Systematic Analysis and Application Prospect of Decision Tree

JunYi Xu *

Department of Automation Science and Engineering, South China University of Technology, Guangdong, 510641, China

* Corresponding Author Email: 202130461854@mail.scut.edu.cn

Abstract. Decision making is common practice for everyone. One must make multiple decisions to move on during his/her lifetime. People are always eager to make the optimum decisions so that they could save energy and step in a right path. Although people try to avoid inferior options that comes with risk and danger, things happen from time to time. In this study, a powerful tool, decision tree, will be introduced to address this problem. With assistance of decision tree, one will make better choices more validly and more efficiently. Some concepts and algorithms of decision tree will be also included to understand examples from the application part. The purpose of this study is to introduce the concept of decision tree, analyze the advantages of decision tree and discuss its future. Although decision tree is widely utilized in many aspects in society, it still has shortcomings like overfitting and underfitting. Fortunately, there are methods such as pruning and random forest to solve these problems. The future of decision tree will be promising.

Keywords: Decision tree, Algorithm, Application.

1. Introduction

Nowadays, people can feel the influence of artificial intelligence in every aspect in their daily routine. With the rapid advancement of AI algorithms, applications of AI theory can be seen in many aspects like estimation of investment risk, analysis of data and so on. Assisting the process of decision making is one of them. Proper decision is vital because it can reduce input and yield ideal output, saving energy and time, avoiding certain risk. That is the exact reason why people devise decision tree, a convenient as well as efficient diagram to improve one's decision-making abilities and prevent adverse consequences [1].

In this study, the history of decision tree will be covered firstly to provide its process of development. Basic concept of decision tree will come next to offer understanding of the term "decision tree" and its fundamental structure and components. Also, comprehension of work pattern of decision tree will be included as well. In following section, interpretations of several classic and most widely used algorithms of decision tree (ID3, C4.5, CART) will be introduced. The process of calculation and principles of determination of decisions will be stressed in this section. The study will also include the discussion of application in three industries (agriculture, medicine, and education) and analyze its significance. Although decision tree has extensive application, there are many issues to settle. The author will propose some solutions of these problems in theoretic way. The purpose of the study is to introduce the concepts, principle and application of decision tree and discuss its prospects.

2. Analysis of decision tree

2.1. The history of decision tree

The history portrays decision trees as an effective tool for analyzing data with a user-friendly tree-like layout that improves comprehension and disseminates data. Decision trees were first used during the formation of written records. Porphyry, a Greek philosopher, created the oldest decision tree called Porphyrian tree, whose form dates to the third century C.E.

Computational decision trees emerged in 20th century, when rapid advances in Artificial Intelligence and Statistical Analysis were witnessed. The first perspective of decision analysis based

on probability and utility was created by Frank P. Ramsey in 1931 [2]. In 1950s, Dr. William Belson conducted surveys on all British citizens representing the British Broadcasting Corporation, which is the roots of modern decision trees [3]. John von Neumann and Oscar Morgenstern made significant contributions to the formalization of the decision theory in 1954. It is known as "expected utility hypotheses" when they assign certain values to various options in order to calculate utility so that the decision-maker can choose the option with the greater value with ease. In 1950, Abraham Wald used game theory techniques to make good statistical decisions. The ideas of Wald framework were endorsed by a large number of other mathematicians of that time. Then, by merging the theories created by Neumann and Wald, David Blackwell, M. A. Girshick, and L. J. Savage made an important advancement [2]. Nowadays, decision trees aid people to analyze future samples in multiple situations where descriptions of certain objects are nonmetric features and provide optimum results according to existing training samples or data set. The timeline of decision tree is shown in Figure 1.

