ANALYSIS OF DECISION TREE ALGORITHM IN MACHINE LEARNING

Mr.N.Selvam[1], R.Saranya[2]

[1]*Assistant Professor, Department of EEE, M Kumarasamy College of Engineering, Karur, TamilNadu*

[2] *PG Student, Department of EEE, M Kumarasamy College of Engineering, Karur, TamilNadu*

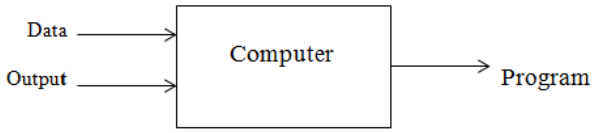
ABSTRACT

Machine learning is the ability to improve the behavior with experience. It is about building computer systems that automatically improve the experience. To make out important classes and predict likely pattern, classification and prediction are most important strategies. In data mining decision tree algorithm is an important classification techniques. As the traditional algorithm of decision tree such as Iterative dichotomiser 3 ,C4.5 and C5.0 has the prosperity of tremendous prediction or classification speed, powerful learning volume and easy to build. However for realistic application these techniques are inadequate. Decision tree techniques are mainly used in E-commerce, social network, web search, and space exploration applications. This paper gives the point of different decision tree technologies such as characteristics, difficulties, preferred standpoint and weakness.

Keywords- *Machine learning, Decision tree, Classification, Prediction.*

1.INTRODUCTION

Machine learning is a branch of science that arrangements with programming the frameworks such that they consequently learn and enhance with understanding. Here, learning implies perceiving and understanding the information and settling on savvy choices in view of the provided information [1]. It is extremely hard to provide food every one of the choices in view of every single conceivable info. To beat this issue, calculations are produced. These calculations construct information from particular information and past involvement with the standards of measurements, likelihood hypothesis, rationale, combinatorial improvement, seek, fortification learning, and control hypothesis [2] . Machine learning is a huge zone and it is very past the extent of this instructional exercise to cover every one of its highlights. There are a few approaches to actualize machine learning systems, however the most generally utilized ones are managed and unsupervised learning [3],[4]. In conventional programming data and program is keep running on the PC to deliver the yield. Be that as it may, in Machine learning as appeared in fig 1 information and yield is keep running on the PC to make a program [5]. This program can be utilized as a part of customary programming.



|Fig 1- Machine learning

There are four types of machine learning they are

1. **Supervised learning**
2. **Unsupervised learning**
3. **Semi-supervised learning**
4. **Reinforcement learning**

If the training data includes desired outputs then it is called as supervised learning or inductive learning but in **unsupervised learning** the Training data does not include desired outputs.**If the t**raining data includes a few desired outputs it is known as **Semi-supervised learning [6].**

