2018 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)

Accuracy vs. Cost in Decision Trees: A Survey

Mona Al Hamad
Department of Information Systems
University of Bahrain
Manama, Bahrain
mona.alhamad@gmail.com

Ahmed M. Zeki
Department of Information Systems
University of Bahrain
Manama, Bahrain
amzeki@uob.edu.bh

***Abstract*—Decision Trees have been applied widely for classification in many fields such as finance, marketing, engineering, and medicine. The increased field of application, made the requirement for understanding various aspects of decision trees in deep. In addition, it is crucial to understand the different type of costs associated with the classification task in a decision tree classifier and their relationship with the classifier's accuracy, as balancing the two is a major concern these days in many fields such as medical diagnosis. This paper introduces the concept of decision trees, presents their various areas of application in data mining, summarizes the standard decision tree algorithms, and identifies their main advantages and disadvantages. It mainly aims to clarify relationship between the classification accuracy and classification cost in decision trees.**

Keywords—decision tree, classification, data mining, cost, accuracy

I. INTRODUCTION

Decision Trees (DTs) are one of the most popular models for classification in many application domains [1]. DT is a "tree-shaped diagram representing a sequential decision process in which attribute values are successively tested to infer an unknown state" [2]. A DT is composed of a root node, internal/test nodes, and leaf nodes having a class or a label. The root node has no incoming edges, while all other nodes have only one incoming edge. Internal nodes also have one outgoing edge [3]. Within each internal node the value of the attribute assigned to the node is tested to determine the next node to go to along the path, and when a leaf node is reached, its label will represent the current class [2].

Different types of costs are associated with the classification task such as misclassification cost, test cost, and computational cost. Various cost-sensitive algorithms were proposed to minimize the total classification cost of a DT construction such as Cost Sensitive Iterative Dichotomiser 3 (CS-ID3), IDX, EG2, and others [4-6], as the standard DT algorithms classify regardless the cost. However, in many real world applications such as medical diagnoses reducing the cost of classification is not the only concern. As a result, balancing between the two (total classification cost and classification accuracy) is the focus of many researchers recently [7-9]. In this paper, the accuracy and cost of a DT will be surveyed.

The rest of this paper is organized as follows: some of the DT applications will be reviewed in section 2. Section 3 will briefly summarize the standard DT algorithms. In section 4, the main advantages and disadvantages of DTs will be discussed, while section 5 will introduce the attribute selection measures. Section 6 will review the classification cost and accuracy and their relationship in a classification task. Finally, section 7 will conclude this paper and discuss findings.

II. DECISION TREE APPLICATIONS

DTs have been successfully utilized in applications of different areas such as education, finance, and healthcare. DT can be applied for mining educational data to discover hidden patterns such as detecting unwanted student behavior, predicting the student performance and learning achievements, and helping educators in planning for future courses [4]. It can also be applied in business for fraud detection by constructing a DT from the normal customer behavior and use it to predict fraudulent financial behavior [5]. Application of DTs in business also include market segmentation to predict the customers who are likely to buy certain products and customer churn to predict the customers who are likely to leave to another competitor [6]. Other applications of DTs include healthcare for medical diagnoses, credit card analysis among others [7].

III. DECISION TREE CONSTRUCTION ALGORITHMS

Several DT algorithms exist to construct or model DTs. This section review the standard and most popular DT algorithms namely: ID3, C4.5, and CART.

A. Iterative Dichotomiser (ID3)

ID3 is a the standard DT algorithm developed by J. Ross Quinlan in 1986. In ID3, only a subset of the dataset is randomly selected to build the DT, while the remaining data are used to test the tree. If the constructed tree is able to classify all testing records correctly, then the tree is considered correct and the process finishes. Otherwise, another subset of the dataset including some of the incorrectly classified records is used to construct a new tree. The process continues until all records in the dataset are classified correctly by the constructed tree [8].

