An Extended Idea about Decision Trees

Feng-Jen Yang
Department of Computer Science
Florida Polytechnic University
Lakeland, Florida, USA
fyang@floridapoly.edu

***Abstract*—Decision trees have been widely recognized as a data mining and machine learning methodology that receives a set of attribute values as the input and generates a Boolean decision as the output. In this paper, I tried two experiments to demonstrate that the fundamental theory of decision trees can be extended to go beyond Boolean decisions. In the first experiment, the decision tree algorithm is extended to be capable of making three-way decisions. Likewise, in the second experiment, the decision tree algorithm is extended to be capable of making four-way decisions. As a result, a conclusion is reached that this extended idea about decision trees can be generalized to make n-way decisions.**

***Index Terms*—Decision Trees, Boolean Decisions, Tree-Based Classifications**

I. Introduction

A decision tree is a tree-based method in which each path starting from the root is representing a sequence of data splitting until a Boolean outcome is reached at the leaf node. In real-life practice, each path in the decision tree is a decision rule that can be easily translated into human languages or programming languages. By taking all paths (rules) into consideration, the entire tree is correspondent to a compound Boolean expression which involves conjunctions and disjunctions to make Boolean decisions.

The capability of decision support is not new in the studies of data mining and machine learning. Technically there are multiple methods capable of supporting decision makings. Nonetheless, a decision tree might be preferred because of its clearness and understandability. Unlike neural networks and support vector machine that are commonly used as black boxes while making decisions, decision trees are much easier to be explained in human languages.

Although conventionally, decision trees are used for supporting Boolean decisions, in this study, I tried a way to extend and generalize the conventional computations of entropies and information gains to go beyond that scope of Boolean decisions. The successful outcomes of this study are demonstrated by performing two experiments in which the first experiment demonstrates the capability of making three-way decisions and the second experiment demonstrates the capability of making four-way decisions.

II. The Best Splitting of Data at Each Stage

Starting from the root, each internal node within a decision tree is representing an attribute that is selected to split data. At each stage of the tree construction, the attribute that has the maximum splitting power is selected and placed along a path until each resultant subset of data is completely purified, i.e., all data in the subset are leading to the same decision. Mathematically, the adopting of maximum splitting power at each stage is a guarantee of having the shortest possible tree at the end.

Conventionally, a decision tree is used for making Boolean decisions in which the splitting power of an attribute is computed as its information gain that, in turn, is computed as its entropy reduction. In the case of pursuing Boolean decisions, the entropy of a data set is computed as:

$$H(Set) = -P_1 \times log_2 P_1 - P_2 \times log_2 P_2$$

$$where\ P_1\ is\ the\ proportion\ of\ the\ 1st\ decision,$$
$$and\ P_2\ is\ the\ proportion\ of\ the\ 2nd\ decision.$$

The information gain of an attribute is computed as:

$$\begin{aligned}&Gain(A)\\&= H(Set) - (w_1 \times H(a_1) + w_2 \times H(a_2) + ... + w_m \times H(a_m))\end{aligned}$$

$$\begin{aligned}&where\ a_1,\ a_2,\ \dots\ ,\ a_m\ are\ the\ different\ values\ of\ attribute\ A,\\&and\ w_1,\ w_2,\ ...,\ w_m\ are\ the\ weights\ of\ the\ subsets\ split\ by\\&using\ the\ values\ of\ attribute\ A.\end{aligned}$$

For the purpose of making Boolean (two-way) decisions, the entropy is computed by using based-2 logarithm because the resultant decision being pursued is one out of the two possible decisions.

A. Extending the Computations to Support 3-Way Decisions

My approach to extending the mathematical rationale from the support of 2-way decisions to 3-way decisions is quite intuitive. Since the resultant decision being pursued is one out the three possible decisions, the entropy will be computed by using based-3 logarithm.