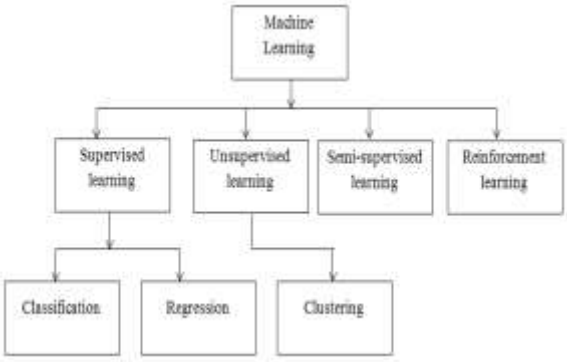


Fig 2 - Types of machine learning

Now a-days for extremely large applications data or information which stored in database can be used.. This dangerous development in information and database has created an earnest requirement for new methods and instruments that can wisely naturally change the prepared information into valuable data and learning. Henceforth information mining has turned into an examination are with expanding significance. Characterization and forecast are two types of information examination that can be utilized to remove models portraying essential information classis or to anticipate future information patterns. Such examination can assist furnish us with better comprehension of the information on the loose [7]. Though arrangement predicts straight out (discrete, unordered) marks, expectation models persistent esteemed capacities. Numerous grouping and forecast techniques have been proposed by scientist in machine learning design acknowledgment and insights. Most calculations are memory occupant, ordinarily accepting a little information measure [8] . Late information mining research has based on such work, creating adaptable arrangement and expectation methods fit for taking care of expansive plate inhabitant data.There are arrangements of information mining strategies, for example, visit design mining,classification, relapse, grouping, affiliation administer mining and numerous more,but out of these classification is as often as possible utilized and no more. In classification, the model is prepared which depicts and differentiate different information classes to foresee the classes whose names are not known. The classification can be performed with different calculations, for example, neural systems, choice trees, relapse, and so on. Because of the significance essential of choice tree for huge informational indexes, in this examination choice tree approach has been utilized.

2. DECISION TREE

A tree can be learned by part the source set into subsets in light of a quality esteem test. This procedure is rehashed on each determined subset in a recursive way called recursive apportioning. The recursion is finished when the subset at a hub has all a similar estimation of the objective variable, or while part never again enhances the forecasts [9] . This procedure of best down acceptance of choice trees (TDIDT) is a case of an eager calculation, and it is by a long shot the most widely recognized technique for taking in choice trees from information.

The choice tree comprises of three basics, root hub, inside hub and leaf hub [10] . Top most central is the root hub. Leaf hub is the terminal major of the structure and the hubs in the middle of is known as the inside hub. Each inward hub indicates test on a trait, each branch speaks to a result of the test, and each leaf hub holds a class mark. Different choice tree calculations are utilized as a part of order like ID3, J48, CART, C5.0, SLIQ, SPRINT, irregular backwoods, arbitrary tree, and so forth. In this work following tree calculations are taken for examination.

ID3 (Iterative Dichotomiser 3) decision tree algorithm is produced by Quinlan . The essential thought of ID3 calculation is to develop the choice tree by utilizing a best down, voracious inquiry through the offered sets to test each trait at each tree hub. In the choice tree technique, data pick up approach is for the most part used to decide reasonable property for every hub of a produced choice tree. Along these lines, we can choose the characteristic with the most astounding data pick up (entropy lessening in the level of greatest) as the test property of current hub. Along these lines, the data expected to characterize the preparation test subset got from later on dividing will be the littlest. In other words, the utilization of this property to parcel the example set contained in current hub will make

the blend level of various kinds for all produced test subsets decrease to a base [11]. Along these lines, the utilization of such a data hypothesis approach will successfully decrease the required partitioning number of protest arrangement. ID3 utilizes just unmitigated ascribes to fabricate a tree demonstrate. This calculation does not create more precise results if the clamor is available in the informational index. Keeping in mind the end goal to get more exact outcomes the compelling pre-preparing is done before show is constructed utilizing ID3.

J48 is a propel rendition of ID3, it chooses target estimation of another test information as for various characteristic benefits of preparing information [3]. The inner hubs of a choice tree are meant by various properties while the branches tell the conceivable estimations of these qualities. The inside hubs tell the reliant variable esteems.

CART, its full form is Classification And Regression Tree (CART) and it was at first proposed by Breiman et al. which was double tree otherwise called (HODA). It is non parametric choice tree. It produces relapse or grouping tree relies upon subordinate variable's compose either numeric or straight out individually. Here the importance of parallel means the hub in a choice tree has two out word branches i.e. gatherings. Gini record is utilized as a part of CART as highlight choice measure. The trait with biggest gini list is utilized to part the records. Truck handles both straight out what's more, numerical esteems alongside missing quality esteems too. It utilizes costcomplexity pruning and furthermore create relapse trees. It assembles the two groupings and relapse trees. It can be executed serially from Hunt's calculation. It additionally utilizes relapse investigation utilizing regressin trees (S.Anupama et al,2011). Over a given timeframe and the arrangement of indicator factors relapse investigation highlight conjectures a needy variable. It gives high characterization and expectation exactness.