B. C4.5 Algorithm

C4.5 is an improved version of ID3 by Quinlan in 1993. C4.5 algorithm has some advantages over ID3 in that it can handle both numerical and categorical attributes. It can also handle the missing attribute values [8]. It selects the attribute selection using information-based criteria and uses pruning to trim the constructed tree in order to handle the overfitting problem that arises in DTs [9].

C. Classifier and Regression Trees (CART)

CART is also a DT algorithm developed by Breiman in 1984 [10]. It can perform classification and regression tasks based on the nature of the input data, categorical or numerical data, where the tree leaves in case of regression tree predicts real numbers [8]. Like C4.5, pruning is also used in CART to handle the overfitting problem. However, unlike ID3 and C4.5, CART allows for constructing binary trees only. At each node, all possible partitions are compared and the one with the highest level of homogeneity is selected [11].

IV. ADVANTAGES AND DISADVANTAGES OF DECISION TREES

DT has various advantages, some of its advantages include: easiness to understand, applied to various real world problems successfully, its ability to build classifiers from

978-1-5386-9207-3/18/$31.00 ©2018 IEEE

datasets containing numerical and non-numerical data, and easiness to be converted to classification rules [12]. In addition, DTs are most commonly used for classification due to their computational efficiency [13], they require no domain knowledge or parameter settings and they provide high classification performance [1].

On the other hand, DTs have some disadvantages. Specifically, any change made in the training data results in a change in the attribute selections and affects the whole tree accordingly. In addition, the output attributes must be single and categorical [14].

V. Attribute Selection Measure

The attribute selection measure is a heuristic for choosing the splitting criterion that makes the optimal split for a class-labeled data into classes. In another word, it selects the best attribute for each node, when building the tree [8]. The attribute selection measure gives ranking to each attribute in the dataset and then selects the attribute with the highest ranking as the splitting attribute. The three popular attribute selection measures are information gain used in ID3 algorithm, gain ratio used in C4.5 algorithm, and gini index used in CART algorithm.

VI. Accuracy vs Cost in Decision Trees

A. Accuracy of DTs

Kamber [12] defined the accuracy of a classifier, such as a DT, as the percentage of instances from the testing set, that were correctly classified by a classifier. Confusion matrix is a tool that can be used for this purpose. It is an $m \times m$ table (where m is the number of classes) that evaluates the ability of a classifier to realize objects of different classes. It combines four different counts in case of binary classification where an object classified as one of two class labels. Those counts are: the number of positive objects classified correctly, also called true positives (TP), the number of negative objects classified correctly, also called true negatives (TN), the number of negative objects incorrectly classified, also called false positives (FP), and the number of positive objects incorrectly classified, also called false negatives (FN) [15]. Using the confusion matrix, accuracy can be calculated as:

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

The accuracy is not sufficient measures in an imbalanced dataset with more majority class instances than minority class instances. In this case, sensitivity and specificity measures are better to be used for evaluating models. Sensitivity is the ability of a classifier to recognize positive instances, where specificity is the ability of a classifier to recognize negative instances [3]. They can be calculated as [3]:

$$Sensitivity = \frac{TP}{TP+FN} \quad (2)$$

$$Specificity = \frac{TN}{FP+TN} \quad (3)$$

B. Cost of DTs

Different types of costs involved in the task of building a classification model such as misclassification cost, test cost, teacher cost, intervention cost, computation cost, and human–computer interaction costs. Among all, Turney [16] mentioned that misclassification and tests cost are the most important classification costs to be considered. Where the misclassification cost is the cost incurred when an instance is not classified correctly, i.e. the penalty received when classifying an instance as class X while it belongs to class Y. Whereas test cost is the cost of acquiring the attribute value [17]. For example, in classifying a medical disease, a patient might be described in terms of temperature, blood test, pulse rate, and others, which have its associated monetary cost [18].

