ENTROPY-BASED DECISION TREES

MIRVJEN ULQINAKU, ANA KTONA

Department of Computer Sciences, Faculty of Natural Sciences, University of Tirana

e-mail: mirvjen.ulqinaku@fshn.edu.al

Abstract

This article offers an overview of the concept of entropy and its relationship with decision trees. Specifically, it explores Shannon's entropy and its influence on information theory, as well as the ID3, C4.5, and C5.0 decision trees that are based on this entropy. The article delves into the subject, providing a comprehensive understanding of the topic.

Key words: *Entropy, Decision Tree, ID3, C4.5, C5.0, Induction, Classification.*

Përmbledhje

Ky artikull jep një përshkrim të përgjithshëm të konceptit të entropisë dhe pemëve të vendimeve. Ai trajton më gjerësisht entropinë e Shannon-it me influencën e saj në teorinë e informacionit, si dhe pemët e vendimeve ID3, C4.5 dhe C5.0, të bazuara në këtë entropi. Artikulli thellohet në subjekt, duke ofruar një kuptim të plotë të tij.

Fjalë kyçe: *Entropi, Pemë Vendimesh, ID3, C4.5, C5.0, Induksion, Klasifikim.*

Introduction

The evolving landscape of artificial intelligence and its vast influence across various theoretical and practice fields, has underscored the significance of understanding and exploring related topics. This article aims to contribute to the practical application and usefulness of these topics in critical fields such as medicine, robotics, geographic information systems, weather and climate forecasting, natural language processing, finance, business, marketing, quality control, and cost management.

With the goal of achieving these benefits, this article aims to focus on the topic of Entropy, which serves as a measure of the degree of disorder within a

physical system and expresses the level of uncertainty of information within a given system. Additionally, the article will provide a comprehensive overview of Decision Trees, a highly intuitive tool that offers a visual representation that is easily interpreted. It will differentiate between two types of decision trees, classification and regression, and compare them according to their distinct characteristics. Furthermore, to provide a comprehensive understanding of decision tree methodology, the article will describe the two phases of constructing classification trees, induction and pruning.

This article centers on decision trees, particularly ID3, C4.5, and C5.0 algorithms, which are entropy-based decision trees. For each of them, the article will outline the advantages and disadvantages, the selection of splitting criterion, the issues and their minimization or avoidance, the types of data they use, their behavior in the presence of uncertainties, noise or missing data. Furthermore, the article will provide recent applications of these algorithms.

To remain within this article subject's framework, only decision trees that use entropy as a tool for classification and prediction will be taken into consideration.

Entropy

Entropy is a scientific concept that was first introduced by Clausius (1868) as a new formulation of the second law of thermodynamics, and subsequently used by Boltzmann as a statistical interpretation of this principle.

Entropy has wide-ranging applications in numerous fields. In physics, it describes the disorder or randomness of particles in a given system. In chemistry, entropy is used to capture the tendency of a system toward increased disorder or randomness. Furthermore, entropy is widely used as a measure of complexity for systems, such as the study of biological systems, where it is used to quantify the complexity of neural networks or gene expression patterns. Entropy also plays a vital role in practical applications, such as cryptography, where it is employed in generating secure encryption keys, as well as in data compression.

Despite having varied definitions in different fields, the essence and the property that unites these fields is the characterization of the degree of uncertainty, disorder and randomness within a system. Shannon (1948) used this fundamental property of entropy to describe complex systems through the

relative frequency of their components, which led to the creation of Information Theory. From an information theory perspective, entropy measures the uncertainty or randomness of an event in a system. In this regard, it is commonly held that information reduces uncertainty, while entropy increases uncertainty. Brillouin (1959) posits that entropy and information exhibit an inverse relationship: specifically, the greater the information available, the smaller the entropy will be, and vice versa. Since the concept of entropy as a measure of uncertainty can sometimes cause confusion, Pearl (1979) suggested that it should be viewed as a measure of the effort required to remove uncertainty from a system.

The subject of interest for this article is the concept of Shannon's entropy, expressed as:

$$H_i = -\sum_{i=1}^{n} p_i \cdot log(p_i) \qquad (1)$$

where $log \equiv log_2$, n is the number of system's different states and p_i is the probability of event *i*. The logarithm used for Shannon's entropy is generally the binary logarithm. That is the reason entropy is measured in binary units (bits). If necessary, the logarithmic base used in Shannon's entropy can be altered, provided that the base is > 1. In such cases, a straightforward multiplication of the entropy by a constant K converts the measure to the new base, Shannon (1948).

