Smart Intelligent Computing and Communication Technology
V.D. Ambeth Kumar et al. (Eds.)

doi:10.3233/APC210018

Comparative Analysis Among Decision Tree vs. Naive Bayes for Prediction of Weather Prognostication

V.Sindhu[a,1] and M. Prakash[b]

[a,1] *Sri Shakthi Institute of Engineering and Technology, Coimbatore Tamilnadu*
[b] *Karpagam College of Engineering, Coimbatore, Tamilnadu*

Abstract: In the previous era, a computer is programmed for some specific task. An electronic device is programmed to do its function electronically. It was done with a target device, the programming environment and the system. We get the necessary intermediate code by running the program with the above said environment and committed into the target device. Thus the device performs the task it was intended to do. In case if we need to change the functionality of the device by the learning experience of the vendor and users, the vendor will upgrade the product. Nowadays in this machine learning era, the devices are programmed in such a way it can learn by its own experience and with the available data it collected it can even manipulate the algorithm by itself with the provided data set. Thus machine learning is ruling this era. We are going to discuss the machine learning algorithms here which was used to predict by itself with the data set collected. Therefore, machine learning is all about learning about computer algorithms that progress its potential through the experience. Thus, Machine learning is presently highly regarded analysis topic and applied to all told application in day to day life. In this paper we have a tendency to extract the knowledge of machine learning algorithms like decision tree, Naive Bayes and enforce the algorithms with sample dataset of weather prognostication.

Keywords —Machine learning, Naive Bayes, Decision tree.

1. Introduction

Machine learning is a programming technique that computers use to upgrade a presentation basis utilizing model information or past expertise. Machine learning will routinely acknowledge the recognized patterns in the information, and thus to utilize the revealed examples to anticipate future information or alternative outcomes of interest. Other words, Machine learning could be a category of algorithmic program that is data driven [1] (i.e) in contrast to traditional algorithmic program. the information that narrates and determines what the "smart response" is. Where as writing the software is chunk let the data do the effort instead.

[1]M. Prakash, *Sri Shakthi Institute of Engineering and Technology, Coimbatore*
Email id:salemprakash@gmail.com

Machine learning is classified into following major types of algorithm. They are
(i)Supervised (ii) Unsupervised (iii)Reinforcement

Supervised: It works below supervising, it is a model can foresee with the assis tance of labeled dataset. Labeled dataset is data which is already famed the target answer is called Labeled dataset. It is a type of ML that uses a known training dataset to make prediction. Supervised learning is categorized two types classification and regression. Classification: Once the yielding variable is downright with two or more classes (yes/no; true/false; red/blue) we have a tendency to create use of categorization [2]. Regression: Once the yielding variable is actual or constant quantity, it is not converted to the equivalent numerical quantity [3]. Unsupervised: No supervision, no training given to machine learning at the side of it acts on the data, which is not labeled. It tries to spot the patterns and provides the response. "It is a type of ML algorithm used to draw inferences from dataset consisting of input data without labeled response". Unsupervised learning is categorized two types clustering and association. Clustering: The method of dividing the object into clusters which are comparative among them and are disparate to the object that is owned by another cluster. Association: It is a rule based machine. It discovers interesting relation ship between variable in large dataset [4]. Reinforcement: It establishes and encourage pattern of behavior. This algorithm was designed as how the brain of human respond to punishment and rewards they learn from outcomes and decide on next action. It has to make lot of small decision without human guidance [5].

2. Decision Tree Terminology

2.1 Decision Tree

It is a deliberative portrayal of all conceivable solution to a verdict dependent on explicit condition. Here, the internal nodes represent test on the attributes. For example, if u call the customer care it will redirect to Intelligent Computer Assistant and it will reply like press 1 for English and press 2 for Tamil, press 3 for Hindi and eventually it will redirect to the authorized person here the corporate using decision tree algorithm to require certain decisions [6]. The basic terminologies are given below:
Root node: This can be a whole population or just a few. This can be fragmented further into more consistent sets. Leaf node: This leaf node cannot be further grouped.
Splitting: It is a process where a sub node or a root node is break down into different elements on a given condition Branch/Sub tree: A branch is formed by splitting a tree
Pruning: The process of cutting the branches that is not needed or not useful. Child/Parent Node: The root node is termed as parental and whatever a parent bears or arises from its nodes forms a child [7][15-21].

2.2. Sample Dataset

To make the comparative analysis between decision tree and "Naive Bayes" the following sample dataset is used [8].

