Decision Tree Regression Algorithm for Mining High-Speed Data Streams

N. SIVAKUMAR[1] S. Anbu[2]

[1]Research Scholar, [2] Professor

[1,2] Department of Computer Science and Engineering,

[1,2] St. Peter's University, Avadi, Chennai, Tamilnadu, India.

***Abstract*— In recent years, advances in hardware technology have facilitated new ways of collecting high speed data continuously. Marvelous and possibly Infinite volumes of data streams are often generated by communication networks, remote sensors and other dynamic environments. Mining these data streams raises new problems in terms of how to mine continuous high speed data items that you can only have one look at. In this paper, we propose algorithm output is a solution for mining high speed data streams. The decision tree algorithm builds regression models in the form of a tree structure. The tree breaks down into smaller subsets while at the same time its associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches each representing values for the attribute tested. Leaf node represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both unconditional and numerical data.**

Keywords—Decision tree regression, Mining Decision tree regression, Regression Tree,

I. INTRODUCTION

Many organizations today produce an electronic record of essentially every transaction they are involved in. These results may in tens or hundreds of millions of records being produced every day. For example Google handles 300 million searches, and Telecomm records produces 400 million call records. Scientific data collections, satellites or an astronomical observation routinely produces gigabytes of data per day.

Data rates of this level have significant consequences for data mining. A few months' worth data can easily add up to billions of records, and the entire history of transactions or observations can be in hundreds of billions.

1a). Data Mining

The field of data mining and knowledge discovery is emerging as a new, fundamental research area with important applications to science, engineering, medicine, business and education.

The data mining process is often characterized as a multi stage iterative process involving data selection, data cleaning and application of data mining algorithms, evaluation and so forth.

1b). Decision Tree

A decision tree is a flowchart like a structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test and each leaf node represents a class label. The path from root node to leaf node represents regression rules.

The decision tree can be one dimensional into **decision rules**, where the outcome is the contents of the leaf node, and the conditions along the path form a conjunction in the if clause. In general, the rules have the form: *if* condition1 & condition2 & condition3 *then* outcome. The advantages of decision tree is given below

- **Easy to understand and interpret:** People are able to understand decision tree models
- **Easy to data preparation:** Other techniques often require data normalization, redundancy and blank can be removed.
- **Able to handle numerical & categorical data:** Numerical values and value of variables could be used
- **Able to validate a model using statistical tests:** That makes it possible to account for the reliability of the model.
- **Robust:** Performs well even if its assumptions are somewhat violated by the true model from which the data were generated.
- **Performance with large datasets:** Large amounts of data can be analyzed using standard computing resources in reasonable time

1c.Decision tree regression

Regression is a data mining function that predicts a value. For example Profit & Loss in sales, mortgage rates, house values, temperature, or distance could all be predicted using regression techniques.

In the model build process, a regression algorithm estimates the value of the target as a function of the predictors for each case in the build data. These relationships between predictors and target are summarized in a model, which can then be applied to a different data set in which the target values are unknown.

Regression models are tested by computing various statistics that measure the difference between the predicted values and the expected values. The historical data for a regression project is typically divided into two data sets: 1. building the model & 2.Testing the model.

II. BACKGROUND STUDY

Hoeffding Tree Algorithm [1] proposed by from Domingos and Hulten's gives the streaming decision tree induction which is called Hoeffiding tree. This gives certain level of confidence on the best attribute to split the tree; hence we can build the model based on certain number of instances. The advantage is high accuracy with small sample but the drawback is performs multiple scans of the same data and it is incremental

Very Fast Decision Tree (VFDT)[2] by Domingos improves the speed and accuracy. It splits the tree using he current best attributes. Although works with high speed data streams, it still cannot handle concept drift in drift in data streams. To adapt concept drift in data streams, VFDT algorithm was further developed into the concept drift-adapting Very Fast Decision Tree(CVFDT) it runs over fixed sliding windows in order to have the most updated classifier.

Ultra – Fast Forest Trees – UFFT

UFFT[3] is an algorithm for supervised classification learning that generates a forest of binary tree. The algorithm is incremental, processing each example in constant time. It is designed for continuous data. It uses analytical techniques to choose the splitting criteria. Merit is, it shows good results with several large and medium size problems. Demerit is the algorithm forgets the sufficient statistics and learns the new concept with only the examples in the new concept.