The computational cost is the cost of the computer resources used in the classification process. This includes the size complexity, time complexity (execution time) and space complexity (memory space) used by a classifier. It can be measured in terms of memory space needed for execution [16]. Dogan and Tanrikulu [19] and Jensen et al. [20] mentioned that the computational cost represents the classifier's speed, i.e. the CPU time used by the classifier to build the model or the time, computational resources, and memory required to build the model.

Zhao et al. [18] mentioned that the total classification cost of a DT is determined by calculating the average total classification cost of the testing cases, where the dataset is divided into training set and testing set. The average total cost (ATC) of classification is calculated by dividing the total classification cost (sum of misclassification cost and test cost) for all testing cases by the number of testing cases as follows [18]:

$$ATC(U) = \sum_{x \in U} \frac{tc(x)+mc(x)}{|U|} \quad (4)$$

where x is a testing case, U is the testing data set, $tc(x)$ and $mc(x)$ are the test cost and misclassification cost of the testing cases, respectively.

To calculate the test cost of a certain case, the cost of each test along the path from the root to the leaf in a DT is added together. If the same test appears twice, it should be added only once.

The misclassification cost is calculated using the classification cost matrix. Having the predicted class of the case (the tree leaf) and the actual class of the case, the matrix used to determine the misclassification cost. This cost is added to the test cost to get the total classification cost. Table 1 shows an example of a classification cost matrix.

Table 1: An example of a Classification Cost Matrix [21]

	Actual Bad	Actual Good
Predict Bad	0	300
Predict Good	800	0

C. Accuracy-Cost Relationship in DTs

Many researchers realized the relationship between the total classification cost, and the classification accuracy. He et al. [22] mentioned that most DM algorithms are exponential in computational complexity. For example, the process of mining big data may take several hours or even days to finish and consumes huge computational resources especially if 100% accuracy is required. Ahmadi also mentioned, "high accuracy is achieved at the expense of increased computational complexity" [23], as the total classification cost is sometimes measured in terms of the time and resources required to build the classification model.

Bohanec and Bratko [24] clarified the relationship between the DT size and the classification accuracy. They noticed through experimentation that as the DT size (number of leaves) increases, the classification accuracy will also increase as shown in Figure 1. However, they noticed that nine

nodes (out of 21 nodes full size tree) are sufficient to provide an accuracy of 99.57% i.e. the remaining 12 nodes account for an accuracy less than 0.5%.

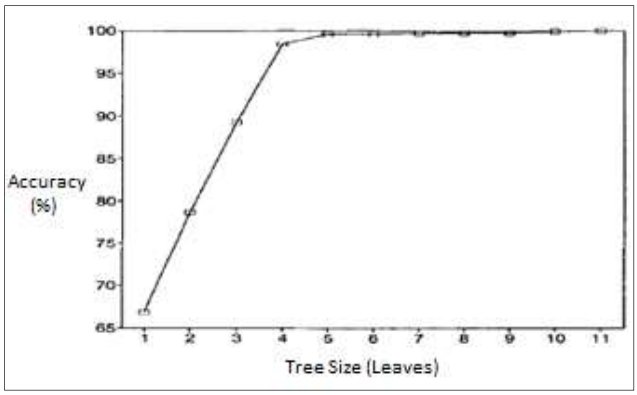

Figure 1: Accuracy vs. Tree Size [24]

Bhowmick et al. [25] confirmed the results of Bohanec and Bratko [24], as they concluded that the larger the DT, the better classification results, but the longer time will be required to construct the classifier. The general rule they came up with is that: the cardinality of feature set affects the cost of constructing and deploying the classifier. Zhou [26] also mentioned that adding more features in a DT will improve the accuracy but will also increase the cost of construction. This asserts that the relationship between the accuracy/cost pairs is proportional.

Turney [9] mentioned that when all test costs are ignored (set to zero), the total average cost of using a DT will be:

$$Cost = p.k \quad (5)$$

where $p \in [0,1]$ is the frequency of errors in the testing set, and k is the error cost from the classification cost matrix. Note that a simple classification cost matrix were used, which means that all errors have the same cost.