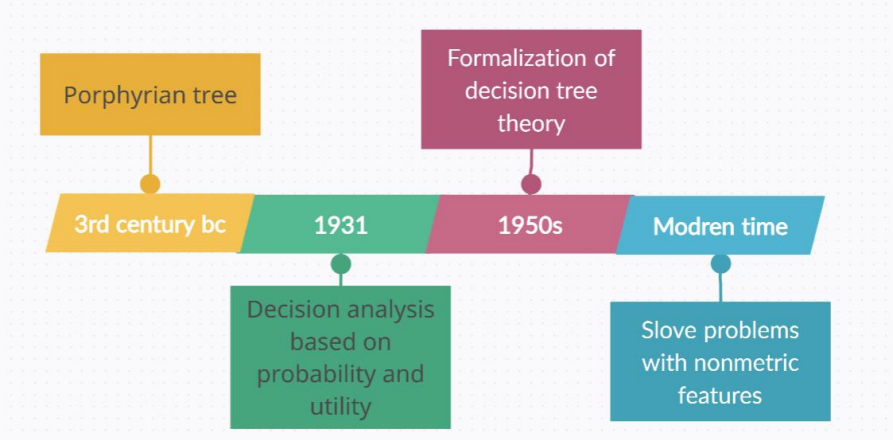

Fig. 1 Timeline of decision tree (Photo/Picture credit: Original)

2.2. Basic concept of decision tree

In order to create categorization models that closely reflect human reasoning and are simple to comprehend, decision tree techniques have been used extensively [4]. The term “decision tree” comes from the resemblance of its structure to branches and trunks of a tree. People tend to make decisions with certain tree-like foundation when solving problems with nonmetric feature. Therefore, decision tree emerged. Unlike traditional problem-solving methods, a decision tree gives a “visual” means of recognizing certain choices or decisions that leads to uncertain outcomes [5]. The purpose of building up decision trees is to set rules to make decisions by learning from certain training samples and data.

The structure of a decision tree contains root node, internal node, leaf node and branch. All the nodes are decision-making point. The root node is the beginning of a tree. The internal nodes are the nodes after the root node and before the leaf node. The leaf node is the node that does not have further splits. The branch is the connection between nodes. The diagram is as follows in Figure 2. It can be said that the work pattern of decision tree is “IF-ELSE” (If conditions of node A is met, choose node A, else consider node B).

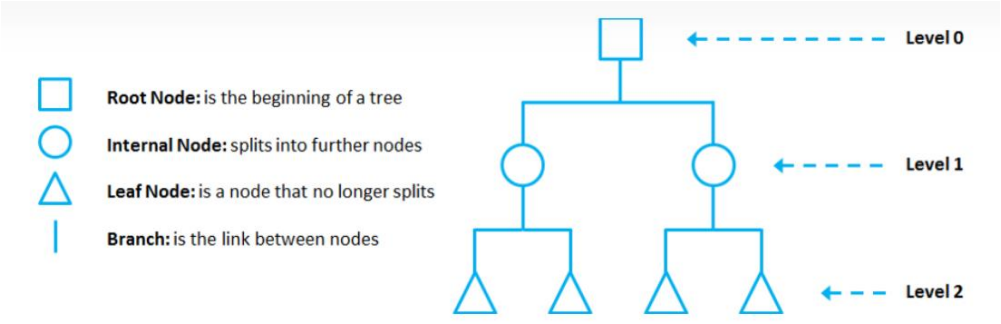

Fig. 2 Structure of decision tree (Photo/Picture credit: Original)

2.3. ID3

ID3 (interactive dichotomizer-3) is the most outstanding and famous method to build decision trees in early phase. Although its term contains dichotomy, it still works when each node splits into more than two further nodes. The prototype of ID3 is concept learning system proposed by Hunt et al, which selects the most highly discernable features to divide data until each leaf node has single type of the data.

Since the foundation of ID3 algorithm is entropy in Shannon information theory, the introduction of entropy, or information entropy according to Shannon information theory is indispensable. Information entropy is used to solve the problem of quantification of information, and by performing specific calculations on the initial confusing data and concept, the precise information entropy is obtained. To put it more simply, information entropy is used to characterize the degree of uncertainty in information, and the lower the entropy, the lower the degree of uncertainty, and the simpler it is for people to obtain specific information.