The entropy of a data set is computed as:

$$H(Set) = -P_1 \times log_3 P_1 - P_2 \times log_3 P_2 - P_3 \times log_3 P_3$$

$$\begin{aligned}&where\ P_1\ is\ the\ proportion\ of\ the\ 1st\ decision,\\&P_2\ is\ the\ proportion\ of\ the\ 2nd\ decision,\\&and\ P_3\ is\ the\ proportion\ of\ the\ 3rd\ decision.\end{aligned}$$

978-1-7281-5584-5/19/$31.00 ©2019 IEEE
DOI 10.1109/CSCI49370.2019.00068

The information gain of an attribute is computed as:

$$
\begin{aligned}
&Gain(A)\\
&= H(Set) - (w_1 \times H(a_1) + w_2 \times H(a_2) + ... + w_m \times H(a_m))
\end{aligned}
$$

where $a_1,\ a_2,\ ...\ ,\ a_m$ *are the different values of attribute* A,
and $w_1,\ w_2,\ ...,\ w_m$ *are the weights of the subsets split by*
using the values of attribute A.

B. Extending the Computations to Support 4-Way Decisions

Based on the same rationale, the computations can be extended to support 4-way decisions. Since now the resultant decision being pursued is one out the four possible decisions, the entropy will be computed by using based-4 logarithm.

The entropy of a data set is computed as:

$$
\begin{aligned}
&H(Set)\\
&= -P_1 \times log_4 P_1 - P_2 \times log_4 P_2 - P_3 \times log_4 P_3 - P_4 \times log_4 P_4
\end{aligned}
$$

where P_1 *is the proportion of the* 1*st decision*,
P_2 *is the proportion of the* 2*nd decision*,
P_3 *is the proportion of the* 3*rd decision*,
and P_4 *is the proportion of the* 4*th decision.*

The information gain of an attribute is still computed in the same manner as:

$$
\begin{aligned}
&Gain(A)\\
&= H(Set) - (w_1 \times H(a_1) + w_2 \times H(a_2) + ... + w_m \times H(a_m))
\end{aligned}
$$

where $a_1,\ a_2,\ ...\ ,\ a_m$ *are the different values of attribute* A,
and $w_1,\ w_2,\ ...,\ w_m$ *are the weights of the subsets split by*
using the values of attribute A.

C. Extending the Computations to Support N-Way Decisions

To be even more generalized, the computations can be extended to support multiple ways of decisions. Since now the resultant decision being pursued is one out the N possible decisions where N $\geq$ 2, the entropy will be computed by using based-n logarithm.

The entropy of a data set is computed as:

$$
\begin{aligned}
&H(Set)\\
&= -P_1 \times log_n P_1 - P_2 \times log_n P_2 - ... - P_n \times log_n P_n
\end{aligned}
$$

where P_1 *is the proportion of the* 1*st decision*,
P_2 *is the proportion of the* 2*nd decision*,
P_3 *is the proportion of the* 3*rd decision*,
and P_n *is the proportion of the nth decision.*

The information gain of an attribute is still computed in the same manner as:

$$
\begin{aligned}
&Gain(A)\\
&= H(Set) - (w_1 \times H(a_1) + w_2 \times H(a_2) + ... + w_m \times H(a_m))
\end{aligned}
$$

where $a_1,\ a_2,\ ...\ ,\ a_m$ *are the different values of attribute* A,
and $w_1,\ w_2,\ ...,\ w_m$ *are the weights of the subsets split by*
using the values of attribute A.

III. The Experiments

As a way of demonstrating the aforementioned computational extensions, the following two experiments are performed and their resultant decision trees are correctly produced. The first experiment is a demonstration that applies this extended idea to create a decision tree for 3-way decision support. The second experiment is a demonstration that applies this extended idea to create a decision tree for 4-way decision support.

A. The First Experiment

In the first experiment a decision tree for 3-way decision is created based on the data set shown in Table 1. Based on the values in attributes A, B and C, this tree-based method will learn the decisions in attribute D.

