Approximation Algorithms for Decision Trees

Aline Saettler [1], **Eduardo S. Laber** [1]

[1]Departamento de Informática – PUC-RIO

Abstract. *Decision trees are a central structure in computer science, having applications in many areas. This thesis contributes to the understanding of this important structure by proving by proving theoretical bounds on its behavior and also providing algorithms to construct decision trees with much stronger guarantee and that can cope with more general situations than the solutions available in the literature.*

1. Introduction

Decision trees are a central structure in computer science, having applications in many areas. In complexity theory is is used to prove lower bounds on the running time of computational problems. In machine learning, decision trees as well as ensemble methods that use them (e.g. Random Forests and Gradient Boosted Trees), are among the most popular methods for classification tasks. This thesis contributes to the understanding of this important structure by proving theoretical bounds on its behavior and also providing algorithms to construct decision trees with much stronger guarantees and that can cope with more general situations than the solutions available in the literature.

We consider a very general model of the decision tree construction problem: we have a set of objects $S = \{s_1, \ldots, s_n\}$ which is partitioned into m classes $C_1, \ldots, C_m$. Objects are characterized by the values they take with respect to a set of tests T. Each test $t \in T$ has a finite number of possible values. Moreover, each test t has also an associated rational positive cost $c(t)$, which is charged whenever it is used, and each object $s \in S$ is associated with a probability $p(s)$.

In general lines, the problems we consider in the thesis consist of designing a procedure for discovering the classification of an unknown object in an efficient way, where efficiency is measured with respect to the (expected) cost of the tests used. Tests are performed to acquire information on the object to be classified. Each new test performed (adaptively chosen from T on the basis of the result of the previous tests) reveals the value taken for the given object. Hence, performing a test restricts the set of possible classifications to those of the objects matching the result of the test. The procedure stops when all objects agreeing with the results of the tests performed belong to the same class, which must also be the class of the object that had to be classified. The probability distribution $\mathbf{p}$ over the set of objects reflects the belief about which objects will have to be classified.

A typical application of the above model is to a diagnosis problem. For instance the objects could represent diseases, e.g., $\{flu, dengue, cancer\}$, classified into infectious and non-infectious, e.g., Infectious $= \{flu\}$ and Non-infectious $= \{dengue, cancer\}$. In this case, tests correspond to medical tests with different costs, and the object probabilities refer to the prior knowledge of the incidence of the three possible diseases among the patients who are going to ask to be diagnosed. The goal is, then, to have a strategy that can be used to quickly/cheaply determine if the disease of a patient (which is not known) is infectious or not.

Object	t_1	t_2	Class	Probability
s_1	0	1	C_1	0.4
s_2	1	0	C_2	0.5
s_3	1	1	C_3	0.1

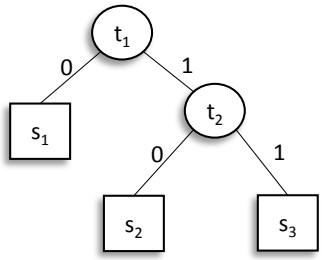

Figure 1. A decision tree D for 3 objects and 2 tests.

Any testing procedure can be represented by a *decision tree*, which is a tree where every internal node is associated with a test. The branches stemming out from a node are associated with the possible outcomes of the test associated with the node. Every leaf is associated with a set of objects that belong to the same class.

Given a decision tree D, rooted at r, we can identify the class of an object s by following a path from r to a leaf as follows: first, we ask for the result of the test associated with r when performed on s; then, we follow the branch of r associated with the result of the test to reach a child r' of r; next, we apply the same steps recursively starting from r'. The procedure ends when a leaf is reached, which determines the class of s. We also say that this is the leaf associated to s.

