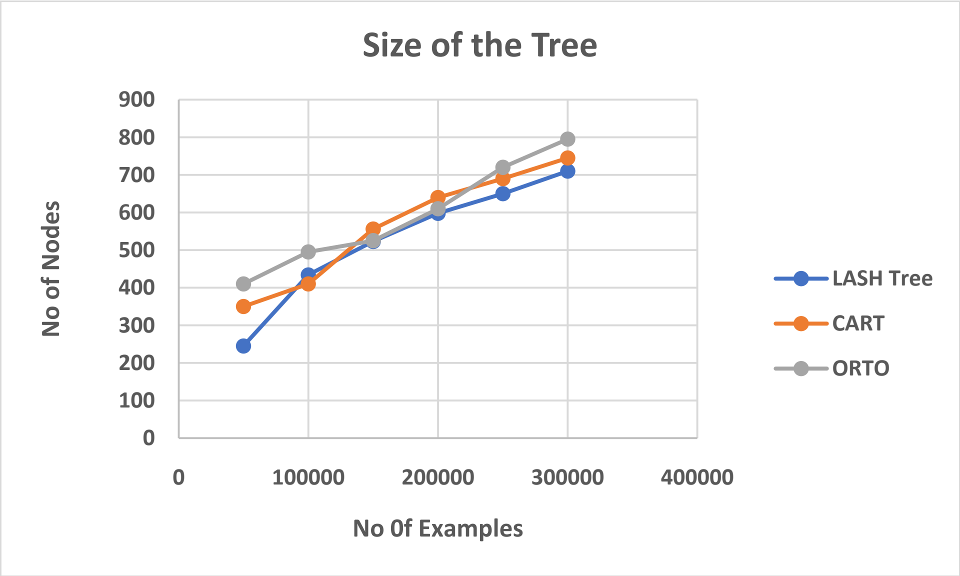


Figure 7: Comparison of size of the tree

V. CONCLUSION

DOI: 10.37200/IJPR/V24I4/PR201415
Received: 22 Sep 2019 | Revised: 13Oct 2019 | Accepted: 15 Jan 2020

In this paper, we have developed algorithms for the implementation of lASH Tree with solar energy data set. Our proposed **LAS**SO Regression Hoeffding**T**ree (LASH Tree) produces better predictions and better insights. It is also compared with other standard model like CART, Hoeffding based Linear Regression Model (ORTO) and proved that the proposed LASH Tree significantly outperforms the existing approaches and it is shown there is improved in accuracy and used less memory usage when compared with other algorithms. The proposed LASH tree can be used as base learner in ensemble approach to improve the prediction accuracy. In future, it is planned to include classification model in the existing LASH Tree to support with both Regression and Classification problems.

REFERENCES

1. Joao Gama , Raquel Sebastiao,” On evaluating Stream Learning Algorithms” in Mach Learn 90: 317 -346 Springer Publication [2013]
2. Lizhe Zun , Yangzi Guo, Adrian Barbo,” A Novel Framework for Online Supervised Learning with Feature Selection” in rXiv:1803.11521v4 [stat.ML] Dec [2018]
3. Robert Tibshirani ,”Regression Shrinkage and selection via lasso” January Journal of Royal Statistics Society, B Series, [1995]
4. Pedro Domingos, Geoff Hulten, “ Mining High-Speed Data Streams”, in KDD 2000, Boston, MA USA © ACM 1-58113-233-6/00/08 [2000]
5. D.Christy Sujatha , Dr,J.Gna Jayanthi, “Meta_LASH Tree : Bagging at Meta Level Using LASSO Regression Hoeffding Tree for Streaming Data” in the third International Conference on Trends in Electronics and Informatics (ICOEI 2019) IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8 ,2019 IEEE.
6. Elena Ikonomovska , Jo˜ao Gama , Bernard ˇZenko “Speeding Up Hoeffding-Based Regression Trees with Options” in Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA, USA,[2011]
7. L. Breiman, J.Friedman, R.Olshen,C.Stone, “Classification and Regression Trees” in Journal of Engineering, Chapman and Hall, New York,[1993]
8. Iman Kamkar ,”Stable feature selection for clinical prediction: exploiting ICD Tree stricture using Tree Lasso “, in Journal of Bio medical Informatics 53, 277-290[2015]
9. Ricardo Pio Monti, Christoforos Anagnostopoulos, and Giovanni Montana “Adaptive regularization for Lasso models in the context of non-stationary data streams” in arXiv:1610.09127v2 [stat.ML] 14 December 2017
10. Feihan Lua and Eva Petkova “A comparative study of variable selection methods in the context of developing psychiatric screening instruments” in Statistics in Medicine , Research Article [2013]
11. Kai Chen ,Yang Chin “An Ensemble Learning Algorithm Based on Lasso Selection” in IEEE conference [2010].
12. Sanjiban Sekar Roy ,Avik Basu “Stock Market Forecasting using LASSO Linear regression model”in Afro-European Conf. for Ind. Advancement, an Advances in Intelligent Systems and Computing , DOI: 10.1007/978-3-319-13572-4_31.Springer International Publication Switzerland [2015]
13. http://dkasolarcentre.com.auhttp://dkasolarcentre.com.au

DOI: 10.37200/IJPR/V24I4/PR201415
Received: 22 Sep 2019 | Revised: 13Oct 2019 | Accepted: 15 Jan 2020

14. Hewageegana h. G. S. P, arawwawala l. D. A. M. , ariyawansa h. A. S, tissera m. H. A, dammaratana i. (2016) a review of skin diseases depicted in sanskrit original texts with special reference to ksudra kushtha. Journal of Critical Reviews, 3 (3), 68-73.
15. Dr.Sundararaju,K., & Rajesh,T. (2016). Control Analysis of Statcom under Power System Faults. *International Journal of Communication and Computer Technologies,* 4(1), 46-50.
16. Ban Maheskumar N., &Prof.Sayed Akhtar, H. (2016). An online and offline Character Recognition Using Image Processing Methods-A Survey. *International Journal of Communication and Computer Technologies,* 4(2), 102-107.
17. Peter, S. An analytical study on early diagnosis and classification of Diabetes Mellitus (2014) Bonfring International Journal on Data Mining, 4 (2), pp. 7-11.
18. Murthy, N.H.S., Meenakshi, M. ANN model to predict coronary heart disease based on risk factors (2013) Bonfring Int. J. Man Mach Interface, 3 (2), pp. 13-18.