From equation 5, Turney [9] calculated the accuracy percentage on the testing set as: $100 \times (1-p)$, which means that there is a linear relationship between the average classification cost and the accuracy percentage, i.e. reducing the cost will increase the accuracy (note that the misclassification cost was the only cost considered in this case) [9].

When Turney [9] considered just the test cost and tried to find its relationship with the accuracy, the hypothesis he made "as the test cost increases, the accuracy will also increase" was confirmed through experimentations.

Many researchers also mentioned the necessity to embed different types of costs –namely misclassification cost and test cost– in classifiers to reduce their total classification cost while having satisfactory accuracy at the same time [27].

Davis et al. [28] modified the existing C4.5 algorithm to build a DT with high accuracy and low test cost. The results showed that the extended algorithm allows for high level of cost reduction (up to six times) compared to the cost-insensitive C4.5 with only 1% reduction in accuracy. Li et al. [1] also worked on minimizing the misclassification cost and maximizing the overall classifier performance of a DT of the minority classes in an imbalanced datasets. The accuracy of the proposed algorithm was compared with the existing cost-sensitive DT algorithms: IDX, CS-ID3, and CS-C4.5. The experimental results showed that the proposed algorithm improved the accuracy performance by 9.32%, 6.16%, and 10.81%, respectively.

Qiu et al. [29] used random attribute selection method, instead of greedy attribute selection used in traditional DTs, to find the best attributes for each splitting in a DT. The aim was to balance between the classification accuracy and efficiency, that is, the proportion of instances classified correctly and the cost of measuring attribute value. The new algorithm was tested on 36 datasets collected from UCI [30] and the results were compared with that of algorithm C4.5. The experimental results showed that the average classification accuracy of the new algorithm (82.73%) was almost the same as C4.5 (82.57%). In addition, the average total test cost of the new algorithm (101.21) was much lower than that of C4.5 (146.73).

In addition, Zhou et al. [26] proposed a new feature selection method that considered the feature "also called test" cost, assuming that different features have different associated costs. It selects the low cost of informative subset of features for constructing the base DT in order to minimize the overall test cost of building the tree while having reasonable classification accuracy. The empirical results showed that the cost reduction reached 30% with slight reduction in accuracy.

Lu et al. [27] embedded the two types of classification costs (i.e. misclassification cost and test cost) in a DT classifier to improve the classification accuracy and reduce the average misclassification cost. They were successful in reducing the average misclassification cost up to 30%, but the drop in accuracy also reached 9.4% compared to rotation forest (RoF) algorithm.

Bhowmick et al. [25] focused on the computational complexity of each features (i.e. the time required to compute each feature) to reduce the total computational complexity of constructing the classification model without sacrifying the accuracy. Their proposed method selects features in increasing order of their computational complexity. The experimental results showed that the new method speeded up the time for building the classifier by a factor of 125 on average.

Table 2 summarizes all reviewed research papers done to minimize the total classification cost and maximize the classification accuracy in a DT.

Table 2: DT Algorithms for Maximizing Accuracy and Minimizing Cost

References	Considered Cost Type	Accuracy vs Cost
[28]	Test cost	High level of cost reduction (up to six times) with only 1% reduction in accuracy.
[1]	Misclassification cost	The accuracy improved by 9.32%, 6.16%, and 10.81% compared to IDX, CS-ID3, and CS-C4.5, respectively.
[29]	Test cost	High level of cost reduction while maintaining same accuracy as C4.5.
[26]	Test cost	The cost reduction reached 30% with slight reduction in accuracy.
[27]	Misclassification cost and Test cost	The reduction in average misclassification cost reached 30%, but the drop in accuracy also reached 9.4% compared to RoF algorithm.

[25]	Computational cost	Speeds up the time for building the DT by a factor of 125 on average with higher accuracy.