Table 1. Sample weather Database

DAY	OUTLOOK	HUMIDITY	WINDY	PLAY
DAY1	SUNNY	HIGH	WEAK	NO
DAY2	SUNNY	HIGH	STRONG	NO
DAY3	OVERCAST	HIGH	WEAK	YES
DAY4	RAIN	HIGH	WEAK	YES
DAY5	RAIN	NORMAL	WEAK	YES
DAY6	RAIN	NORMAL	STRONG	NO
DAY7	OVERCAST	NORMAL	STRONG	YES
DAY8	SUNNY	HIGH	WEAK	NO
DAY9	SUNNY	NORMAL	WEAK	YES
DAY10	RAIN	NORMAL	WEAK	YES
DAY11	SUNNY	NORMAL	STRONG	YES
DAY12	OVERCAST	HIGH	STRONG	YES
DAY13	OVERCAST	NORMAL	WEAK	YES
DAY14	RAIN	HIGH	STRONG	NO

2.3. Building a Decision Tree

To build and decide the tree where to split in a decision tree we must have knowledge in the following terms: Gini index: The decision tree has to be built evaluating the impurity (or purity) in classification and regression tree algorithm is gini index [9].
Information Gain: The information gain is the reduction in entropy after a collection of data is part based on a characteristic developing a decision tree is tied in with discovering quality that returns the utmost information gain.Information gain can be determined by the following formula:Information Gain = Entropy(s)- [(Weighted Average) *Entropy (each feature)] Reduction in variance: It is an algorithm utilized for persistent objective variable. The split with lower change is chosen as the models to part the populace. Chi Square: It is an algorithm to find the quantifiable significance between the sub nodes and parental nodes.

2.4. Measuring the impurity

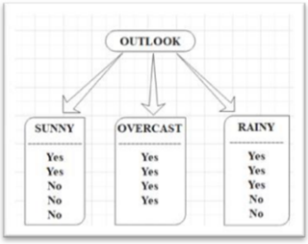
Case 1: Consider the two baskets one of its containing full of lemon and other basket is having name of the fruits as text as lemon if we select each one item randomly in basket probability of getting item will be same. So the impurity will be zero.
Case 2: Consider the two basket one of its containing lemon, apple, orange and other basket is having name of the fruits as text if we select each one item random y in basket probability of getting item will be different. So the impurity will be non-zero [10].

2.5. Entropy

It is termed used for calculating information gain. It is a metric used to
measure the impurity of something. It is a very first step to solve decision tree [11]. Calculate the entropy using the following formula.

$$\text{Entropy}(s) = -P(\text{Yes})\log_2 P(\text{Yes}) - P(\text{No})\log_2 P(\text{No})$$

In the sample dataset out of 14 instance, we have 9 Yes and 5 No
Let us assume Outlook as root node and calculate the entropy for each



$$E(\text{Outlook=Sunny}) = -2/5\log_2 2/5 - 3/5\log_2 3/5 = 0.971$$

$$E(\text{Outlook = Overcast}) = -4/4\log_2 4/4 - 0\log_2 0 = 0$$

$$E(\text{Outlook = Rainy}) = -3/5\log_2 3/5 - 2/5\log_2 2/5 = 0.971$$

Figure 1. Outlook feature

Weighted average is,

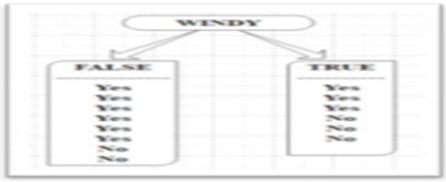


Figure 2. Windy

Let us assume Windy as root node and calculate the entropy for each feature,Similarly Calculate for the Humidity. Weighted average is I(Humidity) = 0.788 and the information gained from Humidity is Gain(Humidity) = 0.152

Outlook	Gain=0.247
Humidity	Gain=0.152
Windy	Gain=0.048

So the Maximum gain is Outlook (i.e.) 0.247 So Outlook is the best root node. Similarly, node to select further should be decided. Complete decision tree will be formed as given below in figure 3 and reducing complexity is shown in figure 4.

$$I(\text{Outlook}) = 5/14 * 0.971 + 4/14 * 0 + 5/14 * 0.971 = 0.693$$

Information gained from Outlook Gain(Outlook)

$$= E(s) - I(\text{Outlook}) = 0.94 - 0.693 = 0.247$$

Weighted average is,

$$I(\text{Windy}) = 6/14 * 1 + 8/14 * 0.811 = 0.892$$

Information gained from Windy.

