Optimal Vacation Destination using Joint Splitting in Binary Trees

Adarsh Sahu
201IT203
Information Technology
NIT Karnataka
Surathkal, India

Kaustubh Vivek Khedkar
201IT128
Information Technology
NIT Karnataka
Surathkal, India

Vishruth M
201IT167
Information Technology
NIT Karnataka
Surathkal, India

***Abstract*—In this paper, we implement joint splitting criteria using two of the most used criterion i.e. Information Gain and Gini index to solve a real-world problem. Using the linear combination of Information Gain and Gini Index, Joint Splitting provides a more accurate index of the purity of a split. This is used to predict the optimal destination.**

***Index Terms*—Decision tree, Gini Index, Information Gain, Joint splitting criterion, Random Forest.**

I. Introduction

Everyone wants their trip to be a success and that starts with a right-fit destination but choosing the optimal vacation destination can feel overwhelming. This project implements a model which will help the user to decide on the vacation destination based on the various criteria provided by the user such as Budget, Weather and if it works as a family gathering spot, etc.

A decision tree is used to predict the unknown input instance by applying a set of decision rules at each node of the tree. A splitting criterion is used in order to decide on node selection and corresponding attribute value. There have been various splitting criteria defined in the literature including the entropy-based method, Bayesian network, Gini Index (GI) and Information Gain (IG) based methods. A rough set-based approach has been proposed to handle the uncertainty present during the process of inducing decision trees. Decision tree classification also extended as ensemble classifier, known as the Random Forest, wherein multiple decision trees cast their votes for predicting an unknown input. All such decisions are then aggregated by a predefined mechanism like majority voting or weightage mechanism.

II. Literature Survey

The following journals and research papers were surveyed for the project.

Vikas Jain, Ashish Phophalia and Jignesh S. Bhatt et al. [1] Investigation of a Joint Splitting Criteria for Decision Tree Classifier Use of Information Gain and Gini Index, TENCON 2018 - 2018 IEEE Region 10 Conference, in this paper they have investigated joint splitting criteria using two of the most used criterion i.e. Information Gain and Gini index.

K. I. Sofeikov; I. Yu. Tyukin; A. N. Gorban; E. M. Mirkes; D. V. Prokhorov and I. V. Romanenko et al. [2] proposed a system that helps in Learning optimization for decision tree classification of non-categorical data with information gain impurity criterion in the 2014 International Joint Conference on Neural Networks (IJCNN)

Shruthi Kothari et al. [3] proposed yoga pose detection using deep learning where in, with the help of DL and ML yoga poses are classified with the help of pre-recorded video and also in real time. The project talks about different pose estimation and methods of detection of key points in a detailed manner and explains various learning models (deep learning models) used for classification of poses.

Mohammad Azad ,Igor Chikalov ,Shahid Hussain and Mikhail Moshkov et al. [4] proposed a system of Entropy-Based Greedy Algorithm for Decision Trees Using Hypotheses in the Special Issue of Rough Set Theory and Entropy in Information Science

Henry E. L. Cagnini; Rodrigo C. Barros and Márcio P. Basgalupp et al. [5] proposed a system for the Estimation of distribution algorithms for decision-tree induction in the 2017 IEEE Congress on Evolutionary Computation (CEC)

III. Problem Statement

To design and implement a system which pick's the optimal vacation destination based on joint splitting by comparing features such as duration of the vacation, personal budget, weather forecast, if their family is joining or not.

IV. Methodology

It is observed that the GI and IG are the two most used splitting measures in the literature. GI can be seen as a minimization problem whereas IG as a maximization problem.

The base paper proposes a new splitting criteria, called Joint Splitting which combines GI and IG linearly. Our project uses Joint Splitting to solve a real-world problem.

Decision tree classification also extended as ensemble classifier, known as the Random Forest, wherein multiple decision trees cast their votes for predicting an unknown input. However, Random Forest is suitable for situations when we have a large dataset and is inconsistent on smaller datasets. Decision trees are much easier to interpret and understand. It is also easier to compare the difference in results when using GI, IG and JS.

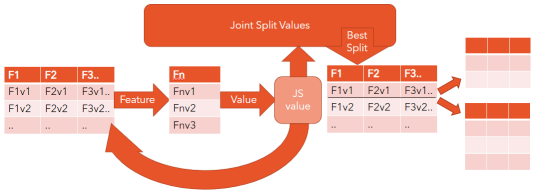

Fig. 1. Flow of the project

Gini Index: Measures the impurity or class inequality present in the given dataset. The GI value lies in the range of [0,1], where 0 corresponds to the equality and 1 corresponds to the inequality. Let D is the dataset, then the GI for dataset D over the attribute X is given as:

$$GI_X(D) = \sum_{j=1}^{k} p(X = x_j)\{1 - \sum_{i=1}^{v} p(Y = y_i/X = x_j)^2\},$$

Fig. 2.