Figure 1 depicts the entropy curve for a two-event system, where the measure exhibits a parabolic shape with its maximum occurring at a probability of 0.5, and its minimum values coinciding with probabilities of 0 and 1, respectively. Since the entropy function provides a quantitative expression of the probabilistic uncertainty associated with the outcomes of a given system, it increases with the degree of uncertainty of the event and decreases as the event becomes more predictable. In the case of Figure 1, the highest level of uncertainty corresponds to a probability of 0.5, while the smallest values occur at probabilities of 0 and 1.

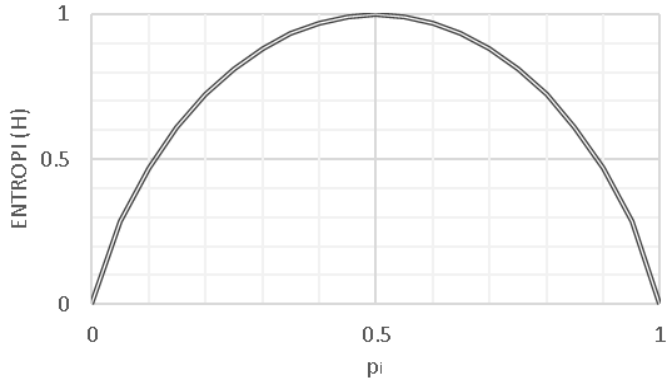

Figure 1: Shannon's entropy for a two-event system

Conditional Entropy: Conditional entropy is a critical concept in information theory, and its relevance will become apparent in the following sections of this article. Consider two dependent systems A and B, with a_i and b_j denoting the possible states of each system, respectively. The conditional probability $P(b_j|a_i)$ represents the likelihood of system B being in state b_j given that system A is in state a_i. In this context, the entropy of system B given that system A is in state a_i can be calculated by applying Equation (1) to the conditional probability distribution:

$$H(B|a_i) = -\sum_{j=1}^{n} p(b_j|a_i) \cdot \log p(b_j|a_i) \qquad (2)$$

Equation (2) provides the mean to quantify the conditional entropy of a system in information theory, revealing the amount of uncertainty that remains in system B when the information pertaining to system A is known. This suggests that entropy is a measure of the degree of uncertainty associated with a given variable. Moreover, as information acquisition serves to diminish uncertainty, the reduction in entropy can be interpreted as an indicator of the amount of information gained about a variable (Perche, 1999).

From a data perspective, entropy serves as a measure of diversity in a given system. Specifically, the greater the number of distinct states present in a system, the larger the system and the more difficult it becomes to ascertain its current state, leading to higher entropy values. Conversely, a system with

fewer states will exhibit lower entropy values as its state can be more readily determined (Marcon, 2019).

As a fundamental concept in information theory, entropy plays a central role in numerous domains such as databases, data mining, machine learning, and artificial intelligence. Irrespective of the specific characteristics of a given database, its composition, or its intended application, there is often a requirement for classification and categorization of data into distinct classes. In this regard, certain types of machine learning algorithms, the Decision Trees, have proven to be highly valuable tools.

Entropy, a foundational construct in information theory, assumes a pivotal role across multiple domains, encompassing databases, data mining, machine learning, and artificial intelligence. Its significance doesn't depend on the specific details of each database, how they're put together, or what they're meant for. Rather, entropy stands as a universal metric, providing a quantitative assessment of diversity or disorder inherent in datasets. In practical terms, there is a common need to organize and sort data into specific groups, regardless of the details of the database. This need is important for good data analysis and decision-making. Here, we introduce Decision Trees, a type of machine learning that's known for being effective. These algorithms are useful for finding detailed patterns in data, some of them using entropy as a guide. Decision Trees do not just help with sorting data; they also give a detailed understanding of the complexity of data. In our review of Decision Trees, we aim to explain how these algorithms handle different data situations, using entropy to improve their effectiveness in various areas.