C5.0 is an augmentation of C4.5 which was at first got from ID3. It is connected on enormous informational collections. It is substantially speedier and memory effective than C4.5. It parts the examples in view of the most extreme data pick up. The example subset that is get from the previous part will be part a while later. This is consistent process until the point when the example subset can not be additionally part. The characteristics/highlights which have less commitment will be rejected. One noteworthy preferred standpoint of C5.0 is it handles multi esteem properties and missing qualities from the informational index.

3.DECISION TREE WORKING METHODOLOGY

1. First choose the training experience or data
2. Choose target fuction that is to be learned.
3. Choose how to represent the target function.
4. Finally choose the learning algorithm to inter the target function.

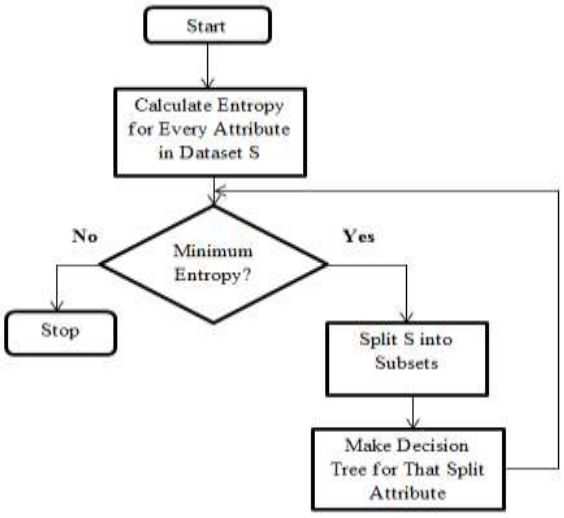


Fig 3 - Working of decision tree

For the most part the classification or regression is the grouping of the tasks as takes after:
1. Set up the preparation informational index utilizing pre-handling on the raw data.
2. Class attribute and classes are identified.
3. Recognize helpful properties for classification (Relevance investigation).
4. Take in a model utilizing preparing cases in Training set.
5. Utilize the model to characterize the obscure information tests.

TABLE 2 - Comparisons between different Decision Tree Algorithms

Algoritms	**ID3**	**C4.5**	**C5.0**	**CART**
Processing speed	Slow	Better	Faster	Average
Types of data	Categorical	Categorical,Numerical	Categorical,dates,time, Numerical	Numerical,Nominal
Boosting	Not Allowed	Not Allowed	Allowed	Allowed
Pruning	No	Pre-pruning	Post-pruning	Pre-pruning
Measure	Entropy, Information Gain	Information Gain ratio,Split info	Information Gain ratio,Split info	Gini diversity index
Missing values	Not Supported	Not Supported	Supported	Supported

3.1 Issues in decision trees

While modelling the decision tree over-fitting is one of the realistic issues faced. When constructing the model with many branches over-fitting is happens because of the presence of outliers and missing information in the dataset. Consequently it will decreases the training set error and increases the test set error results in test set accuracy of the model decreases. In order to overcome these issues two approaches can be introduced.

1. Pre-pruning
2. Post-pruning

Pre-pruning:

Pruning is the process of reducing the size of decision tree in order to minimize the misclassification error rate.
After checking some performance measures, the process of cut down or halting the sub-branch of decision tree at some node is called pre-pruning. It is also called as early stopping because in this techniques it stops the growing of tree prematurely.

Post-pruning

This method allows the decision tree to its entirety. After that ir trim or removes the nodes of decision tree in a bottom-up procedure. It can be done by replacing the node with leaf.

3.2 Advantages:

1. Interpretable initially
2. Suitable for taking care of both all out and quantitative esteems.
3. Universal for taking care of both classification and regression issues
4. Capable of taking care of missing esteems in properties and filling them in with the most plausible esteem.