We define $cost(D, s)$ as the sum of the tests' cost on the path from the root of D to the leaf associated with object s. Then, the *worst-case cost* and the *expected cost* of D are, respectively, defined as

$$cost_W(D) = \max_{s \in S}\{cost(D, s)\} \quad \text{and} \quad cost_E(D) = \sum_{s \in S} cost(D, s)p(s) \tag{1}$$

Figure 1 shows a decision tree D for 3 objects. The instance is given by the set of objects $S = \{s_1, s_2, s_3\}$, with probabilities $p(s_1) = 0.4, p(s_2) = 0.5$ and $p(s_3) = 0.1$. Each object s_i belongs to a different class $C_i = \{s_i\}$ and there are two tests, t_1, t_2, with costs $c(t_1) = 1, c(t_2) = 2$, respectively. Then, for the decision tree D depicted in the figure, we have $cost_W(D) = \max\{1, 1 + 2\} = 3$ and $cost_E(D) = 0.4 \times 1 + 0.5 \times 3 + 0.1 \times 3 = 2.2$.

2. Research Topics

In this thesis, we studied the following topics related to the design of decision trees.

Research Topic 1: Simultaneous optimization of worst-case cost and expected cost. Most of the research on decision tree optimization focuses on building a decision tree that minimizes either the worst-case cost $cost_W$ or the expected cost $cost_E$ [Kosaraju et al. 1999, Adler and Heeringa 2008, Chakaravarthy et al. 2009, Guillory and Bilmes 2009, Guillory and Bilmes 2010, Golovin et al. 2010, Gupta et al. 2010, Bellala et al. 2012]. From an application point of view, the choice of the optimization criterion reflects different assumptions on the data model: a more optimistic perspective on the knowledge of the underlying distribution

might elicit the minimization of the expected cost while a more pessimistic perspective might prefer the conservative minimization of the worst-case cost.

However, the two different optimization criteria can lead to very different trees: a decision tree minimizing the expected cost for a very skewed distribution can have a skewed shape with a very high worst-case cost, even exponentially larger than the worst cost of a decision tree optimized with respect to the worst-case cost. Conversely, optimizing with respect to the worst-case cost can lead to a tree with poor performance in expectation. Examples of this kind of behavior are given in the thesis.

The choice of the "wrong" optimization criterion might have serious consequences in practical applications (see, e.g., [P. Kelle 2014] for such a study in the economics literature). Therefore, it makes sense to look for a trade-off between minimizing these two measures.

Thus, our first and main research topic consisted on investigating the trade-off between the minimization of the worst-case cost and the expected-cost measures for decision trees. More specifically, we addressed the following questions

Q1 Is there an efficient (polynomial) algorithm that builds a decision tree guaranteeing simultaneous approximation for both worst-case and expected cost?

Q2 For every instance I, is there a decision tree whose worst-case cost and expected cost are arbitrarily close, respectively, to the optimal worst-case cost and the optimal expected cost for I?

Note that in the second question we are only concerned with the existence of the decision tree, regardless whether it can be efficiently constructed or not.

Research Topic 2: Decision trees with value dependent costs. In most decision tree problems, a usual assumption is that the testing cost is independent of the test's outcome. However, there are also several scenarios in medical applications where this assumption does not apply. Many diagnostic tests actually consist of a multi-stage procedure, e.g., in a first stage the sample is tested against some reagent to check for the presence or absence of an antigen. If the antigen level is below a certain threshold the test is considered to be negative and no further analysis is performed. Otherwise, the test is necessarily followed by a second stage where several new reagents are used, thus leading to a significantly higher final test cost. Notice that in such a situation no choice can be made by the decision tree strategy between the first and the second stage, so one needs to consider such a two stage procedure as a single test whose cost depends on the outcome. This scenario motivates the following problem:

Q3 Are there efficient algorithms for building decision trees, with provable approximation guarantees when the costs of the tests depend on their outputs?

Research Topic 3: Decision trees with bounded number of misclassifications. In the topics discussed so far we were interested in decision trees that were able to correctly classify each object. However, in order to avoid data overfitting, it is very usual in classification tasks to choose a compact decision tree that is allowed to misclassify some of the objects rather than a more complex one that makes no errors. Therefore, we addressed the following question:

Q4 Given a non-negative integer $k > 0$, are there efficient algorithms for building decision trees with provable approximation guarantees that make at most k classification errors?

3. Our Contributions

In this section we explain our contributions for each of the research topics outlined above.