1 Decision trees

The decision tree is a recursive supervised learning algorithm widely used in classification and regression. Unlike other methods which operate in a single step, the decision tree employs a hierarchical approach to explore, classify, and make decisions. Since the recognition of artificial intelligence as a distinct field, machine learning, including decision trees, has been the central focus of extensive studies, research, and applications. Morgan and Sonquist (1963) were the pioneers in employing decision trees in an explanation and prediction process. Subsequently, several other works, such as Morgan and Messenger (THAID), Kass (CHAID), Breiman et al. (1984) (the well-known CART algorithm), Quinlan (ID3, C4.5 and C5.0) and Rakotomalala (2005), have

contributed to the development of decision trees. According to Santos (2015), the decision tree serves two purposes: exploration – understanding the structure and relationships between instances and attributes of a dataset and generating tree rules (which can be a purpose in itself), and prediction – presenting new information with the aid of established rules.

With the rapid technological advancements of recent years, decision trees have found widespread application in numerous theoretical and real-world domains. In medicine, decision trees aid in the analysis of symptoms, test results, and imaging. Robotics employs decision trees for predicting movements, image and object recognition, and adapting to new environments. Decision trees are also extensively used in natural language processing for text classification, in finance for managing risk and approving loans, in business for predictive analysis (Lee *et al*., 2022), and in marketing for market analysis. Furthermore, decision trees assist in quality and cost control by analyzing production quality and reducing costs. The latest advancements in weather and climate, where the uncertainty of prediction is as crucial as the prediction itself, have also benefited from decision trees (ECMWF[1], 2023). Other notable applications of decision trees include the automation of odor recognition and classification (Mccoy *et al.,* 2003) and their impact on geographic information systems, including the prediction and measurement of urban expansion and remote sensing.

In the context of data sets used in decision trees, it is important to clarify the following terminology: class – a subset of all the records or results that the algorithm attempts to predict; attribute – the data that characterizes a record and is used as input to make decisions; instance – a record in the data set. Decision trees are composed of three main elements: nodes – different values of attributes used to differentiate between classes as much as possible and to minimize diversity within the same class (the first node of the tree and the first element that the tree-creating algorithm will create is called the root, from which the decisions will be derived); branches – different attribute values used for classification; leaves – classes that include objects that are very similar to each other.

[1] European Centre for Medium-Range Weather Forecasts – ECMWF (2023): https://lms.ecmwf.int/pages/index.html

There are two main variants of decision trees: classification and regression. The classification variant of decision trees is designed to create homogeneous subsets of data by recursively partitioning the data based on their attributes. By doing so, classification trees can predict the dependent variable based on the characteristics of several independent variables. On the other hand, regression trees use recursive partitioning to approximate the proportions of classes as well as to predict a target attribute, provided that the latter is continuous in nature, (Xu et al., 2005). They predict the value of a continuous variable based on some other continuous or categorical variables.

Table 1 provides a comparison of decision trees for classification and regression based on several key aspects, (Loh, 2011) (Razi & Athappilly, 2005) (Salih & Abdulazeez, 2021).

Table 1: Comparing classification and regression decision trees

	Classification D.T.	**Regression D.T.**
	Similarities	
Objective	Models and rules prediction	Models and rules prediction
Structure	Hierarchy	Hierarchy
Functioning	Recursive partitioning	Recursive partitioning
Type of procedure	Non-parametric	Non-parametric
Handling of missing data	Suitable	Suitable
Overfitting	Tends to overfit	Tends to overfit
Scaling	Difficulty for large datasets	Difficulty for large datasets
	Differences	
Data type for prediction	Categorical	Continuous
Split criterion	Information gain (Entropy) / Gini index	Variance minimization

Evaluation metrics	Precision, Recall and F1-score	MSE and MAE
Sensitivity to outliers	Suitable	Partially suitable

Table 1 is partitioned into two sections, wherein the first segment displays the commonalities shared by classification and regression trees, and the second section presents their distinguishing features.

As depicted, some of the shared aspects are:

– both trees aim to predict models and rules;

– the structure of both trees is hierarchical. However, it is worth mentioning that in cases where the links between the attribute to be predicted and the predictive attributes need to be displayed, regression trees have better representation in the form of a graph;

– both trees employ recursive partitioning and are non-parametric;

– techniques have been devised for managing missing data in both trees, rendering them more suitable for real models;

– despite these advantages, these types of trees suffer from several limitations such as over-fitting, which pertains to their tendency to become too complex while classifying and fitting all data, leading to good training abilities but reduced generalization of classifications and predictions;

– these types of trees experience difficulty in processing large datasets due to the complexity of the mathematical operations involved, which must be carried out recursively for each instance and for each of its attributes.

