Decision Trees of Machine Learning applied For Adult dataset

Dr.HemaShankari
Associate Professor, Department of Computer Applications,
Women's Christian College, Chennai.

Dr.S.Mathivilasini
Assistant Professor, Department Of Computer Science
Ethiraj College for Women, Chennai.

Dr.S. DilliArasu
Assistant Professor, Department of Computer Application
DRBCCC Hindu College Chennai-600 072

J.Vijayarangam (GF-BITS-PILANI-Chennai)

Abstract-Machine learning is one of the subsets of data mining and its applications are ever growing since the dawn of the decade. It has many tools of analysis at its disposal for any field and decision tree is one such nice tool which can be applied to a numerous list of problems ranging from the fileld of management to fields like Economics, Finance , Banking sector to name a few. This paper is trying to explore the application of the decision tree tool in the Adult dataset from the UCI repository and we employ R programming platform for the analysis.

1.Introduction:

It has become a universal knowledge that Machine Learning is a data based technique and that is why it is one of the most sought after in this era of data mining. Tom Mitchell[1] discusses about Machine Learning in great detail, various aspects of it and he gives the following compact definition, A computer program is said to learn from experienceEwith respect to some class of tasksTand performance measureP, if its performance at tasks in T, as measured by , improves with experience E. (Tom Mitchell, 1998). C.Rudin [2] discusses about Machine Learning for science and society in a nice way. NeeshaJothi[3] discusses data mining in Healthcare in which they talk about various machine learning tools employed in healthcare and Decision tree is one of them. IlyesJenhani[4] discusses in detail the aspects of building a DT as classifiers. In this paper we are exploring the application of Decision trees adult dataset (Table1). This is a simple

classification problem in social science study which has some attributes of humans with the class vector as the gender.

Table1-Adult dataset-sample rows

Age	Work	fnlwgt	Education	Edn-num	Marital Status	Occupation		Race	Salary	Gender
39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical		White	<=50K	Male
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial		White	<=50K	Male
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners		Black	<=50K	Male

We form a decision tree to find whether a person is a male or female based on the values of the feature vectors.We use entropy and information gain for building the decision trees.

For a set , entropy is a measure of impurities in the elements in the set. We can google it as "Number of microscopic configurations that are consistent with the macroscopic quantities that characterize the system" If we are in chemistry domain , an example that would suffice would be the picture below(fig1).

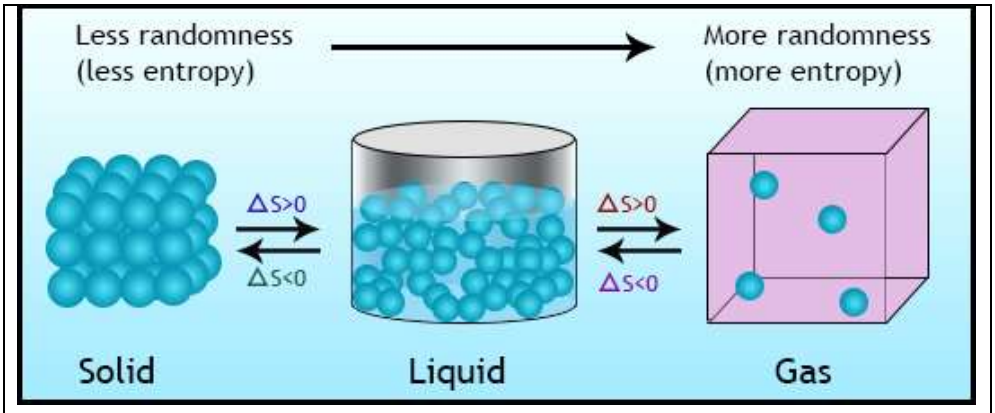

Figure 1 Entropy in Chemistry

It clearly conveys the idea of entropy even from a layman point of view. Shannon is a pioneer in Entropy , from the Information domain. He has a huge part in the theoretical development of entropy models in communication theory. He has given us a nice mathematical formula for calculating entropy for a set.