The interpretation of ID3 algorithm is as follow. For a training set D and feature a (where the number of features are integer v and its value could be $\{a_1, a_2, \ldots, a_v\}$), it can be assumed that the proportion of class i samples in sample set D is p_i (i=1,2,…n), the information entropy is

$$\mathrm{Ent}(D) = -\sum_{i=1}^{|n|} p_i \log_2 p_i \tag{1}$$

$\mathrm{Ent}(D)$ also means entropy impurity. In other words, the less the value of $\mathrm{Ent}(D)$ is, the purer the sample D is. Next, the calculation of conditional information entropy is needed, the result is

$$\mathrm{Ent}(D \mid a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} \mathrm{Ent}(D^v) \tag{2}$$

Here, D^v means the total number of samples that belongs to feature a_v. Finally, information gain is needed to make an optimum decision.

$$\mathrm{Gain}(D, a) = \mathrm{Ent}(D) - \mathrm{Ent}(D \mid a) \tag{3}$$

In most cases, the results of information gain are multiple. The feature with highest information gain will be the first node (root node) because highest value of information gain means greatest reduction in entropy impurity. Then, the second-round calculation will begin with feature (with highest information gain from the first round) removed from sample D and so on until all the features are set as certain nodes.

2.4. C4.5

C4.5 is the improved version of ID3, it solves some problems with ID3. One major problem with ID3 is that if the classifications of single feature are numerous, that is, there are too many forks, the information gain will be extremely high. In this way, ID3 algorithm will believe the classification is correct. Indeed, the classification is true but only restricted to the known samples, it will be sophisticated for ID3 to provide appropriate prediction of future samples. Therefore, C4.5 was revised to deal with this issue.

The interpretation of C4.5 algorithm is as follows. Instead of using information gain, C4.5 utilizes gain ratio to determine the nodes and introduce the “intrinsic value” of feature a to its calculation. The greater the number of feature a is, the bigger “intrinsic value” is.

First, calculation of information gain is needed.

$$\mathrm{Gain}(D, a) = \mathrm{Ent}(D) - \mathrm{Ent}(D \mid a) \tag{4}$$

Then, calculate intrinsic value.

$$\mathrm{IV}(a) = -\sum_{v=1}^{V} \frac{|D^v|}{|D|} \log_2 \frac{|D^v|}{|D|} \tag{5}$$

In the end, calculate gain ratio according to the results of information gain and gain ratio.

$$\text{GainRatio}\ (D,a) = \frac{\text{Gain}\ (D,a)}{IV(a)} \tag{6}$$

2.5. CART

CART (classification and regression tree) is another algorithm which shares the same core ideas with ID3 and C4.5. The main difference is that CART conducts dichotomy at every node, that is, each node only has two child nodes (each node only has two branches, except leaf nodes), forming a binary tree at last. Besides, not only can CART solve classification problem, but also it can construct regression tree to regress continuous variables.

CART uses Gini coefficient to divide attributes and features. Specifically speaking, Gini coefficient indicates the possibility of two samples taken out randomly from the dataset that have different category marks (e.g., the probability of randomly selecting two emails out of 100 that are "spam" and "not spam") The lower the Gini index, the higher the purity of the data set *D*.

The interpretation of C4.5 algorithm is as follows. First, calculate Gini index.

$$\text{Gini}\ (D) = -\sum_{i=1}^{|n|} \quad \sum_{i' \neq i} \quad p_i p_{i'} = 1 - \sum_{i=1}^{|n|} \quad p_i^2 \tag{7}$$

Here, $p_{i'} = (1 - p_i)$. Next, calculate conditional Gini index.

$$\text{GiniIndex}\ (D,a) = -\sum_{v=1}^{V} \quad \frac{|D^v|}{|D|}\text{Gini}\ (D^v) \tag{8}$$

At last, choose the feature with the smallest conditional Gini index among feature *a*.

3. Application

3.1. Application in agriculture

Decision tree is very helpful in agriculture since it can help farmers to decide where to cultivate, what and when to plant and the division of the farmland for various crops. Most farmers used to conduct cultivation according to their experience through years of farming and planting, which is lack of reasoning. But with decision trees, farmers can notice the reasons behind different decisions and learn much knowledge from the decision-making rules. Most importantly, decision trees can help farmers to address professional problems they have never encountered by aiding them to choose proper options.

For instance, it is difficult for farmers to determine the quality of soil elaboratively, which will cause the problem of wasting potential land. To make sure every inch of land is fully utilized, decision trees can be constructed to predict, for example, the density of heavy metal in soil from different areas to divide farmland into cultivable and uncultivable.