In contrast, there are several differences between classification and regression trees, such as:

– the data provided by classification trees are categorical (data that can be divided into categories or groups), while data provided by regression trees are continuous (data that can take infinite values in an assigned interval);

– classification trees use information gain as a criterion for splitting nodes, which includes entropy or the Gini index, whereas regression trees employ variance minimization;

– different evaluation metrics are utilized by each type of tree. Classification trees use techniques such as Precision, Recall, or F1-score, whereas regression trees rely on Mean Squared Error and Mean Absolute Error;

– data often do not follow a specific pattern and may contain what are called outliers, which are better fit by classification trees rather than regression ones.

Although regression decision trees have been the subject of considerable research and interest, this article will solely focus on classification decision trees since they specifically make use of entropy.

1.1 Induction

The process of creating decision trees for classification can generally be divided into two stages: induction and pruning. The induction – or construction phase – is responsible for establishing a set of rules that enable instances (objects) to be classified based on their attributes (Quinlan, 1986). Initially, a training set is used to determine the target attribute for prediction. Subsequently, supporting attributes are defined that will serve to predict (or classify) the target attribute. The next step is to select the attribute that maximizes the homogeneity of the data between the classes to be created. This selection is based on the Information Gain (IG) criterion, where the attribute with the highest IG is chosen. As a reminder, the Information Gain is calculated using the entropy equation (1), expressed as:

$$IG(TA, SA)=H(TA)-\sum_{v\in SA}\frac{TA_v}{TA}\cdot H(TA_v) \qquad (3)$$

In the above equation, TA refers to the target attribute (to be predicted), SA is the supporting (predictor) attribute, and TA_v represents a subset of TA for which SA attribute has the value v.

The first attribute to be selected becomes the root of the tree. Subsequent node selections are performed recursively based on the steps outlined above until the leaves of the tree are reached, and their respective classes are determined. The purpose of the induction phase in a decision tree is to establish a classification system for the objects in the training set that can be applied to other sets not previously tested in the tree.

1.2 Pruning

Breiman *et al*. (1984) have suggested that the performance of decision trees is primarily dependent on the size of the tree. In most cases, decision tree algorithms aim to create models that are best suited to the provided data, which can result in the development of highly complex trees, leading to the undesired phenomenon of overfitting.

One way to prevent overfitting is through the technique of pruning. Pruning involves reducing the size of the tree by removing branches that contribute little to the classification or prediction process. Essentially, pruning serves to control the complexity of the tree, resulting in a sub-tree that avoids the risk of overfitting and allows for generalization of the decision tree. There are three main types of pruning, as identified by Kotsiantis (2013): pre-pruning, post-pruning, and data pre-processing:

– Pre-pruning involves setting a limit for the tree's maximum depth, after which it is cut;

– In post-pruning, the tree grows to its full length and is then pruned by removing branches that do not improve its accuracy;

– Data pre-processing, while not directly related to the tree, involves simplifying the data before it is used by the decision tree.

The literature suggests that there are various pruning techniques available for decision trees, and no single technique has been found to outperform the others significantly (Esposito *et al.*, 1997).

Upon completion of the construction and pruning phases, the resulting tree can be tested on a validation set to evaluate its performance using the metrics described in Table 1. This step is crucial in determining the effectiveness of the decision tree model in accurately classifying new, unseen data.

2. Entropy-Based Decision Trees

Thus far, this article has aimed to provide a comprehensive overview of entropy and decision trees. Moving forward, the focus will be on exploring three decision trees that use entropy as a component in their operation: ID3, C4.5, and C5.0.

2.1 ID3

The ID3 algorithm, developed by Ross Quinlan (Quinlan, 1986), is widely recognized for its simplicity, speed, and interpretability in creating classification decision trees. Due to its straightforwardness and popularity, ID3 has become one of the most widely used decision tree algorithms. The algorithm selects the splitting criterion for each node based on the maximum information gain (equation 3) and the smallest entropy. The objective of ID3 is to classify a target attribute using other attributes, while constructing a tree that seeks to minimize the entropy at each node.

