THINK INDIA JOURNAL

ISSN:0971-1260
Vol-22- Issue-14-December-2019

Dynamic Prophetic Ensemble Predictive Algorithms for Parametric and Non-Parametric Supervised Learning for Extremely Randomized Training Dataset

Dr.D.Kalaivani,

Associate Professor and Head,

Department of Computer Technology,

Dr.SNS Rajalakshmi College of Arts and Science,

Chinnavedampatti, Coimbatore, India

Email : dkalaivani77@gmail.com

Mobile: 98651 61888

Abstract

Prediction technique is one of the strongest subset in the field of Data Mining to find out dynamic patterns or behaviour among the training datasets. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Prediction values are usually continuous when classification values are discrete. Neural Network Analyses, Classification, Regression trees and Decision Trees are major analytic methods for predictive data mining. Decision trees learn from data to approximate a sine curve with a set of if-then-else decision rules. Decision Tree method is one of the non-parametric supervised learning method used for classification and regression. Neural network method is used for classification, clustering, feature mining, prediction and pattern recognition. In more practical terms neural networks are non-linear statistical data modeling tools.

Keywords: **Data Mining, Prediction Techniques, Training datasets, Classification, Regression, Decision Tree.**

Copyright ⓒ 2019Authors

ISSN:0971-1260
Vol-22- Issue-14-December-2019

I. Introduction

Prediction Techniques are based on Mathematical Models such as Simple Statistical Model, Non-Linear Statistical Model and Neural Networks. Decision Trees are simple to understand and to interpret. Data models can be visualized. Decision Trees require little data preparation. Decision Tree method creates rules followed by process. Other techniques often require data normalization, dummy variables need to be created and blank values to be removed. Note however that this module does not support missing values. The cost of using the tree (i.e., predicting data) is logarithmic in the number of data points used to train the tree. They are able to handle both numerical and categorical data. Other techniques are usually specialized in analyzing datasets that have only one type of variable. Decision Trees are able to handle multi-output problems. With respect to other tree-based ensemble methods such as Tree Bagging or Random Forests, Extra-Trees select cut-points at random and use the whole learning sample rather than a bootstrap replica. Their node splitting algorithm depends on two parameters, namely the size K of the random subset of attributes considered at each split, and the minimal (sub) sample size to split a node, n_{min}. In principle, any supervised classification method can be used to grow the subwindow classifiers. Mar´eeet al. exploited a particular decision tree ensemble method, called extremely randomized trees or Extra-Trees. [1]

II. Features of Decision Tree Model

- Uses a white box model. If a given situation is observable in a model, the explanation for the condition is easily explained by boolean logic. By contrast, in a black box model (e.g., in an artificial neural network), results may be more difficult to interpret.
- Possible to validate a model using statistical tests. That makes it possible to account for the reliability of the model.
- Performs well even if its assumptions are somewhat violated by the true model from which the data were generated.

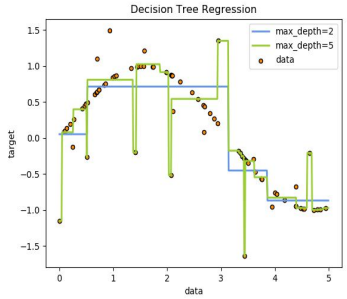

Figure 1. Decision Tree for Linear Regression

Copyright ⓒ 2019Authors

ISSN:0971-1260
Vol-22- Issue-14-December-2019

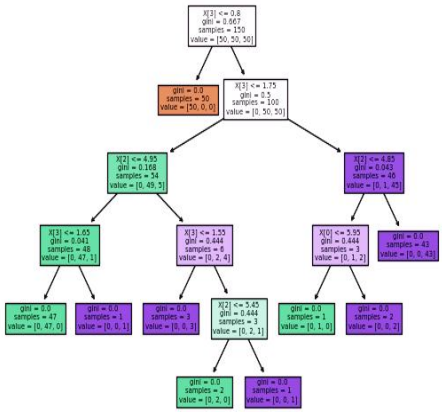

Figure.2 Model of Decision Tree to illustrate gini index

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data (but the resulting classification tree can be an input for decision making) [3].