III. OUR PROPOSED APPROACH

In this paper, a sequence of data mining procedures are applied continuously for design and develop a better approach for mining high speed data streams

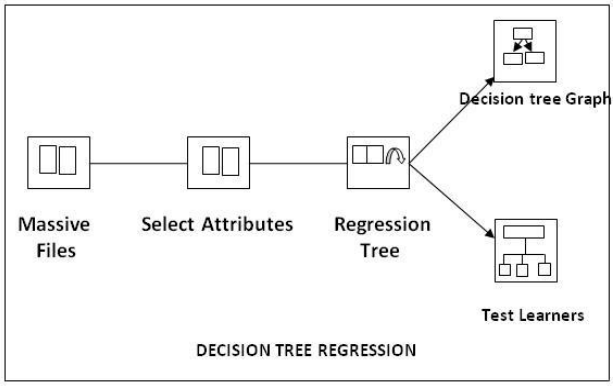

Fig. 1: Proposed model

3a). Massive Files

A regression task begins with a data set in which the target values are known. The dataset is then split on the different attributes. Such data sets are applications in trend analysis, business planning, marketing, financial forecasting, time series prediction, biomedical and drug response modeling, and environmental modeling.

3b). Regression Tree process

Regression analysis seeks to determine the values of parameters for a function that cause the function to best fit a set of data observations that you provided. The following equation expresses these relationships in symbols. It shows that regression is the process of estimating the value of a continuous target (y) as a function (F) of one or more predictors (x_1 , x_2 , ..., x_n), a set of parameters (θ_1 , θ_2 , ..., θ_n), and a measure of error (e). Then the equation as

$$\boldsymbol{Y = F(x, 0) + e} \qquad \textbf{(1)}$$

For Example: Given data on _m' explanatory variables and one explained variable, where explained variable can take **real** values in $\Re^1$, to find a function that gives the —best it fit as

$$\boldsymbol{Given}: (x_1, y_1), (x_2, y_2), \ldots \ldots (x_m, y_m) \in + \Re^n \mathbf{X}\, \Re^1$$

$$\boldsymbol{Find}: f := \Re^n \mathbf{X}\, \Re^1$$

We find "best function " = the expected error on *unseen* data and **(x_1, y_1), (x_2 , y_2),…, (x_m , y_m)** are minimal.

There are different families of regression functions and different ways of measuring the errors. The different types of regressions are given below

3b)1. Linear Regression

A linear regression is a technique which can be used as the relationship between the predictors and the target. This can be approximated with a straight line.

A single predictor regression is the easiest to visualize and it is shown in the following figure

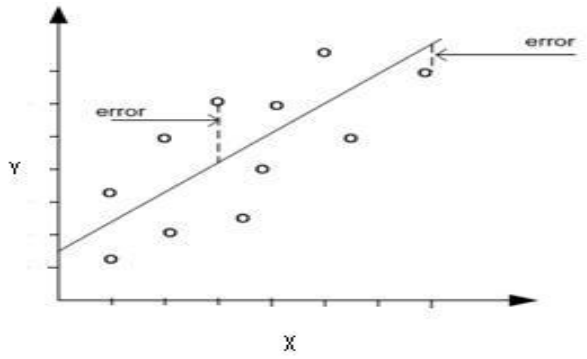

Fig. 2: Single Predictor Linear Regression

The Figure 2 can be expressed with the following equation.

$$y = \emptyset 2x + \emptyset 1 + e \qquad (2)$$

The regression parameters in (2) as follows ($\emptyset_2$) Slope line — Represents the angle between a data point and the regression line ($\emptyset_1$) y intercept — Represents the point where x crosses the y axis (x = 0)

3b)2. Nonlinear Regression

The nonlinear regression expresses the relationship between x and y and these cannot be approximated with a straight line. In this case, a nonlinear regression technique may be used. Alternatively, the data could be pre-processed to make the relationship linear. This model define y as a function of x using an equation that is more complicated than the linear regression equation. In the Figure 3, the x and y have a nonlinear relationship.