TABLE I
The Data Set for a 3-Way Decision Support

A	B	C	D
a1	b1	c1	x
a1	b1	c2	y
a1	b1	c2	y
a1	b2	c2	z
a1	b2	c2	z
a1	b2	c2	z
a1	b2	c3	x
a2	b2	c3	y
a2	b2	c3	y
a3	b2	c3	x
a3	b2	c3	y
a3	b2	c3	x

If each different value in attribute D is treated as a particular kind of decision. In the original data set, there are 4 out of the 12 examples belonging to decision x; 5 out of the 12 examples belonging to decision y; and 3 out of the 12 examples belonging to decision z. So, the proportions of these three decisions are 4/12, 5/12 and 3/12 respectively. The entropy of this data set is computed as:

$$
\begin{aligned}
&H(Set)\\
&= -(4/12) \times log_3(4/12) - (5/12) \times log_3(5/12) - (3/12) \times log_3(3/12)\\
&= 0.980834
\end{aligned}
$$

To compute the information gain of attribute A. The original data set has to be grouped by using the values of attribute

A. This grouping results in three subsets in which the 1st subset consists of those examples in which A = a1; the 2nd subset consists of those examples in which A = a2; and the 3rd subset consists of those examples in which A = a3. These three subsets are shown in Table 2, Table 3 and Table 4.

TABLE II
THE 1ST SUBSET GROUPED BY ATTRIBUTE A

A	B	C	D
a1	b1	c1	x
a1	b1	c2	y
a1	b1	c2	y
a1	b2	c2	z
a1	b2	c2	z
a1	b2	c2	z
a1	b2	c3	x

TABLE III
THE 2ND SUBSET GROUPED BY ATTRIBUTE A

A	B	C	D
a2	b2	c3	y
a2	b2	c3	y

TABLE IV
THE 3RD SUBSET GROUPED BY ATTRIBUTE A

A	B	C	D
a3	b2	c3	x
a3	b2	c3	y
a3	b2	c3	x

There entropy of these three subsets are computed as follows:

$$\begin{aligned}&H(a1)\\&= -(2/7) \times log_3(2/7) - (2/7) \times log_3(2/7) - (3/7) \times log_3(3/7)\\&= 0.982141\end{aligned}$$

$$\begin{aligned}&H(a2)\\&= -0 - (2/2) \times log_3(2/2) - 0 = 0\end{aligned}$$

$$\begin{aligned}&H(a3)\\&= -(2/3) \times log_3(2/3) - (1/3) \times log_3(1/3) - 0\\&= 0.57938\end{aligned}$$

Since 7 out of the 12 cases are grouped into the 1st subset, 2 out of the 12 cases are grouped into the 2nd subset, and the rest of the 3 out the 12 cases are grouped into the 3rd subset, the weights for these thee subsets are 7/12, 2/12 and 3/12 respectively. The information gain of attribute A is, thus, computed as:

$$\begin{aligned}&Gain(A)\\&= H(Set) - [(7/12) \times H(a1) + (2/12) \times H(a2) + (3/12) \times H(a3)]\\&= 0.263073\end{aligned}$$

To compute the information gain of attribute B. The original data set has to be grouped by using the values of attribute B. This grouping results in two subsets in which the 1st subset consists of those cases in which B = b1, and the 2nd subset consists of those cases in which B = b2. These two subsets are shown in Table 5 and Table 6.

TABLE V
THE 1ST SUBSET GROUPED BY ATTRIBUTE B

A	B	C	D
a1	b1	c1	x
a1	b1	c2	y
a1	b1	c2	y

TABLE VI
THE 2ND SUBSET GROUPED BY ATTRIBUTE B

A	B	C	D
a1	b2	c2	z
a1	b2	c2	z
a1	b2	c2	z
a1	b2	c3	x
a2	b2	c3	y
a2	b2	c3	y
a3	b2	c3	x
a3	b2	c3	y
a3	b2	c3	x

There entropy of each subset and the information gain of attribute B are computed as follows:

$$\begin{aligned}&H(b1)\\&= -(1/3) \times log_3(1/3) - (2/3) \times log_3(2/3) - 0\\&= 0.57938\end{aligned}$$

$$\begin{aligned}&H(b2)\\&= -(3/9) \times log_3(3/9) - (3/9) \times log_3(3/9) - (3/9) \times log_3(3/9)\\&= 1\end{aligned}$$

Since 3 out of the 12 cases are grouped into the 1st subset, and 9 out of the 12 cases are grouped into the 2nd subset, the weights for these two subsets are 3/12 and 9/12 respectively. The information gain of attribute B is, thus, computed as:

$$\begin{aligned}&Gain(B)\\&= H(Set) - [(3/12) \times H(b1) + (9/12) \times H(b2)]\\&= 0.085989\end{aligned}$$

In the same manner, to compute the information gain of attribute C. The original data set has to be grouped by using the values of attribute C. This will result in three subsets in which the 1st subset consists of those cases in which C = c1; the 2nd subset consists of those cases in which C = c2; and the 3rd subset consists of those cases in which C = c3. These three subsets are shown in Table 7, Table 8 and Table 9.