The disadvantages of decision trees include:

- Decision-tree learners can create over-complex trees that do not generalize the data well. This is called overfitting. Mechanisms such as pruning (not currently supported), setting the minimum number of samples required at a leaf node or setting the maximum depth of the tree are necessary to avoid this problem.
- Decision trees can be unstable because small variations in the data might result in a completely different tree being generated. This problem is mitigated by using decision trees within an ensemble.

III. Use of Gini index in data modeling

The Gini Coefficient or Gini Index measures the inequality among the values of a variable. Higher the value of an index, more dispersed is the data. Alternatively, the Gini coefficient can also be calculated as the half of the relative mean absolute difference. Predictive analytics encompasses a variety of statistical techniques from data mining, predictive modelling, and machine learning, that analyze current and historical facts to make predictions about future or otherwise unknown events [4]. The predictive analytics used to predict trends, improve performance, drive decision making, and predict the behavior [5] .Data mining classification approach is used to predict the class label by using training data set. Classification approaches are categorized into Decision Tree, Back propagation Neural Network. Support Vector Machine (SVM) Rule based Classification and Bayesian Classification. Decision Tree (DT) approach is used to analyze the data in the form decision rules [6]

Graphviz:

Graphviz is open source graph visualization software. Graph visualization is a way of

Copyright ⓒ 2019Authors

ISSN:0971-1260
Vol-22- Issue-14-December-2019

representing structural information as diagrams of abstract graphs and networks. It has important applications in networking, bioinformatics, software engineering, database and web design, machine learning, and in visual interfaces for other technical domains. The decision tree structure can be analyzed to gain further insight on the relation between the features and the target to predict. In this example, we show how to retrieve:

- the binary tree structure;
- the depth of each node and whether or not it's a leaf;
- the nodes that were reached by a sample using the decision path method;
- the leaf that was reached by a sample using the apply method;
- the rules that were used to predict a sample;
- the decision path shared by a group of samples.

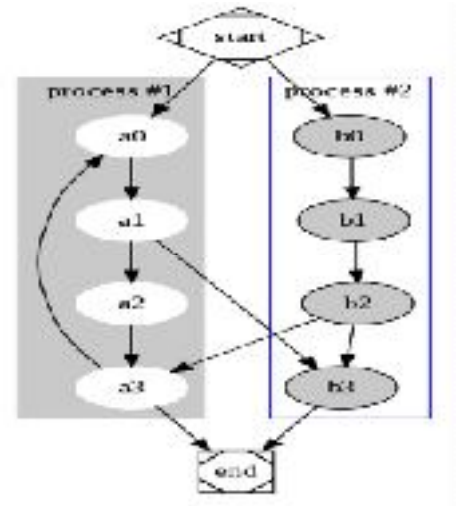

Figure 3. Decision Tree illustrated using Graphviz- A Graphical Visualization Tool

IV. Roadmap – Hierarchical Method

A product roadmap is a guide that allows communicating the vision for the product's future. Some roadmaps show short-term visions of a product, while others show the life cycle of a product over a longer time.

dot - "hierarchical" or layered drawings of directed graphs. This is the default tool to use if edges have directionality.

nexto - "spring model" layouts. This is the default tool to use if the graph is not too large (about 100 nodes) and you don't know anything else about it. Nexto attempts to minimize a global energy function, which is equivalent to statistical multi-dimensional scaling.

fdp - "spring model'' layouts similar to those of neato, but does this by reducing forces rather than working with energy.

sfdp - multiscale version of fdp for the layout of large graphs.

twopi - . Nodes are placed on concentric circles depending their distance from a given root node.

circo - This is suitable for certain diagrams of multiple cyclic structures, such as certain telecommunications networks.

V. A Random Forest classifier.

A random forest is a meta-estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.Predictive analytics speaks to the key aspects of product development: time to market, cost, quality, and compliance. [7]

Copyright © 2019Authors

ISSN:0971-1260
Vol-22- Issue-14-December-2019

Parameters

n_estimatorsinteger, optional (default=100)

The number of trees in the forest.The default value of n_estimators changed from 1 to 100 in 0.22.