Information Gain: The concept of IG is based on minimizing randomness present in the dataset at each step. The objective function is to maximize the IG by splitting the training data arriving at the node into multiple child subsets.

$$IG(X = x_j) = \underset{\forall j}{\operatorname{argmax}}\{H(Y) - H(Y/(X = x_j))\},$$

Fig. 3.

$$H(Y) = -\sum_{i=1}^{v} p(Y = y_i) log(p(Y = y_i))$$

$$H(Y/(X = x_j)) = -\sum_{j=1}^{k} p(X = x_j)$$
$$\left\{-\sum_{i=1}^{v} p\left\{\frac{Y = y_i}{X = x_j}\right\} logp\left\{\frac{Y = y_i}{X = x_j}\right\}\right\}$$

Fig. 4.

Initially a dataset is prepared containing various destinations and features. This dataset will be fed to prepare the tree.

To split the dataset, a feature and particular splitting value will be taken. The Joint Split value of the split formed by this value is calculated and stored. Higher the value, better the split. This is repeated for all values and features and the split which returns the highest value for Joint Split.

The dataset is divided into two based on the threshold and a binary node is formed. When Joint Split index reaches value 1.0, it is said to be a pure split. If depth limit is given, the tree terminates.

The quality of the split is determined by the method used to calculate impurity/uncertainty in split.

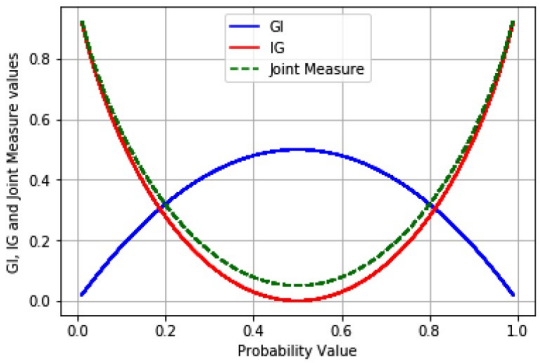

Fig. 1. Characteristic curves of Gini Index, Information Gain and a joint measure.

Fig. 5. Comparison of curves

Joint Splitting: To explore the space between these two criteria (GI and IG splitting criteria), we have formulated a splitting criterion using a convex combination of GI and IG as shown in Equation. The corresponding curves of GI, IG and joint split function are shown in Figure 1. Let D be a dataset having N number of instances. Each instance is having M number of attributes. Initially, a random splitting point is selected as threshold from the given set of attributes. The joint split criterion is applied over the selected attribute value.

$$J_{value}(X) = \underset{i}{\operatorname{argmax}}\{\alpha * IG(X_i) + (1 - \alpha) * \{1 - GI(X_i)\}\}$$

Fig. 6.

V. Result and Analysis

We have investigated a joint splitting criterion in order to design a better classifiers using decision trees. It has been observed that the proposed joint splitting criterion is working satisfactory.

As mentioned in the based paper joint splitting is better than information gain and gini impurity individually .The splits provided by joint splitting index have higher purity value and result in better decision trees .

This better split allows us to make better decision trees with minimal data . Thus, here joint splitting criterion is used in order to design a better classifiers using decision trees helping us to arrive at best option here which is the location using less data input.

VI. Conclusion

This project model predict whether the vacation destination is optimal to the user based on the factors such as:

- Duration of the vacation
- Personal budget
- Weather forecast
- If their extended family is joining

VII. Future work

The next steps would be to improve the usability of the model by:

- Increasing the size of the dataset
- Implementing a better interface
- Displaying the tree after a new tree is built

VIII. Base Paper

Vikas Jain, Ashish Phophalia and Jignesh S. Bhatt "Investigation of a Joint Splitting Criteria for Decision Tree Classifier Use of Information Gain and Gini Index", TENCON 2018 - 2018 IEEE Region 10 Conference; https://ieeexplore.ieee.org/document/8650485

IX. References

[1] K. I. Sofeikov; I. Yu. Tyukin; A. N. Gorban; E. M. Mirkes; D. V. Prokhorov and I. V. Romanenko"Learning optimization for decision tree classification of non-categorical data with information gain impurity criterion" 2014 International Joint Conference on Neural Networks (IJCNN); https://ieeexplore.ieee.org/document/6889842/;

[2] Mohammad Azad ,Igor Chikalov ,Shahid Hussain and Mikhail Moshkov "Entropy-Based Greedy Algorithm for Decision Trees Using Hypotheses" Special Issue Rough Set Theory and Entropy in Information Science; https://www.mdpi.com/1099-4300/23/7/808

[3] Henry E. L. Cagnini; Rodrigo C. Barros and Márcio P. Basgalupp "Estimation of distribution algorithms for decision-tree induction" 2017 IEEE Congress on Evolutionary Computation (CEC) https://ieeexplore.ieee.org/document/7969549

[4] Shagufta Tahsildar "Gini Index For Decision Trees" https://blog.quantinsti.com/gini-index/

[5] S. Sivagama Sundhari "A knowledge discovery using decision tree by Gini coefficient"2011 International Conference on Business, Engineering and Industrial Applications https://ieeexplore.ieee.org/document/5994250