This work has done by Wei Cheng et al. In their study, to anticipate the spatial distribution of soil heavy metal levels in the severely contaminated area, the classification and regression tree (CART) has been incorporated into ArcGIS using ArcObjects and Visual Basic for Application (VBA) [3]. They took dataset of quality of soil in Fuyang County, situated at the north of Zhejiang Province. The overall CART accuracy of assigning samples to the right Pb (a type of heavy metal element) classes is 89.62% and 85.71%, which indicates decision trees in this model make relatively proper prediction about the distribution of soil with higher density of heavy metal [6]. This application can be used in different areas so long as local soil information is acquirable.

Another application is to boost the development of agriculture by aiding the sponsors to make beneficial investment. As it is illustrated by JingGui Lu et al, they analyze Lin Xian's agricultural output value data of 20 years and set certain rules for agriculture development [7]. The decision tree based on decision making rules is shown in Figure 3 (Area here means planting area, Frost-free and Sunshine means duration of these two kinds of weather).

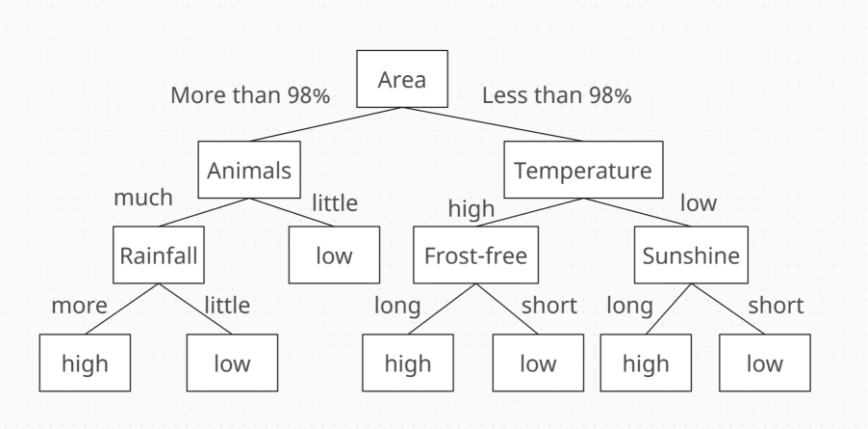

Fig 3. Reference decision tree based on data in Lin Xian [7]

The largest influence on agricultural output value may be seen to be from the planting area, which is congruent with reality. The value of agricultural produce will invariably be reduced if there is no sowing area, regardless of how much influence other factors have [7].

Decision trees can help farmers to make optimum options for their cultivation, directly influencing individuals in agriculture and getting the utmost out of land. In addition, decision trees assist sponsors with their investment in agriculture to make sure it is relatively most profitable. Meanwhile, profitable land also represents excellent agricultural output. Under this circumstance, decision trees create a win-win situation. Besides, prediction of soil quality can call for protection of land and treatment of polluted soil, which is beneficial to the advancement of agriculture.

3.2. Application in medicine

In medicine, it is significant for doctors to inspect the results of body examination to diagnose. They need to know about the value of certain index to see what kind of disease a patient has. The threshold value of the index is determined by data mining (a process of knowledge discovery in databases to discover the useful information from the existing database). Decision trees can make excellent classification so it can be used in this aspect.

Research conducted by Meng-Chang Tsai et al took samples of Breast Tissue, Haberman's Survival, Liver Disorders, Parkinson, Pima Indians Diabetes as data set and compare C4.5 with other two algorithms (back propagation neural network, BPNN and support vector machine SVM), the result in Table 1 shows that C4.5 has a more accurate result overall. Therefore, decision trees are suitable to detect various features of different diseases [8]. The concrete data are shown in Table 1.

Table 1. The percentage of average classification accuracy of the 3 data sets

	BPNN	SVM	C4.5
Breast Tissue	21.60	51.82	60.35
Haberman's Survival	73.50	73.50	70.91
Liver Disorders	57.97	58.26	64.93
Parkinsons	75.38	86.15	82.56
Pima Indians Diabetes	65.10	76.65	76.22
Average Accuracy	58.71	69.28	70.99

Decision trees can also be advantageous to diagnose and specific treatments. In their decision-making process for the management of chronic wounds, Letourneau et al. adopted a decision tree technique. Two groups of home care nurses in significant urban areas provided the information for the study. When the decision tree was first introduced to one group, it was measured immediately; for the other group, it was measured two years later. It was combined with visual case studies and the chronic wound management decision tree (CWMDT). They conclude that a decision tree can help with decision-making by directing the nurse through assessment and treatment alternatives [9].