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

For example, Consider a random variable *x* having 8 possible states, each of which is equally likely. In order to communicate the value of *x* to a receiver, we would need to transmit a message whose length is given by the entropy.

H(x)=-8*[1/8]*log2[1/8]= 3 bits.

Hence send it through 3 bits.

information gain is a measure of the reduction in the overall entropy of a set of instances that is achieved by testing on a descriptive feature.Software packages use the entropy idea in different forms for building a decision tree.

2. Decision tree (DT):

A Decision tree (fig2) is a graphical structure with nodes branching to grow. Based on the position, the nodes are called root node, Intermediate node and leaves or decision nodes. These are applied for classification and regression problems.

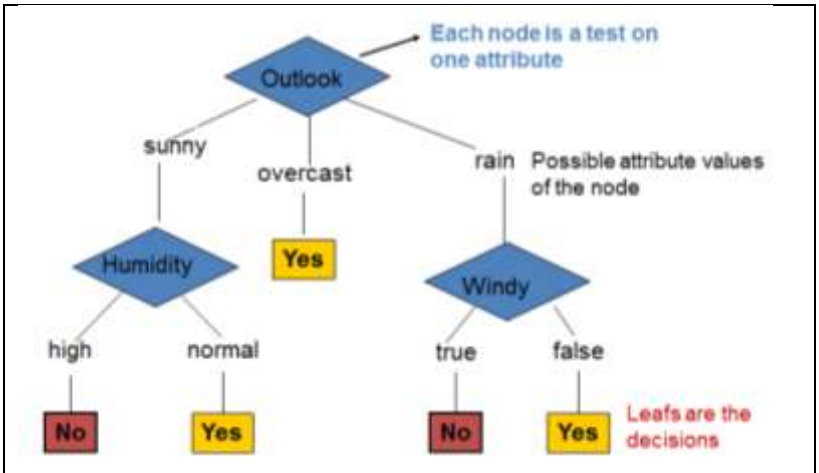

Fig2-Decision tree

The above is a DT for the famous weather dataset (Table 2). It has four feature vectors, Day, Outlook, Humidity and wind and one class vector Play, which is about deciding whether to play golf or not based on the weather conditions. The first branching is using the Outlook feature and so it is called the root node. The next two branches based on Humidity and windy are

intermediate nodes and the final nodes of No and Yes can be called the Leaves and those are the classification values. Any sequence of nodes from the root node to a leaf node is called a rule.

Table 2-Weather dataset-sample rows

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes

3. Build a DT for Adult dataset.

For the present discussion, we are using the adult dataset as described in the introduction. For simplicity we have reduced the number of feature vectors to 5(Table3). The gender is numeric with MALE given by '0' and FEMALE given by '1'.

Table 3- Adult dataset with 5 features and one class vector (Gender).

	Age	Work hrs perweek	Work	Education	Marital status	Gender
1	39	40	0	0	0	0
2	50	13	2	0	1	0
3	38	40	1	2	2	0
4	53	40	1	2	1	0
5	28	40	1	0	1	1
6	37	40	1	1	1	1

When we built a decision tree using R, we obtained the result (fig3)

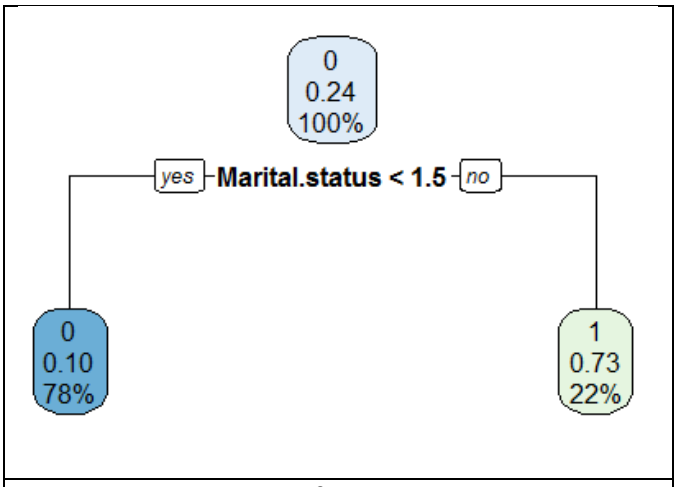

Fig3- DT for Adult data

Just for clarity purpose and Illustration, we also built a decision tree for the inbuilt IRIS dataset in R(fig4).