3.1. Results for Research Topic 1

Simultaneous optimization of worst-case cost and expected cost. Our first result is polynomial algorithm that builds a single decision tree whose worst-case cost and expected cost are at most $O(\log n)$ times larger than the minimum possible worst-case cost and the minimum possible expected cost, respectively. In fact, this is the first guarantee for the expected cost that is independent of the values of the probabilities. Previously, only a $O(\log 1/p_{\min})$ approximation was known [Golovin et al. 2010] and [Bellala et al. 2012], where p_{min} is the minimum positive probability among the objects in S. Moreover, our guarantees are tight for both metrics, that is, unless $\mathcal{P} = \mathcal{NP}$, it is impossible to obtain approximations better than $O(\log n)$ for either metric [Chakaravarthy et al. 2009, Laber and Nogueira 2004]

At a high level, our construction consists of building a root-to-leaf path that splits the input instance into smaller ones, for which decision trees are recursively constructed and attached as children of the nodes in the path. Moreover, such a path is obtained via the combination of approximated solutions for two different submodular optimization programs.

We shall mention that our approach is related to the one used by Gupta *et al.* [Gupta et al. 2010] for obtaining the $O(\log n)$ approximation for the expected cost in the identification problem (a particular case of our problem). However, in addition to solving a more general problem under two metrics simultaneously, our algorithm is much simpler than the one presented in [Gupta et al. 2010]. First, it is more transparently linked to the structure of the problem, which remained somehow hidden in [Gupta et al. 2010] where the result was obtained via an involved mapping from adaptive TSP. Second, our algorithm avoids expensive computational steps and some non-intuitive/redundant steps that are used to select the tests for the root-to-leaf path in the tree. In order to achieve these simplifications we had to prove some non-trivial results such as the Lemma 2 of the thesis that provides an upper bound on the cost of the root-to-leaf path.

Trade-off between worst-case and expected-case minimization. We completely characterize the trade-off curve between optimizing the worst-case cost and expected cost mestrics. More precisely, we show that for every $\rho > 0$ and every instance I there exists a decision tree D with worst-case cost at most $(1+\rho)OPT_W(I)$ and expected cost at most $\left(\frac{1}{1-e^{-\rho}}\right) OPT_E(I)$, where $OPT_W(I)$ (resp. $OPT_E(I)$) denotes the cost of the decision tree with minimum worst-case cost (resp. minimum expected cost) for the instance I. Moreover, we also show that this is the best possible trade-off attainable, in the sense that there are infinitely many instances for which we cannot obtain a decision tree with both worst-case cost smaller than $(1+\rho)OPT_W(I)$ and expected cost smaller than $\frac{1}{1-e^{-\rho}}OPT_E(I)$.

To obtain the upper bound, we present a general procedure that merges decision trees built according to different optimization criteria: given a parameter $\rho > 0$, a decision tree D_W with worst-case cost W and a decision tree D_E with expected cost E, our merging procedure produces a decision tree D with worst-case cost at most $(1+\rho)W$ and expected cost at most $\frac{1}{1-e^{-\rho}}E$. The first upper bound follows from a simple analysis of our procedure. On the other hand, to obtain the latter, we first formulate a max min fraction optimization problem, say $\mathcal{O}$, and we argue that its optimal objective value provides an upper bound on the expected cost. Then, we show that the optimum value of $\mathcal{O}$ can be obtained via the analysis of a certain linear program (LP). Finally, we guess a solution for the LP and prove that it is optimal using LP duality.

For the lower bound, we construct a probability distribution based on the optimal solution of the linear program and use it as a starting point for constructing non-trivial instances that guarantee that the upper bound is essentially tight. This construction is arguably the most involved result in the thesis.

3.2. Results for Research Topic 2

We present algorithms for the minimization of the worst-case cost for the generalized setting where the cost of a test depends of its outcome. We provide a greedy algorithm and prove that it is an $O(\log(n))$-approximation for the case of binary tests. This bound is the best possible under the assumption that $\mathcal{P} \neq \mathcal{NP}$. We then present a second algorithm that attains an $O(n)$ approximation for multiway tests and value-dependent costs.

3.3. Results for Research Topic 3

In the setting where k misclassifications are allowed, we design algorithms with provable approximations for *oblivious decision trees*, a class of decision trees where every node in the same level is associated with the same attribute. Oblivious decision trees have been studied in the context of feature selection [Kohavi and Li 1995].