The issue of overfitting is a significant challenge faced by the ID3 algorithm, as noted in the pruning section.

Another challenge is the method of selecting attributes, which relies on information gain. Kononenko *et al.* (1984) have demonstrated that this criterion favors multi-valued attributes, which can result in biased attribute selection. To mitigate this issue, a condition can be imposed to limit each trial to only two outcomes. Nonetheless, this tendency to favor attributes with a greater variety of values can result in good performance on training sets when used to predict the majority class, but may lead to poor predictions on unseen data.

In addition to over-fitting, ID3 and other tree-building algorithms face another problem, which is noise in the data. Noise in the data refers to inaccuracies in the attribute values or class labels, and it can seriously affect the accuracy of the decision tree. Since noise in the data is inevitable, to address this issue Quinlan (1986) proposed two modifications to the ID3 algorithm: (a) enable the algorithm to determine the optimal point at which to stop testing attributes to improve classification/prediction accuracy; (b) enable the algorithm to handle insufficient, incomplete or incorrect attributes. To achieve the latter, three approaches are suggested: (i) replacing an unknown attribute value with its most frequent value; (ii) using Bayesian method; (iii) using a decision tree to determine unknown values. Quinlan (1986) found that the decision tree approach was the most effective, while value substitution and the Bayesian method yielded poorer results – despite the latter having a smaller error rate. Nevertheless, ID3 remains sensitive to missing data, even with these modifications.

ID3 faces challenges when working with continuous data. The algorithm is designed to handle categorical data and cannot handle continuous data

directly. As a solution, Han *et al.* (2012) propose the use of discretization techniques, which involve dividing continuous data into discrete intervals, allowing the algorithm to handle these data types. However, this approach can lead to loss of information and reduce the accuracy of the resulting decision tree. Therefore, selecting an appropriate discretization method is critical to ensuring the accuracy of ID3 when dealing with continuous data.

ID3 is a popular and widely used algorithm for classification due to its simplicity, comprehensibility, efficiency, and accuracy. Its applications span across various fields, including Geographic Information Systems (GIS). In GIS, ID3, and entropy are the basis for the work of Li and Claramunt (2006), who highlight the difference between conventional and spatial entropy. Other notable works that have used ID3 include Wang *et al.* (2017) with a new approach for selecting the splitting attribute, and Soni *et al.* (2017) with an improved version of ID3 for emotion-based text classification. These studies demonstrate the versatility of the ID3 algorithm and its potential for solving various classification problems in different fields.

2.2 C4.5

C4.5 algorithm, again developed by Ross Quinlan (Quinlan, 1993), represents a significant improvement over its predecessor, ID3. Like ID3, C4.5 is a powerful algorithm that creates classification decision trees for classification and prediction purposes. Given that this algorithm builds upon the ID3 framework, the details of its construction and operation will not be reiterated in this section.

The C4.5 algorithm is another decision tree creating algorithm that uses entropy. However, unlike ID3 – which uses information gain as a splitting criterion – C4.5 employs the gain ratio. To determine the gain ratio, it is necessary to introduce a new concept called Split Information. Split Information represents a normalization of the information gain derived from the ID3 algorithm and indicates the potential information obtained from splitting the data set. It is expressed as:

$$Split\ Info.(S,TA) = -\sum_{i=1}^{n}\left(\frac{|S_i|}{|S|}\cdot log_2\frac{|S_i|}{|S|}\right) \qquad (4)$$

Where: TA represents the target attribute for testing, S denotes the entire

dataset, and S_i represents a subset of S for which it takes on values ranging from i through n for attribute TA.

Once the Split Information has been determined and the information gain has been obtained using Equation (3), the gain ratio is defined as follows:

$$Gain\ Ratio(TA) = \frac{IG(TA)}{Split\ Info.(S,TA)} \qquad (5)$$

Following the testing of all attributes, the one with the highest gain ratio is selected for splitting. However, it is possible for the Split Information to become 0 or approach this value. In such cases, all or almost all instances are grouped into a single branch, resulting in a perfect split, where the entropy is 0 and there is no uncertainty in the result. This scenario can occur as a result of overfitting. According to Han *et al.* (2012) if this value approaches 0, it can lead to unstable gain ratio. They suggest that to avoid this issue, one option is to impose the condition that the information gain for a given test must be at least as large as the average information gain of all tests performed. Another strategy is to control the complexity of the tree by applying pruning techniques, which have been previously discussed in this article.