TABLE VII
THE 1ST SUBSET GROUPED BY ATTRIBUTE C

A	B	C	D
a1	b1	c1	x

TABLE VIII
THE 2ND SUBSET GROUPED BY ATTRIBUTE C

A	B	C	D
a1	b1	c2	y
a1	b1	c2	y
a1	b2	c2	z
a1	b2	c2	z
a1	b2	c2	z

There entropy of each subset and the information gain of attribute C are computed as follows:

$$H(c1)$$
$$= -(1/1) \times log_3(1/1) - 0 - 0 = 0$$

$$H(c2)$$
$$= -0 - (2/5) \times log_3(2/5) - (3/5) \times log_3(3/5) = 0.612602$$

$$H(c3)$$
$$= -(3/6) \times log_3(3/3) - (3/6) \times log_3(3/6) - 0 = 0.63093$$

Since 1 out of the 12 cases are grouped into the 1st subset; 5 out of the 12 cases are grouped into the 2nd subset; and the rest of the 6 out the 12 cases are grouped into the 3rd subset, the weights for these thee subsets are 1/12, 5/12 and 6/12 respectively. The information gain of attribute C is, thus, computed as:

$$Gain(C)$$
$$= H(Set) - [(1/12) \times H(c1) + (5/12) \times H(c2) + (6/12) \times H(c3)]$$
$$= 0.665369$$

At this stage, attribute C has the highest information gain and, thus, chosen to be the root of the tree as shown in Figure 1.

The construction of this decision tree is continued along the branch of C=c2, i.e. to further split the data in Table 8. Within this branch if the data are grouped by using the values of attribute A, the Table 8 is not further divided. All of the 5 cases are staying in the same subset. So, the in formation gain of attribute A along this branch is zero, i.e., $Gain(A) = 0$.

TABLE IX
THE 3RD SUBSET GROUPED BY ATTRIBUTE C

A	B	C	D
a1	b2	c3	x
a2	b2	c3	y
a2	b2	c3	y
a3	b2	c3	x
a3	b2	c3	y
a3	b2	c3	x

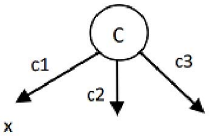

Fig. 1. After C is chosen to be the root.

Along the same branch of C=c2, if the data are grouped by using the values of attribute B, the Table 8 is further divided into two subsets in which the 1st subset consists of those cases in which B = b1, and the 2nd subset consists of those cases in which B = b2. These two subsets are shown in Table 10, and Table 11, an the information gain of attribute B along this branch is computed as $Gain(B) = 0.612602$.

TABLE X
THE 1ST SUBSET WITHIN THE BRANCH OF C=C2 GROUPED BY ATTRIBUTE B

A	B	C	D
a1	b1	c2	y
a1	b1	c2	y

TABLE XI
THE 2ND SUBSET WITHIN THE BRANCH OF C=C2 GROUPED BY ATTRIBUTE B

A	B	C	D
a1	b2	c2	z
a1	b2	c2	z
a1	b2	c2	z

At this stage, within the branch of C=c2, attribute B has the highest information gain and, thus, chosen to be placed along this branch as shown in Figure 2.

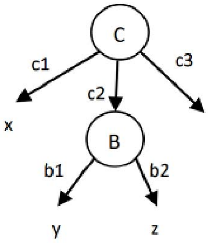

Fig. 2. After B is placed along the path of C=c3.

In the same manner, the construction of this decision tree is continued along the branch of C=c3, i.e. to further split the data in Table 9. Within this branch if the data are grouped by using the values of attribute A, the Table 9 is further divided into three subsets in which the 1st subset consists of those cases in which A = a1; the 2nd subset consists of those cases in which A = a2; and the 3rd subset consists of those cases in which A=a3. These three subsets are shown in Table 12, Table 13, and Table 14, and the information gain of attribute A along this branch is computed as $Gain(A) = 0.63093$.