Criterion string, optional (default="gini=1").The function to measure the quality of a split. Supported criteria are "gini" for the Gini impurity and "entropy" for the information gain. Note: this parameter is tree-specific.

max_depthinteger or None, optional (default=None)

The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.

min_samples_splitint, float, optional (default=2)

The minimum number of samples required to split an internal node:

If int, then consider min_samples_split as the minimum number.

If float, then min_samples_split is a fraction and ceil(min_samples_split * n_samples) are the minimum number of samples for each split.Added float values for fractions.

min_samples_leafint, float, optional (default=1)

The minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least min_samples_leaf training samples in each of the left and right branches. This may have the effect of smoothing the model, especially in regression.

If int, then consider min_samples_leaf as the minimum number.

If float, then min_samples_leaf is a fraction and ceil(min_sample_leaf * n_sample) are the minimum number of samples for each node. Added float values for fractions.

min_weight_fraction_leaffloat, optional (default=0)

The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Samples have equal weight when sample_weight is not provided_maxint_feature, float, string or None, optional (default="auto"). **If float, then max_features is a fraction and int(max_features * n_features) features are considered at each split.**

If "auto", then max_features=sqrt(n_features).

If "sqrt", then max_features=sqrt(n_features) (same as "auto").

If "log2", then max_features=log2(n_features).

If None, then max_features=n_features. The search for a split does not stop until at least one valid partition of the node samples is found.

max_leaf_nodesint or None, optional (default=None)

Grow trees with max_leaf_nodes in best-first fashion. Best nodes are defined as relative reduction in impurity. If None then unlimited number of leaf nodes.

min_impurity_decreasefloat, optional (default=0.)

A node will be split if this split induces a decrease of the impurity greater than or equal to this value. The weighted impurity decrease equation is the following:

N_t / N * (impurity - N_t_R / N_t * right_impurity - N_t_L / N_t * left_impurity)

where N is the total number of samples, N_t is the number of samples at the current node, N_t_L is the number of samples in the left child, and N_t_R is the number of samples in the right child. N, N_t, N_t_R and N_t_L all refer to the weighted sum, if sample_weight is passed.

min_impurity_splitfloat, (default=1e-7)

Threshold for early stopping in tree growth. A node will split if its impurity is above the threshold, otherwise it is a leaf. min_impurity_split has been deprecated in favor of min_impurity_decrease in

Copyright ⓒ 2019Authors

ISSN:0971-1260
Vol-22- Issue-14-December-2019

0.19. The default value of min_impurity_split will change from 1e-7 to 0 in 0.23 and it will be removed in 0.25. Use min_impurity_decrease instead.

bootstrapboolean, optional (default=True)
Whether bootstrap samples are used when building trees. If False, the whole datset is used to build each tree.

n_jobsint or None, optional (default=None)
The number of jobs to run in parallel. fit, predict, decision_path and apply are all parallelized over the trees.

random_stateint, RandomState instance or None, optional (default=None)
Controls both the randomness of the bootstrapping of the samples used when building trees (if bootstrap=True) and the sampling of the features to consider when looking for the best split at each node (if max_features < n_features). verboseint, optional (default=0).Controls the verbosity when fitting and predicting.

warm_startbool, optional (default=False)

When set to True, reuse the solution of the previous call to fit and add more estimators to the ensemble, otherwise, just fit a whole new forest. class_weightdict, list of dicts, "balanced", "balanced_subsample" or None, optional (default=None).Weights associated with classes in the form {class_label: weight}. If not given, all classes are supposed to have weight one. For multi-output problems, a list of dicts can be provided in the same order as the columns of y.