As an advancement over ID3, C4.5 allows for the utilization of both categorical and continuous data. To process continuous data, C4.5 employs a discretization technique, where the continuous data is partitioned into a series of discrete intervals or "bins". Subsequently, these bins are treated in the same manner as categorical data, enabling the decision tree to make decisions. Nevertheless, the efficacy of this technique hinges on the type of the training set and the used parameters (Quinlan, 1996a).

Another means by which the C4.5 algorithm mitigates against overfitting is through the implementation of pruning during or after tree construction. When pruning is applied after the tree has been built, pessimistic pruning, as described by Quinlan (1987), is one possible technique that may be used. This method involves testing each sub-tree of the tree constructed against a validation set, with the aim of identifying the error rate of each. Pruning is only performed if the sub-tree's error rate exceeds that of the entire tree, based on the validation set.

In the real world, uncertainty and missing data are common occurrences that directly impact the accuracy of decision tree predictions. It is therefore

important to consider both of these factors for the most accurate and efficient classifications. To this end, Jenhani *et al.* (2009) propose a clustering method for dealing with uncertainty and is applied in the attribute selection criterion of the C4.5 algorithm – gain ratio. This method can be achieved through two strategies, namely baseline and hierarchical. On the other hand, Twala et al. (2008) address the issue of missing data and propose a method that groups missing data for an attribute being considered for splitting, into a separate set – thus both categorical and continuous data can be handled.

The C4.5 algorithm has become a popular and widely used decision tree algorithm, which has led to numerous studies focused on its improvement (Kotsiantis, 2013). One approach to improving the algorithm is presented by Yao *et al.* (2005), which suggests using attributes with greater than average gain ratio and merging high entropy branches to reduce overfitting and increase accuracy. Ruggieri (2002) proposes an initial arrangement of the data as an approach to improving the algorithm. Muslim *et al.* (2018) presents a new method called Particle Swarm Optimization for breast cancer diagnosis. These studies, along with others, demonstrate ongoing efforts to enhance the performance and accuracy of the C4.5 algorithm.

2.3 C5.0

C5.0 is the third decision tree algorithm developed by Ross Quinlan as an improvement over the C4.5 algorithm. This algorithm is capable of creating binary or multi-branch decision trees based on several attributes and training sets. The created decision trees can then be tested on several validation sets and finally used for classification and prediction on new sets. At each node of the tree, the splitting criterion is determined using the information gain (equation 3). This indicates the amount of entropy reduction obtained by splitting the data according to a given attribute, similar to the ID3 algorithm.

C5.0 is a highly effective tree-building algorithm that performs classification and prediction with nearly the same accuracy as other machine learning algorithms. It boasts several strengths, as noted by Lantz (2015), including its efficiency in comparison to earlier versions, as well as its ability to handle both large and small data sets. C5.0 also excels at treating categorical and continuous data, and its split and pruning criterion optimization leads to faster tree creation. However, like its predecessors, C5.0 is susceptible to overfitting. To minimize overfitting, post-pruning is typically used.

Despite its strengths, the C5.0 algorithm has several weaknesses. For example, while the algorithm has reduced its tendency to favor multi-valued attributes, this bias still exists. Additionally, small changes in the training set can result in significant changes in the decision logic. Finally, as the size of the tree increases, it can become more challenging to interpret.

The C5.0 algorithm has been enhanced with several new techniques, including the cost of attribute misclassification and boosting, as noted by Bujlow *et al.* (2012). Misclassification cost is a technique that assigns different costs to instance misclassifications. Given that classification errors can vary, the implications of misclassification will also differ. As a result, algorithms assign costs to these errors. This approach enables the algorithm to consider both information gain and the costs associated with incorrect classifications, thereby constructing a tree that maximizes information gain while minimizing misclassifications.

Another technique, known as boosting (Quinlan, 1996b), involves constructing multiple decision trees that use all available instances in the training set and assign weights based on their importance. Each tree makes a contribution to the weights of the instances, allowing subsequent trees to focus more on the most relevant ones. Boosting prioritizes misclassified instances by giving them greater weight, resulting in improved classification accuracy and more accurate predictions.