TABLE XII
THE 1ST SUBSET WITHIN THE BRANCH OF C=C3 GROUPED BY ATTRIBUTE A

A	B	C	D
a1	b2	c3	x

TABLE XIII
THE 2ND SUBSET WITHIN THE BRANCH OF C=C3 GROUPED BY ATTRIBUTE A

A	B	C	D
a2	b2	c3	y
a2	b2	c3	y

TABLE XIV
THE 3RD SUBSET WITHIN THE BRANCH OF C=C3 GROUPED BY ATTRIBUTE A

A	B	C	D
a3	b2	c3	x
a3	b2	c3	y
a3	b2	c3	x

Along the same branch of C=c3, if the data are grouped by attribute B, the Table 9 is not further divided. All of the 6 example are staying in the same table the in formation gain of attribute A along this branch is zero, i.e., $Gain(B) = 0$.

At this stage, within the branch of C=c3, attribute A has the highest information gain and, thus, chosen to be placed along this branch as shown in Figure 3. Since each path is now leading to a particular kind of decision, the entire decision tree is completely constructed.

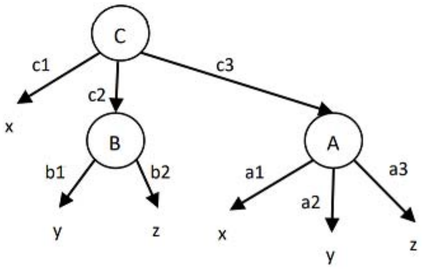

Fig. 3. The eventual 3-way decision tree.

B. The Second Experiment

In the Second experiment a decision tree for 4-way decision is created based on the data set shown in Table 14. Based on the values in attributes A, B and C, this tree-based method will learn the decisions in attribute D.

If each different value in attribute D is treated as a particular kind of decision. The proportions of these four decisions are 2/8, 2/8, 2/8, and 2/8 respectively. The entropy of this data set is computed as:

TABLE XV
THE DATA SET FOR A 4-WAY DECISION SUPPORT

A	B	C	D
a1	b1	c1	w
a1	b1	c1	w
a1	b2	c1	x
a1	b2	c1	x
a2	b2	c1	y
a2	b2	c1	y
a2	b2	c2	z
a2	b2	c2	z

$$H(Set)$$
$$= -(2/8) \times log_4(2/8) - (2/8) \times log_4(2/8) - (2/8) \times log_4(2/8)$$
$$= 1$$

To compute the information gain of attribute A, the original data set has to be grouped by using the values of attribute A. This grouping results in two subsets in which the 1st subset consists of those cases in which A = a1, and the 2nd subset consists of those cases in which A = a2. The entropies of these two subsets and the information gain of attribute A are computed as:

$$H(a1) = -(2/4) \times log_4(2/4) - (2/4) \times log_4(2/4) - 0 - 0$$
$$= 0.5$$

$$H(a2) = -0 - 0 - (2/4) \times log_4(2/4) - (2/4) \times log_4(2/4)$$
$$= 0.5$$

$$Gain(A) = H(Set) - [(4/8) \times H(a1) + (4/8) \times H(a2)]$$
$$= 0.5$$

To compute the information gain of attribute B, the original data set has to be grouped by the using values of attribute B. This grouping results in two subsets in which the 1st subset consists of those cases in which B = b1, and the 2nd subset consists of those cases in which B = b2. The entropy of these two subsets and the information gain of attribute B are computed as:

$$H(b1) = -(2/2) \times log_4(2/2) - 0 - 0 - 0$$
$$= 0$$

$$H(b2) = -0 - (2/6) \times log_4(2/6) - (2/6) \times log_4(2/6) - (2/6) \times log_4(2/6)$$
$$= 0.972481$$

$$Gain(B) = H(Set) - [(2/8) \times H(b1) + (6/8) \times H(b2)]$$
$$= 0.405639$$

To compute the information gain of attribute C, the original data set has to be grouped by using the values of attribute C.