$$\text{Gain(Windy)} = E(s) - I(\text{Windy}) = 0.94 - 0.892 = 0.04.$$

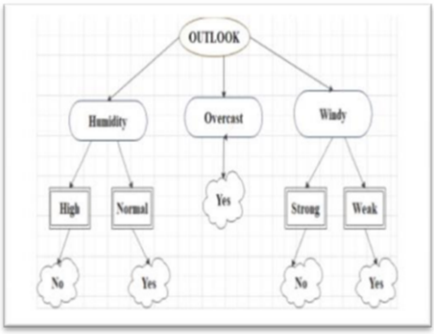


Figure 3. Pruning Tree

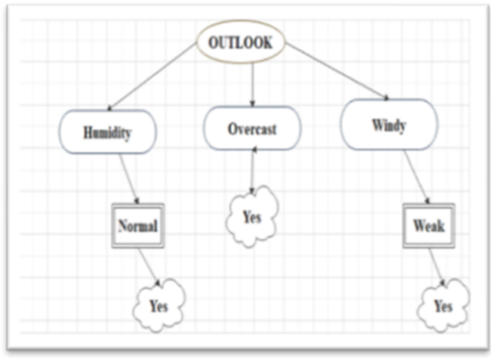


Figure 4. Reducing Complexity

3. What is Naive Bayes

This is algorithm is quite a normal but a phenomenal as for as prediction models are concerned [12]. It is one of the hierarchical classification methods of Bayes theorem. Naive and Bayes are the two categories of this algorithm [13]. An assumption like one feature which is considered here is no more a relation to another feature in this class, though that feature is dependent to other. So each feature in this class contributes with its functionality to the probability. It is used in large dataset[14].

3.1 Bayes theorem

"Given a Hypothesis H and evidence E, Bayes theorem states that the relationship between the probability of the hypothesis before getting the evidence P(H) and the probability of the hypothesis after getting the evidence P(H/E) is" [14].
P(H/E) = P(E/H). P(H) /P(E)

3.1.1 Classification Steps

From the above dataset the frequency table is designed for Outlook, Humidity, Windy [14]

Table 3. Frequency Table

Frequency Table		Play	
		Yes	No
OUTLOOK	Sunny	3	2
	Overcast	4	0
	Rainy	3	2

Frequency Table		Play	
		Yes	No
HUMIDITY	High	3	4
	Normal	6	1

Frequency Table		Play	
		Yes	No
WINDY	Strong	6	2
	Weak	3	3

Next, Likelihood is calculated for Outlook

Table 4. Likelihood for Outlook

Likelihood Table		Play		
		Yes	No	
OUTLOOK	Sunny	3/10	2/4	5/14
	Overcast	4/10	0/4	4/14
	Rainy	3/10	2/4	5/14
		10/14	4/14	

The Probability of getting P(Yes/Sunny) =0.591 and P(No/Sunny) =0.40.
Similarly,
Likelihood for humidity is P(Yes/High) = 0.42 and P(No/High) =0.58.
Likelihood for Windy is P(Yes/Weak) = 0.75 and P(No/Weak) =0.25.

Suppose we have a day with following values
Outlook = Rain
Humidity = High
Windy = Weak
Play = ?

"Likelihood of Yes on that day
= P (Outlook = Rain/Yes) * P (Humidity = High/Yes) * P (Windy = Weak/Yes)*P(Yes)
= 2/9 * 3/9 * 6/9 * 9/14
= 0.0199

Likelihood of No on that day
=P (Outlook = Rain/No) * P (Humidity=High/No) *P (Windy = Weak/No) *P(No)
= 2/5 * 4/5 *2/5 *5/14
= 0.0166

So, P(Yes) = 0.0199 / (0.0199 + 0.0166) = 0.55 and P(No)=0.45
Our model predicts that there will be 55 percentage chances to play the game."

4. Conclusion

This paper gave an exposure in the Machine learning algorithms like decision tree and Naive Bayes. We have used a sample weather dataset and using entropy we have calculated information gain and measure the impurity. Weighted average was calculated for the root nodes outlook, windy and humidity. The maximum gain is measured for the respective root nodes and the algorithm shuffles the value of different root nodes to discover the best root node. This technique was used to discover the maximum gain using the decision tree. The working technique of these algorithms were elaborated and discussed in this paper with the calculated values. The future work can be extended for other machine learning algorithms.

References

[1] Sindhu, V., Nivedha, S., &Prakash, M.(2020). An Empirical Science Research On Bioinformatics In Machine Learning.Journal of Mechanics of Continua and MathematicalSciences.

[2] Sorower MS. A literature survey on algorithms for multi-label learning. Oregon State University, Corvallis, Dec2010.

[3] Wang, X., Lin, X., & Dang, X. (2020). Supervised learning in spiking neural networks: A review of algorithms and evaluations. Neural Networks, 125,258-280.

[4] Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., &Aljaaf, A. J. (2020). A systematic review on supervised and unsupervised machine learning algorithms for data science. Supervised and Unsupervised Learning for Data Science,3-21.

[5] Levine, S., Kumar, A., Tucker, G., & Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. arXiv preprintarXiv:2005.01643.