To conclude, decision trees are efficient when it comes to analyzing data from patients and diagnose and steps of treatment. The process of data mining with utilization of decision trees can be regarded as decision making by setting proper range of certain index and classifying different diseases. As for diagnose and treatment, it is believable that people with little or even no medical knowledge can make proper assessment of current situation and cure patients or themselves by using decision tree.

3.3. Application in education

When it comes to education, individualized teaching will be more effective since every student has his/her strength and weakness. However, it is hard for educators to make accurate evaluations. Decision trees can analyze students' performance in different subjects to make education plans for each of them.

Chen LiXia et al took back 2184 copies of questionnaire which collected scores in mid-exam and final exam from different students. They considered that ID3 algorithm cannot process continuous attributes and divided the scores into several grade sections. "Because of the particularity of admission scores, we consider the score to be 'excellent' when it is above 75, to be 'good' when it is from 65 to 74, to be 'pass' when it is from 50 to 64 to be 'fail' when it is below 50" [10]. They used admission scores as dependent variable and sexuality as independent variable to draw a decision tree with SPSS (Statistical Package for The Social Sciences) [10]. The decision tree is shown in Figure 4.

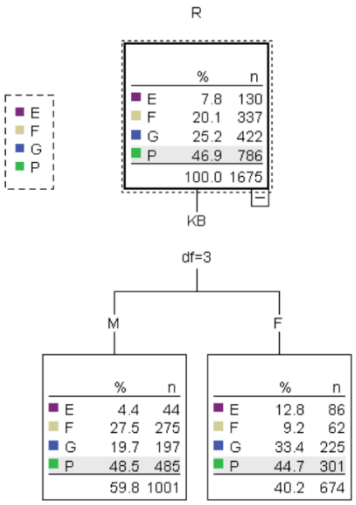

Fig. 4 Decision tree in SPSS [10]

When the sexuality is male, reference decision tree is shown in Figure 5.

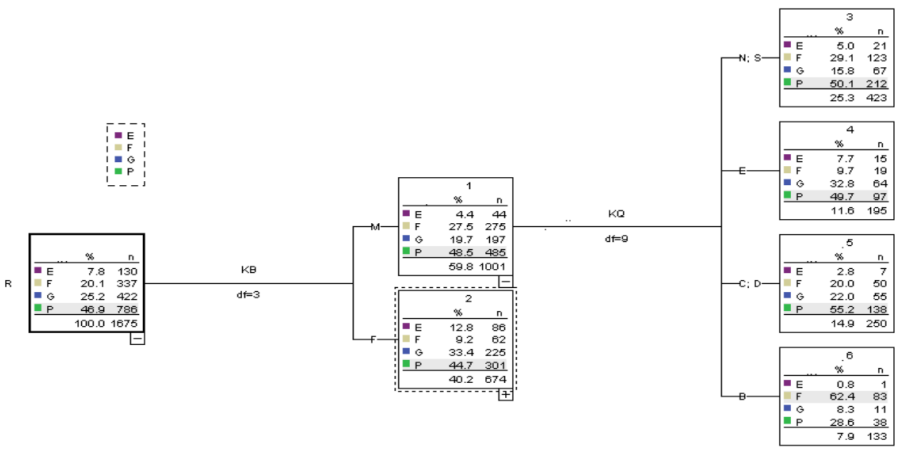

Fig. 5 Decision tree based on the data from male [10]

When the sexuality is female, reference decision tree is shown in Figure 6.