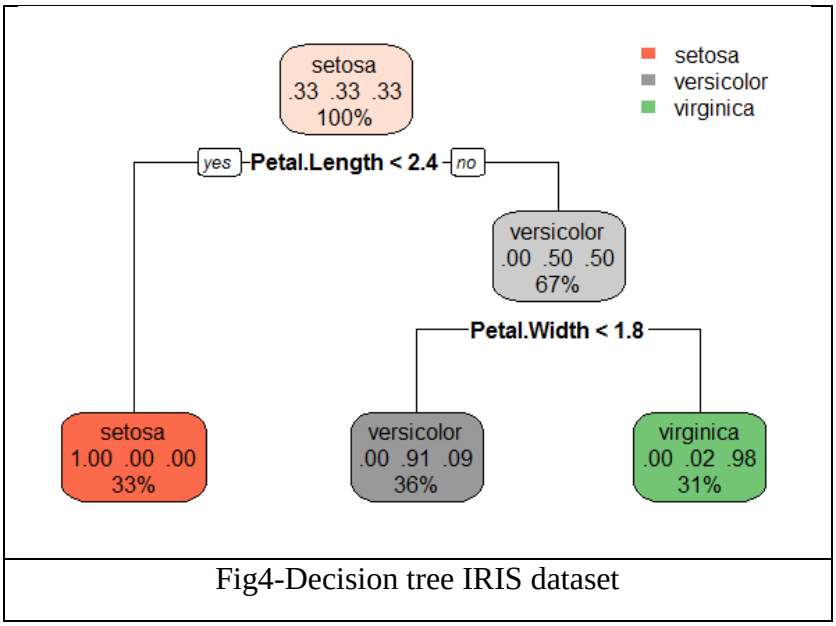

Fig4-Decision tree IRIS dataset

4. Analysis and Conclusion:

From the figure 3 we can deduce that with 50 rows-5 feature vectors-one class vector, the only feature vector which is the optimal one , as deduced by the package, is the Marital status. It says, when the marital status is <1.5, it is a MALE(78% of the data) and otherwise it is FEMALE(22% of the data). The decision to use marital status as the feature vector for branching of the tree is based on the entropy and information gain calculation for all the feature vector vectors and the best one being used for branching. The first chosen feature vector is the root and later features

form the remaining tree with the final nodes without branching are called the leaves. In our tree the gender nodes are the leaves.

Just for comparison, we have built a decision tree for the IRIS dataset in R and it is in figure 4. It is a dataset with four features, the sepal length, sepal width, petal length and petal width and they fall in one of three species. We can clearly see there are two branches , one based on petal length and one based on petal width and in the leaf nodes, we have the class values, Setosa, Versicolor and Virginica. So, we could see that Decision trees are simple to create and are efficient to solve simple classification problems.

References:

1.TomM.Mitchell, Machine Learning, McGrawHill,2017.

2..C.Rudin, KL.Wagstaff, Machine Learning for science and society, 2014, Springer.

3.NeeshaJothi et al., Data mining in healthcare, A review, 2015, Elsevier.

4.IlyesJenhani et al.,Decision trees as possibilistic classifiers, International Journal of Approximate reasoning,volume 48,Issue 3, Aug2008,784-807.

5.RajeevRastogi, KyuseokShim, PUBLIC: A Decision Tree Classifier that Integrates Building and Pruning,Data Mining and Knowledge Discovery, 4, 315–344, 2000.

6.Agrawal, R., Imielinski, T., and Swami, A. 1993. Database mining: A performance perspective. IEEE Transactions on Knowledge and Data Engineering, 5(6):914–925.

7.Agrawal, R., Ghosh, S., Imielinski, T., Iyer, B., and Swami, A. 1992. An interval classifier for database mining applications. In Proc. of the VLDB Conference, Vancouver, British Columbia, Canada, August, pp. 560–573.