VII. CONCLUSION

The amount of data is growing worldwide and the need for data analysis became necessary to discover hidden pattern knowledge. DT is one of the essential tools used for this purpose. This paper first discussed the basic concept of DTs, some DT applications, the standard DT algorithms, the main DT advantages and disadvantages, and the attribute selection measure used for DT construction. The main focus of the paper is to review different type of costs consumed in building a DT, how to calculate them, different accuracy measures for evaluating the performance of a classifier, and the relationship between the classification accuracy and cost. Based on the survey made, the relationship between the classification accuracy and cost in DTs found to be proportional. In addition, among different type of costs, most researchers focused on either test cost or on misclassification cost for constructing cost-sensitive DT.

Most of the researches either focused on balancing test cost and classification accuracy or on balancing misclassification cost and classification accuracy. Among all the reviewed researches, only one research found for balancing both types of cost (test cost and misclassification cost) and the classification accuracy. It was successful in decreasing the total classification cost but the accuracy was degraded too.

REFERENCES

[1] F. Li, X. Zhang, X. Zhang, C. Du, Y. Xu and Y. Tian, "Cost-sensitive and Hybrid-attribute Measure Multi-decision Tree over Imbalanced Data Sets," *Information Sciences,* vol. 422, pp. 242-256, 2018.

[2] S. Ohta, R. Kurebayashi and K. Kobayashi, "Minimizing False Positives of a Decision Tree Classifier for Intrusion Detection on the Internet.," *Journal of Network and Systems Management,* vol. 16, no. 4, pp. 399-419, 2008.

[3] L. Rokach and M. Oded, Decision Trees : Theory and Applications, Singapore: World Scientific, 2008.

[4] B. Bakhshinategh, O. Zaiane, S. ElAtia and D. Ipperciel, "Educational Data Mining Applications and Tasks: A Survey of the Last 10 Years," *Education and Information Technologies,* vol. 23, no. 1, pp. 537-553, 2017.

[5] R. Chattamvelli, Data Mining Algorithms, Oxford: Alpha Science International, 2011.

[6] G. Sharma, "A Study on Data mining Algorithms for Tourrism Industry," *International Journal of Latest Trends in Engineering and Technology,* vol. 7, no. 1, pp. 580-587, 2016.

[7] S. Hussain, "Survey on Current Trends and Techniques of Data Mining Research," *London Journal of Research in Computer Science and Technology,* vol. 17, no. 1, pp. 7-16, 2017.

[8] A. Urso, A. Fiannaca, M. Rosa, V. Ravì and R. Rizzo, Encyclopedia of Bioinformatics and Computational Biology, Palermo: Elsevier, 2018.

[9] P. Turney, "Cost-sensitive Classification: Empirical Evaluation of a Hybrid Genetic Decision Tree Induction Algorithm," *Journal of Artificial Intelligence Research,* vol. 2, pp. 369-409, 1995.

[10] A. AlMana and M. Aksoy, "An Overview of Inductive Learning Algorithms," *International Journal of Computer Applications,* vol. 88, no. 4, pp. 20-28, 2014.

[11] M. Berry and M. Browne, Lecture Notes in Data Mining, Hackensack: World Scientific, 2006.

[12] H. Kamber, Data Mining: Concepts and Techniques, 2nd ed., Amsterdam: Elsevier, 2007.

[13] Y. Chen, C. Wu and K. Tang, "Time-constrained Cost-sensitive Decision Tree Induction," *Information Sciences,* vol. 354, pp. 140-152, 2016.

[14] R. Roiger and M. Geatz, Data Mining: A Tutorial-Based Primer, Addison Wesley, 2003.

[15] M. Sokolova and G. Lapalme, "A Systematic Analysis of Performance Measures for Classification Tasks," *Information Processing and Management,* vol. 45, no. 4, pp. 427- 437, 2009.

[16] P. Turney, "Types of Cost in Inductive Concept Learning," in *17th International Conference on Machine Learning*, California, 2000, pp. 15-21.