Since its initial development, the C5.0 algorithm has undergone multiple modifications and improvements, and has been widely implemented in both commercial and open-source applications. Notable studies employing C5.0 include Pang and Gong's (2009) assessment of individual loans, Kristóf and Virág's (2022) prediction of E.U. bank failures, Guo *et al*. (2021) modeling of areas affected by landslides, Khadiev et al. (2019) proposal of a quantum version based on this algorithm, among others.

Conclusions

This article provided a summary of some key aspects related to the concept of entropy and its relationship with Decision Trees.

While the term "entropy" can have varying interpretations across different fields, it is generally regarded as a measure of the uncertainty and randomness present within a system. The focus of this article was on Shannon's entropy,

which serves as the foundation of Information Theory.

The article also touches on the topic of conditional entropy, which represents the degree of uncertainty within a system when additional information is available about another system. These concepts of entropy have been employed in the creation of various types of decision trees.

This article has attempted to provide a modest introductory overview of decision trees, since they constitute a vast area of study within machine learning and artificial intelligence. Specifically, the article focuses on decision trees that use the concept of entropy as a means of splitting instance attributes within datasets.

This article also touches on the two primary stages of decision tree creation, namely induction and pruning.

This article has examined the ID3, C4.5, and C5.0 tree algorithms, all of which use the concept of entropy as a means of splitting instance attributes within datasets. For each algorithm, the article presents an overview of their respective advantages and disadvantages, along with their avoidance techniques, ongoing improvements, and recent proposals and applications.

In conclusion, decision trees based on the entropy splitting criterion represent a powerful tool for classification and prediction tasks. While classifications may seem like a small part of procedural functions, they play a crucial role in more complex tasks such as robotic movements and computer game development.

Future Works

Beyond the important role of entropy discussed in this article, it is worth noting that it has recently found new applications in Geographic Information Systems (GIS) for showing the concentration and spatial distribution of variables.

While the current article had a focus on the objective uses of entropy in decision trees, future research will delve more deeply into this specific application of entropy in GIS.

References

Breiman L., Friedman J. H., Olshen R. A., Stone C. J. (1984): Classification and Regression Trees (1st ed.). Routledge

Brillouin, L. (1959): La science et la théorie de l'information. Masson

Bujlow, T., Riaz, T., & Pedersen, J. M. (2012): A method for classification of network traffic based on C5. 0 Machine Learning Algorithm. In 2012 international conference on computing, networking and communications (ICNC) (pp. 237-241). IEEE

Clausius R. (1868): Théorie mécanique de la chaleur. Eugène Lacroix, Paris

Esposito, F., Malerba, D., Semeraro, G., & Kay, J. (1997): A comparative analysis of methods for pruning decision trees. IEEE transactions on pattern analysis and machine intelligence, 19(5), 476-491

Guo, Z., Shi, Y., Huang, F., Fan, X., & Huang, J. (2021). Landslide susceptibility zonation method based on C5. 0 decision tree and K-means cluster algorithms to improve the efficiency of risk management. Geoscience Frontiers, 12(6), 101249

Han, J., Kamber, M., & Pei, J. (2012): Data mining concepts and techniques third edition. University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University

Jenhani I., Benferhat, S., & Elouedi, Z. (2009): On the use of clustering in possibilistic decision tree induction. In Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 10th European Conference, ECSQARU 2009, Verona, Italy, July 1-3, 2009. Proceedings 10 (pp. 505-517). Springer Berlin Heidelberg

Khadiev, K., Mannapov, I., & Safina, L. (2019): The quantum version of classification decision tree constructing algorithm C5. 0. arXiv preprint arXiv:1907.06840

Kononenko I., Bratko I., Roskar E. (1984): Experiments in automatic learning of medical diagnostic rules. Technical Report, Jozef Stefan Institute

Kotsiantis, S. B. (2013): Decision trees: a recent overview. Artificial Intelligence Review, 39, 261-283

Kristóf, T., & Virág, M. (2022): EU-27 bank failure prediction with C5.0 decision trees and deep learning neural networks. Research in International Business and Finance, 61, 101644