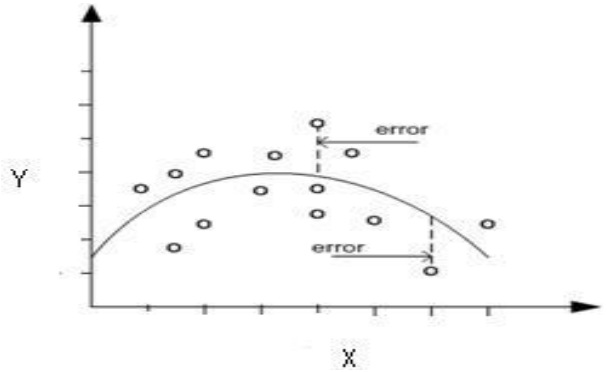

Fig. 3: Single Predictor with non linear regression

3c). Multiple Linear Regressions

The term multiple linear regressions refer to linear regression with two or more predictors. In this case, we examine the process of multiple linear regressions. In this context we give the significant to split the data into 2 categories: 1. the training data set and 2. Validation data set. These are being able to validate the multiple linear regression models. After this examine, we know that the method for identifying subsets of the independent variables to improve predictions.

There is a continuous random variable called the dependent variable, ***Y***, and a number of independent variables $\mathbf{x_1, x_2, .. x_p}$ Our purpose is to predict the value of the dependent variable or response variable using a linear function of the independent variables. The values of the independent variables or predictor variables are known quantities for purposes of prediction. The model is given as the following equation.

$$\boldsymbol{Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \ldots + \beta_p x_p + \varepsilon,} \qquad (3)$$

Where

$\boldsymbol{\varepsilon}$ – Random variable with mean and equal to 0.
$\boldsymbol{\beta_0, \beta_1, \ldots\ldots \beta_p}$ a Co – efficient with unknown values.
p – We estimate (p+2) unknown values from variables **n**- the number of data consists
n give us the values $\boldsymbol{y_i, x_{i1}, x_{i2}, \ldots, x_{ip}; i = 1, 2, \ldots, n}$

The estimates for the β coefficients are computed so as to minimize the sum of squares of differences between the predicted values at the observed values in the data. The sum of squared differences is given by the formula.

$$\boldsymbol{Y_i = \beta_0 - \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \ldots - \beta_p x_{ip})^2} \qquad (4)$$

We denote the values of the coefficients that minimize this expression by $\hat{\beta}\mathbf{0}, \hat{\beta}\mathbf{1}, \hat{\beta}\mathbf{2}, \ldots, \hat{\beta}\boldsymbol{p}$. These are our estimates for the unknown values and are called Ordinary Least Squares

Once we have computed the estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ we can calculate an unbiased estimate $\hat{\sigma}\mathbf{2}$ for $\boldsymbol{\sigma 2}$ using the formula:

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n} + (y_i - \hat{\beta}_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip})^2$$

(5)

We plug in the values of $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots \hat{\beta}_p$ in the linear regression model to predict the value of the dependent value from known values of the independent values, $x_1, x_2, \ldots, x_p$. The predicted value, $\hat{Y}$, is computed from the equation

$$\hat{Y} = \hat{\beta}_0 - \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots \ldots - \hat{\beta}_p x_p)^2 \quad (6)$$

This equation is the best predictions possible in the sense that they will be unbiased and will have the smallest expected squared error compared to any unbiased estimates.

IV. ERROR REDUCTION PROCESS

In this approach, the minimum error pruning has been proposed. The most common method for building a regression model based on a sample of an unknown regression surface consists of trying to obtain the model parameters that minimize the least squares error criterion.

The following formula helps to calculate the error in the regression model

$$\mathbf{Err(t)} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{y_i} - r(\boldsymbol{\beta}, \mathbf{x_i}))^2 \quad (7)$$

Where

t Set of leaves on tree
n is the sample size
(x_i,y_i) is the data point
$r(\beta,x_i)$ is the prediction of the regression model $r(\beta,x_i)$ for (x_i,y_i)

For Example the following figure 4 shows that the two regression examples. You can see that in Graph A1, the points are closer to the line than they are in Graph A2. Therefore, the predictions in Graph A1 are more accurate than in Graph A2

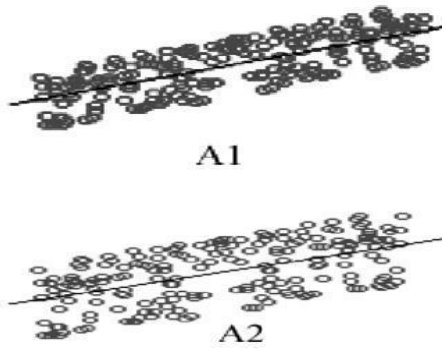

Fig. 4: Variations of two regressions

4a). Estimation of Accuracy in Regression

In the regression problem there are 2 estimates of the accuracy are used: 1. Estimate of Re substitution, 2. Test sample estimate. These estimates are defined here.