5. High-performing with respect to seeking down a fabricated tree, in light of the fact that the tree traversal calculation is proficient notwithstanding for enormous informational collections.
6. Easy to understand.
7. Easy to produce rules

3.3 Disadvanatages:

1. May experience the ill effects of overfitting
2. Classifies by rectangular partitioning
3. Does not deal with non-numeric information
4. Can be very substantial so pruning is important
5. Decision trees can be unsteady.
6. It can be hard to control the measure of the tree.
7. In some unpredictable cases, part information into classes won't not be useful.
8. Information gain is prone to prefer attributes with a large number of different values**.**

3.4 Applications:

- **Web search: positioning page in light of what you are well on the way to tap on.**
- **Computational biology: balanced outline sedates in the PC in view of past tests.**
- **Finance: choose who to send what charge card offers to. Assessment of hazard on layaway offers. Instructions to choose where to contribute cash.**
- **E-trade: Predicting client beat. Regardless of whether an exchange is false.**
- **Space exploration: space tests and radio stargazing.**
- **Robotics: how to deal with vulnerability in new situations. Self-ruling. Self-driving auto.**
- **Information extraction: Ask inquiries over databases over the web.**
- **Social networks: Data on connections and inclinations. Machine figuring out how to separate an incentive from information.**
- **Debugging: Use in software engineering issues like troubleshooting. Work concentrated process. Could recommend where the bug could be.**

4. CONCLUSION

As per our analysis, the efficiency of the decision tree is completely trust in entropy, information gain for classification techniques and standard deviation, standard deviation reduction for prediction techniques. By utilizing the decision tree properly different types of work has been completed. In any case it is fixed or unvarying in character. Some advanced techniques in machine learning cut down the issues like reproduction, managing constant data. This analysis gives some primary key knowledge to the student and researcher about benefits, difficulties and issues of decision tree algorithm.

5.REFERENCES

[1] Jiawei Han And Micheline Kamber,Data Mining Concept and Techniques, Copyright 2006, Second Edition.
[2] Chen Jin, Luo De-lin and mu Fen-xiang An improve ID3 Decision tree algorithm. IEEE 4th International Conference on computer Science & Education.
[3] Devashish Thaku, Nisarga Makandaiah and Sharan Raj D (2010). Re Optimization of ID3 and C4.5 Decision tree. IEEE Computer & Communication Technology.
[4] Gordan.V.Kass(1980). An exploratory Technique for inverstigation large quantities of categorical dataApplied Statics, vol 29, No .2, pp. 119-127.
[5] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone (1984). *Classification and Regression Trees.* Wadsworth International Group, Belmont, California.
[6] Quinlan J. R. (1986). Induction of decision trees. *Machine Learning*, Vol.1-1, pp. 81-106.
[7] Zhu Xiaoliang, Wang Jian YanHongcan and Wu Shangzhuo(2009) Research and application of the improved algorithm C4.5 on decision tree.

[8] Prof. Nilima Patil and Prof. Rekha Lathi(2012), Comparison of C5.0 & CART Classification algorithms using pruning technique.
[9] Baik, S. Bala, J. (2004), A Decision Tree Algorithm For Distributed Data Mining.
[10] Kyu Park, Kyoung Mu Lee and Sang Uk Lee (1999). Perceptual grouping of 3D features in aerial image using decision tree classifier. In Proc. of 1999 International Conference on Image Processing, Vol. 1, pp. 31 - 35.
[11] Sean D. MacArthur, Carla E. Brodley, Avinash C. Kak and Lynn S. Broderick (2002). Interactive content-based image retrieval using relevance feedback. Computer Vision and Image Understanding, pp. 55-75.
[12] Selvam Nallathambi.,Karthikeyan Ramasamy:Prediction of Electricity Consumption based on DT and RF: Application on USA Country Energy Consumption. IEEE conference on Electrical,Instrumentation and communication Engineering (ICEICE) 2017 pages 1-6.
[13] Monanapriya Muthukumar., Selvam Nallathambi.: Remote Sensor Networks for Condition Monitoring: An Application on Railway Industry. IEEE conference on Electrical,Instrumentation and communication Engineering (ICEICE) 2017 pages 1-6.
[14] Selvam Nallathambi, "Analysis of five-level Dc-Dc converter with capacitor for high gain application", International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST), ISSN:2395-695X, vol.3, Special Issue.24, pp. 352-356, March2017