[6] Wibowo, A. H., &Oesman, T. I. (2020, February). The comparative analysis on the accuracy of k-NN, Naive Bayes, and Decision Tree Algorithms in pre- dicting crimes and criminal actions in Sleman Regency. In Journal of Phys- ics: Conference series (Vol. 1450,p.012076).

[7] Wei, Y., Wang, X., & Li, M. (2020). Intelligent Medical Auxiliary Diagnosis Algorithm Based on Improved Decision Tree.Journal of Electrical and Computer Engineering, 2020.

[8] Rajalakshmi, V., &Parasan, N. (2020). Weather Prediction Using Data Mining Approach. Studies in Indian Place Names, 40(71),1508-1517.

[9] Pathak, A. R., Welling, A., Shelar, G., Vaze, S., &Sankar, S. (2020). A Framework for Performing Prediction and Classification Using Machine Learning. In Proceedings of ICETIT 2019 (pp. 893-906).Springer,Cham.

[10] Sindhu, V., &Prakash, M. (2019, November). A Survey on Task Scheduling and Resource Allocation Methods in Fog Based IoT Applications. In International Conference on Communication and Intelligent Systems (pp. 89-97). Springer,Singapore.

[11] Kelleher, J. D., Mac Namee, B., &D'arcy, A. (2020). Fundamentals of ma- chine learning for predictive data analytics: algorithms, worked examples, and case studies.MITpress.

[12] Lee, I., & Shin, Y. J. (2020). Machine learning for enterprises: Applications, algorithm selection, and challenges. IEEE Business Horizons,63(2),157-170.

[13] Chen, W., Zhang, S., Li, R., &Shahabi, H. (2018) . Performance evaluation of the GIS-based data mining techniques of best-first decision tree, random for- est, and naïve Bayes tree for landslide susceptibility modeling. Science of the total environment,644,1006-1018.

[14] Lestari, F. P., Haekal, M., Edison, R. E., Fauzy, F. R., Khotimah, S. N., &Haryanto, F. (2020, March). Epileptic Seizure Detection in EEGs by Using Random Tree Forest, Naïve Bayes and KNN Classification. In Journal of Physics: Conference Series (Vol. 1505, No. 1, p. 012055). IOPPublishing.

[15] V. D. Ambeth Kumar, S. Sharmila, Abhishek Kumar, A. K. Bashir, Mamoon Rashid, Sachin Kumar Gupta &Waleed S. Alnumay . A novel solution for finding postpartum haemorrhage using fuzzy neural techniques. Neural Computing and Applications (2021) (https://doi.org/10.1007/s00521-020-05683-z)

[16] AnkitKumar,VijayakumarVaradarajan,AbhishekKumar, PankajDadheech, SurendraSinghChoudhary, V.D. AmbethKumar, B.K.Panigrahi, KalyanaC.Veluvolug. Black hole attack detection in vehicular ad-hoc network using secure AODV routing algorithm.Microprocessors and Microsystems, In Press,(https://doi.org/10.1016/j.micpro.2020.103352)

[17] V.D.Ambeth Kumar. A Cognitive Model for Adopting ITIL Framework to Improve IT Services in Indian IT Industries. Journal of Intelligent Fuzzy Systems. (DOI: 10.3233/JIFS-189131)

[18] Ambeth Kumar.V.D .Efficient Data Transfer in Edge Envisioned Environment using Artificial Intelligence based Edge Node Algorithm. Transactions on Emerging Telecommunications Technologies (Accepted - Inpress)(DOI: 10.1002/ett.4110)

[19] V. D. AmbethKumar,S. Malathi,AbhishekKumar,Prakash M and Kalyana C. Veluvolu.Active Volume Control in Smart Phones Based on User Activity and Ambient Noise.Sensors 2020, 20(15), 4117; https://doi.org/10.3390/s20154117

[20] Ambeth Kumar.V.D and M.Ramakrishan (2013).Temple and Maternity Ward Security using FPRS", Journal of Electrical Engineering & Technology, ,Vol. 8, No. 3, PP: 633-637. [http://dx.doi.org/10.5370/JEET.2013.8.3.633]

[21] V.D.Ambeth Kumar and M.Ramakrishan (2013) .A Comparative Study of Fuzzy Evolutionary Techniques for Footprint Recognition and Performance Improvement using Wavelet based Fuzzy Neural Network. for the International Journal of Computer Applications in Technology, Vol.48, No.2,pp.95 – 105. [DOI: ttp://dx.doi.org/10.1504/IJCAT.2013.056016]

[22] B. Aravindh; V.D.Ambeth Kumar; G. Harish; V. Siddartth, " A novel graphical authentication system for secure banking systems", IEEE (ICSTM), Pages: 177 – 183, 2-4 Aug. 2017, DOI: 10.1109/ICSTM.2017.8089147