1. Estimate of Re substitution**:** It is the estimate of the expected squared error using the predictor of the continuous dependent variable. This can be computed as the formula 4.4. Where the learning sample is (x_i,y_i),i = 1,2,...,N and d is the predictor

$$\mathbf{R}(d) = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{y_i} - \mathbf{d}(\mathbf{x_i}))^2 \quad (8)$$

2 Estimate test sample**.** The total number of cases is divided into two subsamples. The test sample estimate of the mean squared error is computed in the formula 4.5. Where N is subsample size of x_i and N_2 is subsample size of y_i:

$$\mathbf{R^{t5}}(d) = \frac{1}{\mathbf{N_2}} \sum_{(x_1,y_1)=Z_i}^{n} (\mathbf{y_i} - \mathbf{d}(\mathbf{x_i}))^2 \quad (9)$$

V. DECISION TREE REGRESSION (DTR) ALGORITHM

Input: Load the High speed data sets
Output:

Step 1: Parse the loaded data sets

Step 2: Split the data sets as

1. Training data sets
2. Test data sets

Step 3: Create a tree node and initialize

MaxDepth=100

MaxBins=30

Step 4: Train the Decision tree regression model

Step 5: Evaluate the model & Compute Error

Predict = Training_data.map

Predict= Test_data.map

Calculate the error using formula (7)

Step 6: Repeat Step 3 to step 5 Until final Node

Step 7: Print: _Learned Regression Model‘

Step 8: Print: _Error Report‘

Step 9: Save the model

The above algorithm performs regression using a decision tree with variance as an impurity measure and a maximum tree depth of 100. The error is computed at the end of evaluation to fit the goodness.

VI. SIMULATION & RESULTS

The algorithm is written in MATLAB-R2014B software with higher configuration system. The mining high speed data is taken for experiment which is available in 'www.archive.ics.uci.edu/ml/ machine-learning-databases/ housing/ housing.data', filename)'. Various numbers of transactions are generated and the results are stored.

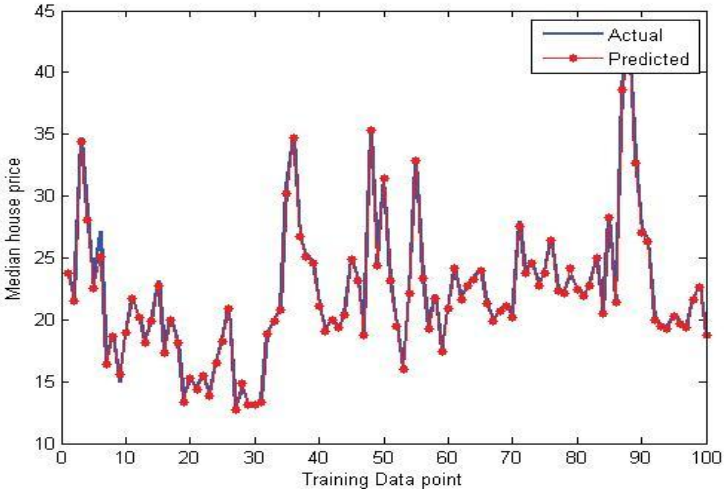

Fig. 5: Plot the training data

Figure 5 shows that the training data sets with 100 points. There are the process of building the dataset is completely automated. First the framework randomly generates multiple values for each independent variable. For each independent variable, the user defines a number of generated values.