Lantz B. (2015): Machine Learning with R (2nd ed.). Packt Publishing, Birmingham

Lee, C. S., Cheang, P. Y. S., & Moslehpour, M. (2022). Predictive analytics in business analytics: decision tree. Advances in Decision Sciences, 26(1), 1-29

Li, X., & Claramunt, C. (2006): A spatial entropy- based decision tree for classification of geographical information. Transactions in GIS, 10(3), 451-467

Loh, W. Y. (2011): Classification and regression trees. Wiley interdisciplinary reviews: data

mining and knowledge discovery, 1(1), 14-23

Marcon E. (2019): Mesure de la biodiversité et de la structuration spatiale de l'activité économique par l'entropie. Revue Economique, Presses de Sciences Po, 2019, 70 (3), pp.305

Mccoy, S. A., Martin, T. P., & Baldwin, J. F. (2003). Learning rules for odour recognition in an electronic nose. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 11(05), 517-543

Morgan J., Sonquist J.A. (1963): Problems in the Analysis of Survey Data, and a Proposal, Journal of the American Statistical Association, 58:415-435

Muslim, M. A., Rukmana, S. H., Sugiharti, E., Prasetiyo, B., & Alimah, S. (2018): Optimization of C4. 5 algorithm-based particle swarm optimization for breast cancer diagnosis. In Journal of Physics: Conference Series (Vol. 983, No. 1, p. 012063). IOP Publishing

Pang, S. L., & Gong, J. Z. (2009): C5.0 classification algorithm and application on individual credit evaluation of banks. Systems Engineering-Theory & Practice, 29(12), 94-104

Pearl, J. (1979): Entropy, information and rational decisions. Policy Analysis and Information Systems, Special Issue on Mathematical Foundations, 3(1), 93-109

Perche Paul-Benoît (1999): Méthodes d'induction par arbres de décision dans le cadre de l'aide au diagnostic (Thèse de Doctorat, Université des Sciences et Technologies de Lille), Lille

Quinlan J. R. (1986): Induction of Decision Trees. Kluwer Academic Publishers, Boston. Machine Learning 1: 81-106

Quinlan, J. R. (1987): Simplifying decision trees. International Journal of Man-Machine Studies 27, 221-234

Quinlan J. R. (1993): C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo

Quinlan J. R. (1996a): Improved use of continuous attributes in C4. 5. Journal of artificial intelligence research, 4, 77-90

Quinlan, J. R. (1996b). Bagging, boosting, and C4. 5. In Aaai/Iaai, vol. 1 (pp. 725-730)

Rakotomalala R. (2005). Arbres de décision. Revue Modulad, 33, 163-187

Razi, M. A., & Athappilly, K. (2005): A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. Expert systems with applications, 29(1), 65-74

Ruggieri, S. (2002): Efficient C4.5 [classification algorithm]. IEEE transactions on knowledge and data engineering, 14(2), 438-444

Salih, A. A., & Abdulazeez, A. M. (2021): Evaluation of classification algorithms for intrusion detection system: A review. Journal of Soft Computing and Data Mining, 2(1), 31-40

Santos F. (2015): Arbres de décision. CNRS, UMR 5199 PACEA, Bordeaux

Shannon, C.E. (1948): A Mathematical Theory of Communication. Bell System Technical Journal, 27, 379-423

Soni, V. K., & Pawar, S. (2017): Emotion based social media text classification using optimized improved ID3 classifier. In 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS) (pp. 1500-1505). IEEE

Twala B. E., Jones, M. C., & Hand, D. J. (2008): Good methods for coping with missing data in decision trees. Pattern Recognition Letters, 29(7), 950-956

Wang, Z., Liu, Y., & Liu, L. (2017): A new way to choose splitting attribute in ID3 algorithm. In 2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC) (pp. 659-663). IEEE

Xu, Min & Watanachaturaporn, Pakorn & Varshney, P.K. & Arora, Manoj. (2005): Decision tree regression for soft classification of remote sensing data. Remote Sensing of Environment. 97. 322-336. 10.1016/j.rse.2005.05.008

Yao, Z., Liu, P., Lei, L., & Yin, J. (2005, June). R-C4. 5 Decision tree model and its applications to health care dataset. In Proceedings of ICSSSM'05. 2005 International Conference on Services Systems and Services Management, 2005. (Vol. 2, pp. 1099-1103). IEEE