This grouping results in two subsets in which the 1st subset consists of those cases in which C = c1, and the 2nd subset consists of those cases in which C = c2. The entropies of these two subsets and the information gain of attribute C are computed as:

$$H(c1) = -(2/6) \times log_4(2/6) - (2/6) \times log_4(2/6) - (2/6) \times log_4(2/6) - 0$$
$$= 792481$$

$$H(c2) = -0 - 0 - 0 - (2/2) \times log_4(2/2)$$
$$= 0$$

$$Gain(C) = H(Set) - [(2/8) \times H(c1) + (6/8) \times H(c2)]$$
$$= 0.405639$$

At this stage attribute A has the highest information gain and, thus, chosen to be the root of the tree as shown in Figure 4.

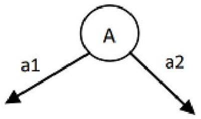

Fig. 4. After A is chosen to be the root.

Along the branch of A = a1, if the data are grouped by using the values of attribute B, the set is further divided into two subsets. The entropy of each subset and the information gain of attribute B within this branch is computed as:

$$H(b1) = -(2/2) \times log_4(2/2) - 0 - 0 - 0 = 0$$

$$H(b2) = -0 - (2/2) \times log_4(2/2) - 0 - 0 = 0$$

$$Gain(B) = H(Set) - [(2/4) \times H(b1) + (2/4) \times H(b2)]$$
$$= 0.5$$

Along the same branch of A=a1, if the data are grouped by using the values of attribute C, the set is not further divided. So the information gain of attribute C within this branch is zero, i.e., $Gain(C) = 0$.

At this stage, within the branch of A=a1, attribute B has the highest information and, thus, chosen to be placed along this branch, as shown in Figure 5.

Along the branch of A = a2,if the data are grouped by attribute B, the set is not further divided. So the information gain of attribute B within this branch is zero, i.e., $Gain(B) = 0$.

Along the same branch of A=a2, if the data are grouped by attribute C, the set is further divided into two subsets. the entropy of each subset and the information gain of attribute C is computed as:

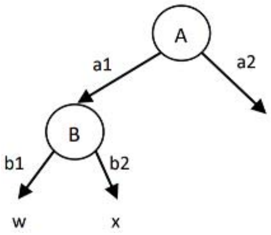

Fig. 5. After B is placed along the path of A=a1.

$$H(c1) = -0 - 0 - (2/2) \times log_4(2/2) - 0 = 0$$

$$H(c2) = -0 - 0 - 0 - (2/2) \times log_4(2/2) = 0$$

$$Gain(C) = H(Set) - [(2/4) \times H(c1) + (2/4) \times H(c2)]$$
$$= 0.5$$

At this stage, within the branch of A=a2, attribute C has the highest information and, thus, chosen to be placed along this branch. Eventually, the decision tree is shown in Figure 6. Since each path is now leading to a particular kind of decision, the entire decision tree is completely constructed.

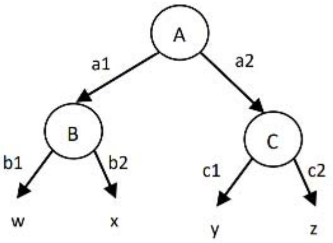

Fig. 6. The eventual 4-way decision tree.

IV. Conclusion

This study was inspired by a personal curiosity about tree-based data splitting. In the theory of tree structures, there are no limitations neither on the number of branches that a tree node could have nor on the number of different values that a leaf node could have. However, the discussions in most of textbooks of artificial intelligence, data mining, and machine learning are limited to the application of using decision trees for Boolean decisions only [1, 2, 3]. After accumulating enough curiosity, I started to try out this extended idea about decision trees and found that the mathematical rationale could be generalized to support more than Boolean decisions.

References

[1] M. Negnevitsky, ”Artificial Intelligence: A Guide to Intelligent Systems,” 3rd Ed., Pearson Education Canada, 2011.

[2] H. Witten, E. Frank and M. A. Hall, ”Data Mining: Practical Machine Learning Tools and Techniques,” 3rd Ed., Morgan Kaufmann, 2011.

[3] P. Flach, ”Machine Learning: The Art and Science of Algorithms that Make Sense of Data,” 1st Ed., Cambridge University Press, 2012.