We present an algorithm that, given a parameter $0 < \epsilon < 1/2$, builds an oblivious decision tree with worst-case cost at most $(3/(1-2\epsilon))\ln(n)OPT(I)$ and makes at most (k/ϵ) errors, where $OPT(I)$ and n are, respectively, the optimal worst-case cost and the number of objects for instance I. Our tree is obtained via a randomized rounding of a linear program. The logarithmic factor in the cost of the tree is the best possible attainable, even for $k = 0$, unless $\mathcal{P} = \mathcal{NP}$.

4. Products

The results developed in this thesis were reported in the following papers:

1. Aline M. Saettler, Eduardo S. Laber, Ferdinando Cicalese: Trading off Worst and Expected Cost in Decision Tree Problems. ISAAC 2015: 223-234; Qualis B1.
2. Ferdinando Cicalese, Eduardo S. Laber, Aline M. Saettler: Diagnosis determination: decision trees optimizing simultaneously worst and expected testing cost. Int. Conference on Machine Learning (ICML) 2014: 414-422; Qualis A1.
3. Aline M. Saettler, Eduardo S. Laber, Ferdinando Cicalese: Trading Off Worst and Expected Cost in Decision Tree Problems. Algorithmica 79(3): 886-908 (2017), Special issue of ISAAC 2015. This is an extended version of ISAAC paper including all proofs and some extra results; Qualis A2.

4. Ferdinando Cicalese, Eduardo S. Laber, Aline M. Saettler: Decision Trees for Function Evaluation: Simultaneous Optimization of Worst and Expected Cost. Algorithmica 79(3): 763-796 (2017). The extended version of the ICML paper including all the proofs; Qualis A2.
5. Aline M. Saettler, Eduardo S. Laber, Ferdinando Cicalese: Approximating decision trees with value dependent testing costs. Inf. Processing Letters (2016); Qualis A2.
6. Aline M. Saettler, Eduardo S. Laber and Felipe A. M. Pereira. Decision Tree Classification with Bounded Number of Errors. Inf. Processing Letters (2017). Qualis A2.

We shall highlight that Paper 1 was the recipient of the **Best Paper Award** of the ISAAC conference, out of 180 submissions. ISAAC is a traditional conference in the area of Algorithms which is currently in its 28th edition.

We shall also mention that ICML, the conference where Paper 2 was published, is the premier machine learning conference, arguably one of the 'hottest' fields in Computer Science nowadays. Moreover, in 2014, there were two cycles of acceptance and the paper was one of the 85, out of 577 submitted, that was accepted in the first cycle.

References

Adler, M. and Heeringa, B. (2008). Approximating optimal binary decision trees. APPROX '08 / RANDOM '08, pages 1–9.

Bellala, G., Bhavnani, S. K., and Scott, C. (2012). Group-based active query selection for rapid diagnosis in time-critical situations. *IEEE Trans. Inf. Theor.*, 58(1):459–478.

Chakaravarthy, V. T., Pandit, V., Roy, S., and Sabharwal, Y. (2009). Approximating decision trees with multiway branches. ICALP, pages 210–221.

Golovin, D., Krause, A., and Ray, D. (2010). Near-optimal bayesian active learning with noisy observations. In *NIPS 2010*.

Guillory, A. and Bilmes, J. (2009). Average-case active learning with costs. In *ALT'09*, pages 141–155.

Guillory, A. and Bilmes, J. (2010). Interactive submodular set cover. In *ICML*, pages 415–422.

Gupta, A., Nagarajan, V., and Ravi, R. (2010). Approximation algorithms for optimal decision trees and adaptive tsp problems. In *ICALP'10*, pages 690–701.

Kohavi, R. and Li, C.-H. (1995). Oblivious decision trees, graphs, and top-down pruning. In *IJCAI*, pages 1071–1079.

Kosaraju, Przytycka, and Borgstrom (1999). On an optimal split tree problem. In *WADS: 6th Workshop on Algorithms and Data Structures*.

Laber, E. S. and Nogueira, L. T. (2004). On the hardness of the minimum height decision tree problem. *Discrete Applied Mathematics*, 144(1):209–212.

P. Kelle, H. Schneider, H. Y. (2014). Decision alternatives between expected cost minimization and worst case scenario in emergency supply second revision. *International Journal of Production Economics*, pages 250–260.