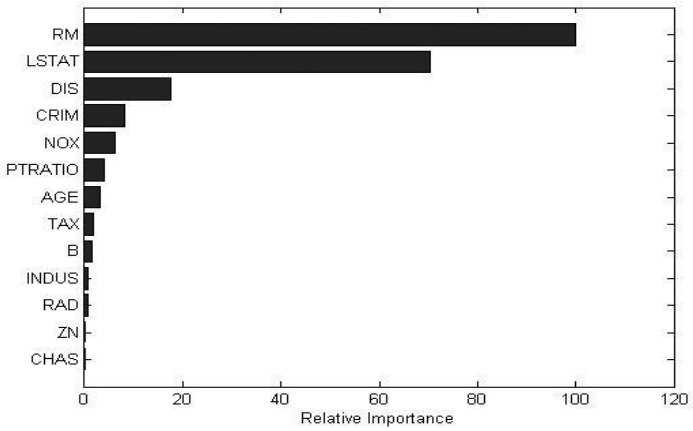

Fig. 6: Plot the sorted predictors

Figure 6 shows that the sorted predictors. The predictive model output is used to analyze current data and historical facts in order to better understand to identify the risk and opportunities for an organization. It uses a number of decision tree regression techniques and machine learning to help analysts make future predictions.

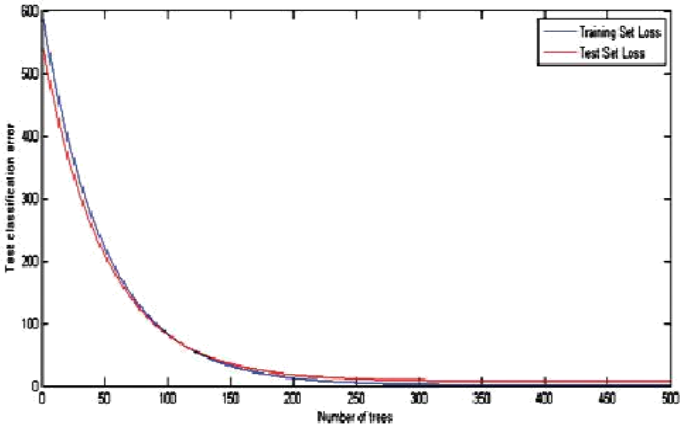

Fig. 7: Plot the error

Figure 7 shows that the calculated the Standard Error of the Estimate of the Pearson's product correlation coefficient P and the difference between the means of the predicted and the measured amounts of re-transmitted data. The error rate is reduced to 0.004%. This indicates strong positive linear correlation between the predicted and the measured values.

VII. CONCLUSION

In this paper we proposed DTR and analyzed to solve mining high speed data streams. At the first we presented a decision tree regression concepts based on data mining applications. In this algorithm, we classify various types of regressions to train and test the tree datasets. The DTR algorithm evaluated efficiently. Later we discussed the error reduction methods. Results are shown that it is necessary to adopt many approximation and summarization techniques from the field of computational theory. Furthermore, there are still wide areas for further researchers

REFERENCES

[1].P. Domingos and G. Hultan _mining high speed data streams'. Proceedings of International conference on knowledge discovery and data mining, ACM press 2000

[2].Pedro Domingos _Mining high speed data streams' proceedings of International conference on Artificial Intellignence, University of Washington, 2001

[3].R. Duda, P. Hart, and D. Stork. Pattern classification, Willey and sons, 2012.

[4].Breiman L., Friedman J. H., Olshen R. A., and Stone, C.

J. on Classification and Regression Trees 1984. Wadsworth Gey S, Lebarbier E, Using CART to detect multiple change-points in the mean for large samples, on statistics for Systems Biology, 2008

[5].Rokach, Lior; Maimon, O. Data mining with decision trees: theory and applications. 2008.

[6].Strobl, C.; Malley, J.; Tutz, G. "An Introduction to Recursive Partitioning‖: Rationale, Application and Characteristics of Classification and Regression Trees, 2013

[7].Barros, Rodrigo C., Basgalupp, M. P., Carvalho, F., Freitas, Alex A on ― A Survey of Evolutionary Algorithms for Decision-Tree Induction‖. 2012.

[8].Barros R. C., Cerri R., Jaskowiak P. A., Carvalho, A. C. P. L. F., ― A bottom-up oblique decision tree induction algorithm‖. on Intelligent Systems Design and Applications 2011.

[9].Rokach, L.; Maimon, O. "Top-down induction of decision trees classifiers-a survey". 2005.