[17] O. Marbán, E. Menasalvas and C. Fernández-Baizán, "A Cost Model to Estimate the Effort of Data Mining Projects (DMCoMo)," *Information Systems,* vol. 3, no. 1, pp. 133-150, 2008.

[18] H. Zhao, X. Li, Z. Xu and W. Zhu, "Cost-sensitive Decision Tree with Probabilistic Pruning Mechanism.," in *International Conference on Machine Learning and Cybernetics*, Guangzhou, 2015, pp. 81-87.

[19] N. Dogan and Z. Tanrikulu, "A Comparative Analysis of Classification Algorithms in Data Mining for Accuracy, Speed and Robustness," *Information Technology and Management,* vol. 14, no. 2, pp. 105-124, 2012.

[20] U. Jensen, P. Kugler, M. Ring and B. Eskofier, "Approaching the Accuracy–cost Conflict in Embedded Classification System Design," *Pattern Analysis and Applications,* vol. 19, no. 3, pp. 839-855, 2015.

[21] L. Deng and J. Song, "Decision Tree Classification Algorithm based on Cost and Benefit Dual-sensitive," in *IEEE International Conference on Electro/Information Technology*, 2014, pp. 320 - 323.

[22] Q. He, X. Zhu, D. Li, S. Wang, J. Shen and Y. Yang, "Cost-effective Big Data Mining in the Cloud: A Case Study with K-means.," in *IEEE 10th International Conference on Cloud Computing (CLOUD)*, 2017, pp. 74-81.

[23] A. Ahmadi, O. Dehzangi and R. Jafari, "Brain-computer Interface Signal Processing Algorithms: A Computational Cost vs. Accuracy Analysis for Wearable Computers.," in *Ninth International Conference on Wearable and Implantable Body Sensor Networks*, London, 2012, pp. 40-45.

[24] M. Bohanec and I. Bratko, "Trading Accuracy for Simplicity in Decision Trees," *Machine Learning,* vol. 15, no. 3, pp. 223-250, 1994.

[25] S. Bhowmick, B. Toth and P. Raghavan, "Towards Low-cost, High-accuracy Classifiers for Linear Solver Selection," in *International Conference on Computational Science, LNCS 5544*, 2009, pp. 463-472.

[26] Q. Zhou, H. Zhou and T. Li, "Cost-sensitive Feature Selection using Random Forest: Selecting Low-cost Subsets of Informative Features," *Knowledge-Based Systems,* vol. 95, pp. 1-11, 2016.

[27] H. Lu, L. Yang, K. Yan, Y. Xue and Z. Gao, "A Cost-sensitive Rotation Forest Algorithm for Gene Expression Data," *Neurocomputing, ,* vol. 228, p. 270–276, 2017.

[28] J. Davis, J. Ha, C. Rossbach, H. Ramadan and E. Witchel, "Cost-sensitive Decision Tree Learning for Forensic Classification," The University of Texas at Austin., 2006.

[29] C. Qiu, L. Jiang and C. Li, "Randomly Selected Decision Tree for Test-cost Sensitive Learning," *Applied Soft Computing,* vol. 53, pp. 27-33, 2017.

[30] "UCI Machine Learning Repository: Data Sets," 2018. [Online]. Available: https://archive.ics.uci.edu/ml/datasets.html. [Accessed 20 July 2018].

[31] H. Lu, L. Yang, K. Yan, Y. Xue and Z. Gao, "A Cost-sensitive Rotation Forest Algorithm for Gene Expression Data," *Neurocomputing,* vol. 228, p. 270–276, 2017.

[32] C. Ling, Q. Yang, J. Wang and S. Zhang, "Decision Trees with Minimal Costs," in *International Conference on Machine Learning*, Canada, 2004, pp. 1-8.

[33] U. M. L. R. Datasets, 2018. [Online]. Available: https://archive.ics.uci.edu/ml/datasets.html.