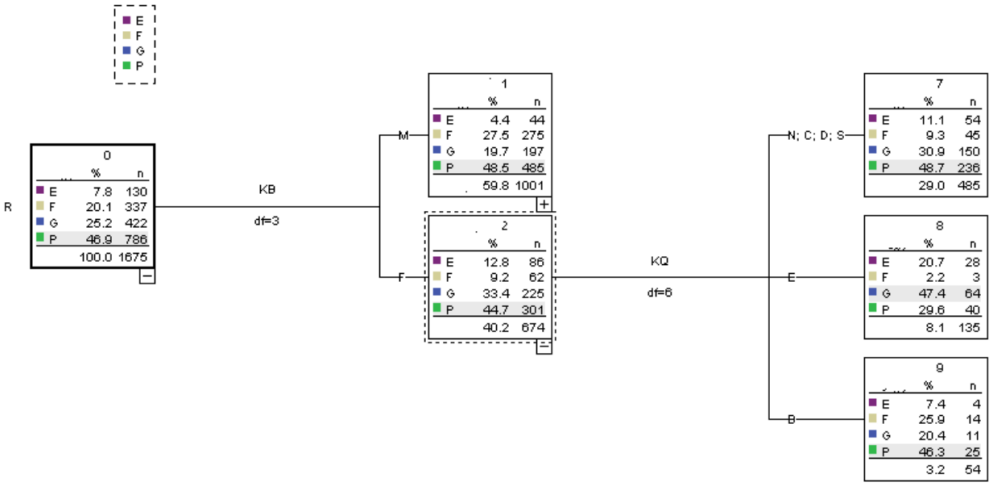

Fig. 6 Decision tree based on the data from female [10]

From the pictures, it is indicated that the students in Eastern China have higher English level. And students in North-western China are obviously weak at English. Generally, girls have higher scores than boys [10].

Based on the models of decision tree above, it can be concluded that decision tree is of considerable significance to make individualization of education possible, since it uncovers the potential of students and make them realize what their strengths and weaknesses is. Meanwhile, educator can formulate proper education policy according to different circumstances like economy standard, number of students or average literacy level.

4. Conclusion

Humankind started to make different decisions from the very beginning of history. However, people desire the best choice all the time so they need to have efficient tools to cope with the problem. Decision tree is one of them. It is created to divide samples and help people to make optimum options. Researchers and scholars apply the theory and principle of decision tree to practice to solve multiple problems in many aspects. However, these algorithms will have some problems like overfitting and underfitting, which does not meet the requirement of prediction of future samples or even the basic need to classify the known samples. Fortunately, there are ways to solve the problems, such as pruning, which is a process of elimination of surplus leaf nodes to accelerate the process of decision making and random forest that consists of addition of plenty decision to tree to field a choice with highest votes. Besides, combination of decision tree and other algorithms like PSO (particle swarm optimization) speeds up the process of making decisions with superior accuracy compared to single decision tree algorithm. In conclusion, decision tree will have greater influence in daily life and further utilization in every industry in near future.

References

[1] Rachel Cravit. What is a Decision Tree and How to Make One. Venngage, 2023

[2] Mithun Sridharan, Harshita Thakur, Ashutosh Singh. Decision trees – a Robust decision-making framework. Think Insights, 2023

[3] Kapil Mittal, Dinesh Khanduja, Puran Chandra Tewari. An insight into “decision tree analysis”. WWJMRD, 2017, 3(12): 111-115.

[4] Serka. B. Kotsiantis. Decision trees: a recent overview. Springer link, 2011

[5] Letícia Fonseca. Decision Tree Analysis Examples and How to Use Them. Venngage, 2022

[6] Wei Cheng, Ke Wang, Xiuying Zhang. A comgis based decision tree model in agricultural application. Institution of Remote Sensing & Information System Application, Zhejiang University, 2008.

[7] JingGui Lu, Yi Liu, and Xiaoli Li. The decision tree application in agricultural development. Taiyuan University of Technology, 2011.

[8] Meng-Chang Tsai, Kun-Huang Chen, Chao-Ton Su, and Hung-Chun Lin. An application of PSO algorithm and decision tree for medical problem. 2nd International Conference on Intelligent Computational Systems, 2012.

[9] Vili Podgorelec, Peter Kokol, Bruno Stiglic, Ivan Rozman. Decision Trees: an overview and their use in medicine. Journal of Medical Systems, 2002, 26(5).

[10] Chen LiXia, Zhang ZhiGang. Application of decision tree algorithm in the analysis of student achievement. National Conference on Information Technology and Computer Science, 2015.