VI. Graphical Representation of the Gini Index (Lorenz curve)

The Gini coefficient is usually defined mathematically based on the Lorenz curve, which plots the proportion of the total income of the population (y-axis) that is cumulatively earned by the bottom x% of the population. The line at 45 degrees thus represents perfect equality of incomes. Fitting a curve through data is also referred as finding relationship through predictors. Precision is also called as positive predictive value. In pattern recognition and information retrieval with binary classification, precision is the fraction of retrieved instances that are appropriate. Documents referred as instances and a set of specified documents returned by a search engine referred as tasks in an information retrieval system or equivalently, "relevant" and "not relevant" are two categories assigned to each document

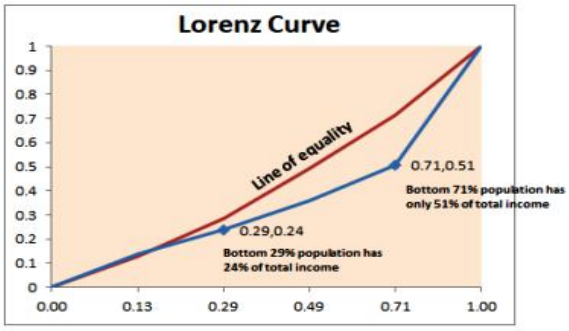

Figure.4 Illustration of Lorenz Curve (The line at 45 degrees thus represents perfect equality of incomes)

Table.1 Predicting the Equality of income

Copyright ⓒ 2019Authors

ISSN:0971-1260
Vol-22- Issue-14-December-2019

Buckets	Predicted value	N
1	460	83
2	343	104
3	406	134
4	501	143
5	1662	185

B=Sum(Area)	0.410506
A=0.5-B	0.089494
Gini=A/(A+B)	0.178987

Area=(Cum Pred dist+Prev cum dist)*0.5*Freq dist

Mean Squared Error is calculated as:

$$\bar{y}_m = \frac{1}{N_m} \sum_{i \in N_m} y_i$$

$$H(X_m) = \frac{1}{N_m} \sum_{i \in N_m} (y_i - \bar{y}_m)^2$$

Mean Absolute Error is calculated as:

$$\bar{y}_m = \frac{1}{N_m} \sum_{i \in N_m} y_i$$

$$H(X_m) = \frac{1}{N_m} \sum_{i \in N_m} |y_i - \bar{y}_m|$$

where X_m is the training data in node m

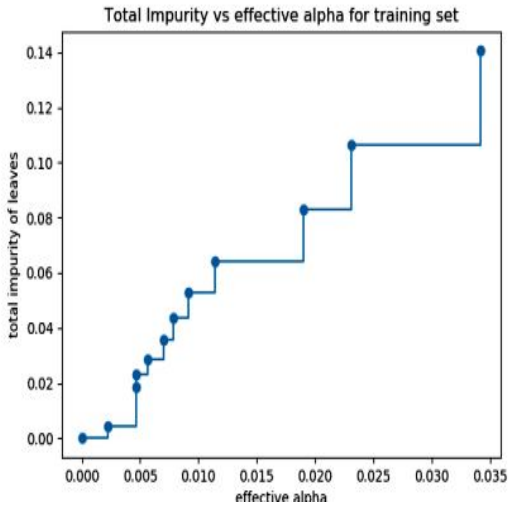

Figure.5 Complexity Pruning

VII Result Analysis

There are different techniques used for Predicting data and result is continuous sometimes discrete also. Error Back Propagation comes under Neural Networks. An algorithm known as error back propagation uses this error value to change the weights The weight change from the input layer unit i to hidden layer unit j is: i_j j i $\Delta w = \eta < \delta < x k$ *k where δ j = o_j 1(− oj) $\sum w_{jk} < \delta$ The weight change from the hidden layer unit j to the output layer unit k is: where w_{jk} k o j $\Delta = \eta$,$\delta < 1$.

VIII Conclusion

The prediction techniques should have the capacity to handle over fitting and under-fitting. Whichever technique you use, it should have the following qualities: – Ability to make correct predictions on data that is not in the original training data. And should possess the ability to provide a certainty measure with its predictions. How well a solution performs depends on both the data and the person who built it. Over Fitting – A data mining predictor can capture the structure of the data so well that irrelevant details are picked up and used when they are not generally true. Data Quality and Quantity – Insufficient data or data that does not capture the relationship between predictors and predicted can produce a very poor solution. Depending on the nature of training datasets the performance of Algorithm relies.

References:

[1] M. Dumont et al, Fast multi-class image annotation with random sub windows and multiple output randomized trees,

Copyright ⓒ 2019Authors

ISSN:0971-1260
Vol-22- Issue-14-December-2019

International Conference on Computer Vision Theory and Applications 2009

[2] Rajan Gupta and Chaitanya Pathak, A Machine Learning Framework for Predicting Purchase by Online Customers based on Dynamic Pricing, Procedia Computer Science 36(2014) 599-605

[3]https://en.wikipedia.org/wiki/Decision_tree_learning

[4]https://en.wikipedia.org/wiki/PredictiveAnalytics

[5] S.Banumathi and A.Aloysius, Predictive Analytics Concepts In Big Data- A Survey 8(8), September-October 2017

[6] D.Kalaivani et al., Customer Behavior Predictive Modeling for Online Shopping using Tuned Decision Tree Method, Helix Vol. 7(5): 1994-1999

[7]https://www.cognizant.com/whitepapers/predictive-analytics-accelerating-and-enriching-product-development

[8] M Krishna Satya Varma et al.Comparative Study of Different Data Mining Prediction Algorithms,7(2)http://dx.doi.org/10.21172/1.72.562 394,2016

[9] Shubpreet Kaur, a. D. (2015). Future Trends of Data Mining in Predicting the Various Diseases in Medical Healthcare System. International Journal of Energy, Information and Communications , 17-34 .

[10] Shubpreet Kaur, D. R. (2015). Future Trends of Data Mining in Predicting the Various Diseases in Medical Healthcare System. International Journal of Energy, Information and Communications, 17-34.

[11]. Shukla, S. B. (2014). A Literature Review in Health Informatics Using Data Mining Techniques. International journal of software and hardware research engineering, 2(2).

[12] Sitar-Taut, V. e. (2009). "Using machine earning algorithms in cardiovascular disease risk evaluation. "Journal of Applied Computer Science & Mathematics.

[13]. U. Fayyad, G. P.-S. (1996). The KDD process of extracting useful knowledge form volumes of data.commun. ACM, 39(11), 27-34.

[14]. WHO. (n.d.). Retrieved May 25, 2015, from http://www.who.int/influenza/human_animal_interface/en.

[15] Yadav, S. K. (2012). "Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification. World of Computer Science and Information Technology (WCSIT).

[16] Yafeng Lu, Robert Kr ¨uger, Dennis Thom, Feng Wang, Steffen Koch, Thomas Ertl, Ross Maciejewski. “Integrating Predictive Analytics and Social Media”. November 9- 2014. IEEE [19].

[17] Fangzhou Sun, Yao Pan, Jules White, Abhishek Dubey. “Real-time and Predictive Analytics for Smart Public Transportation Decision Support System”. 2016. IEEE. [20].

[18] Chandra J, DrNachamai.M, DrAnitha S Pillai. “An Integrated Approach to Predictive Analytics for Stock Indexes and Commodity Trading Using Computational Intelligence”. Vol. 5, No. 8, 142-148, 2015. World of Computer Science and Information Technology Journal (WCSIT).

[19]. HodaMoghimi, Stephen Vaughan, Steven McConche, NilminiWickramasinghe.” How Do Business Analytics and Business Intelligence Contribute to Improving Care Efficiency?”. 2016. 49th Hawaii International Conference on System Sciences.

[20]. Suraj Pandey, Surya Nepal, Shiping Chen.” A Test-bed for the Evaluation of Business Process Prediction Techniques”. October 2011. International Conference on Collaborative Computing: Networking, Applications and Worksharing.

[21]. Johannes Kunze von Bischhoffshausen, Markus Paatsch, Melanie Reuter, Gerhard Satzger, Hansjoerg Fromm. “.An Information System for Sales Team Assignments Utilizing Predictive and Prescriptive Analytics”. 2015. IEEE 17th Conference on Business Informatics.

Copyright ⓒ 2019Authors