Experimental Study of Totally Optimal Decision Trees

Abdulla Aldilaijan[1], Mohammad Azad[2], and Mikhail Moshkov[2]

[1] Brown University
Department of Computer Science
Providence, Rhode Island 02912 USA
abdulla_aldilaijan@brown.edu
[2] King Abdullah University of Science and Technology (KAUST)
Computer, Electrical and Mathematical Sciences & Engineering Division
Thuwal 23955-6900, Saudi Arabia
{mohammad.azad, mikhail.moshkov}@kaust.edu.sa

Abstract. In this paper, we experimentally study the existence of totally optimal decision trees (which are optimal relative to two cost functions simultaneously) for nine decision tables from UCI Machine Learning Repository. Such trees can be useful when we consider decision trees as algorithms for problem solving or as a way for knowledge representation. For cost functions, we use depth, average depth, and number of nodes. We study not only exact but also approximate decision trees based on five uncertainty measures: entropy, Gini index, misclassification error, relative misclassification error, and number of unordered pairs of rows with different decisions. To investigate the existence of totally optimal trees, we use an extension of dynamic programming that allows us to make multi-stage optimization of decision trees relative to two cost functions. Experimental results show that totally optimal decision trees exist in many cases. However, the behavior of graphs describing how the number of decision tables with totally optimal decision trees depends on their accuracy is mainly irregular, but we can see some trends, in particular, an upward trend when accuracy is decreasing.

Keywords: decision tree, uncertainty measure, cost function, totally optimal tree, multi-stage optimization

1 Introduction

In this paper, we experimentally investigate the existence of totally optimal decision trees that are optimal relative to two cost functions simultaneously. For example, decision trees that have minimum depth and minimum number of nodes at the same time. Such decision trees can be useful if they are considered as algorithms for problem solving (see, for example, [7]) or as a way for knowledge representation (see, for example, [3, 11]). When decision trees are considered as algorithms for problem solving we would like to simultaneously minimize its time complexity (depth or average depth) and its space complexity (number of nodes).

When decision trees are considered as a way for knowledge representation we would like to simultaneously minimize the number of nodes (to make the decision tree more understandable) and the depth (to make understandable conjunctions of conditions corresponding to the paths from the root to terminal nodes).

We have two powerful tools for the study of totally optimal decision trees that are based on extensions of dynamic programming. These are (i) multi-stage optimization of decision trees relative to two criteria [8] and (ii) construction of the set of Pareto optimal points for bi-criteria optimization problem (totally optimal decision tree exists if and only if there is exactly one Pareto optimal point) [2]. In this paper, we use the first tool which is less time consuming. For experimenting, we use Dagger system created in KAUST to implement various extensions of dynamic programming [1, 2].

One of the main goals of this paper is to experimentally study how the existence of totally optimal decision trees depends on the accuracy of such trees. We work with nine decision tables from UCI Machine Learning Repository [6]. For cost functions, we consider the depth h, average depth h_{avg}, and number of nodes L which describe the time complexity in the worst case, time complexity in the average case, and space complexity of decision trees, respectively. We study not only exact but also approximate decision trees defined based on five uncertainty measures: entropy ent, Gini index $gini$, misclassification error me, relative misclassification error rme, and number of unordered pairs of rows with different decisions R. We consider 101 values of the threshold α from 0.00 to 1.00 with the step 0.01. This threshold describes the accuracy of decision trees: $\alpha = 0.00$ means 100 percent accuracy (exact decision trees), and $\alpha = 1.00$ means 0 percent accuracy.

Experimental results show that totally optimal decision trees exist in many cases. The behavior of graphs describing how the number of decision tables with totally optimal decision trees depends on decision tree accuracy is mainly irregular. However, we notice an upward trend when accuracy is decreasing.

One of the main areas of applications of the obtained results (existence of totally optimal decision trees in many cases) is the rough set theory [9, 10] in which decision trees are widely used.

Note that totally optimal decision trees computing Boolean functions were considered in [5] including exact and approximate decision trees defined based on relative misclassification error rme uncertainty measure. The following two statements were proved in that paper:

- As $n \to \infty$, almost all Boolean functions with n variables have totally optimal exact decision trees relative to depth and number of nodes.
- As $n \to \infty$, almost all Boolean functions with n variables have totally optimal exact decision trees relative to depth and average depth.

Note also that totally optimal exact decision trees for decision tables with many-valued decisions relative to different combinations of cost functions were studied in [4].

This paper consists of four sections. In Sect. 2, we consider main notions and tools. Section 3 is devoted to the consideration of experimental results. Section 4 contains short conclusions.

2 Main Notions and Tools

In this section, we discuss the notions of decision table, uncertainty measure, decision tree, cost function, and totally optimal decision tree. We consider also a way to prove the existence of totally optimal decision trees.

2.1 Decision Tables and Uncertainty Measures

A *decision table* is a rectangular table T with $n \geq 1$ columns filled with numbers from the set $\omega = \{0, 1, 2, \ldots\}$ of nonnegative integers. Columns of the table are labeled with *conditional* attributes $f_1, \ldots, f_n$. Rows of the table are pairwise different, and each row is labeled with a number from ω which is interpreted as a decision (a value of the *decision* attribute d). Rows of the table are interpreted as tuples of values of conditional attributes. We denote by $\mathcal{T}$ the set of all decision tables.

A decision table is called *empty* if it has no rows. The table T is called *degenerate* if it is empty or all rows of T are labeled with the same decision. Let $D(T)$ be the set of decisions attached to rows of T. We denote by $N(T)$ the number of rows in the table T and, for any $t \in \omega$, we denote by $N_t(T)$ the number of rows of T labeled with the decision t. By $mcd(T)$ we denote the *most common decision* for T which is the minimum decision t_0 from $D(T)$ such that $N_{t_0}(T) = \max\{N_t(T) : t \in D(T)\}$. If T is empty then $mcd(T) = 0$.

For any conditional attribute $f_i \in \{f_1, \ldots, f_n\}$, we denote by $E(T, f_i)$ the set of values of the attribute f_i in the table T. We denote by $E(T)$ the set of conditional attributes for which $|E(T, f_i)| \geq 2$.

Let T be a nonempty decision table. A *subtable* of T is a table obtained from T by removal of some rows. Let $f_{i_1}, \ldots, f_{i_m} \in \{f_1, \ldots, f_n\}$ and $a_1, \ldots, a_m \in \omega$. We denote by $T(f_{i_1}, a_1) \ldots (f_{i_m}, a_m)$ the subtable of the table T containing the rows from T which at the intersection with the columns $f_{i_1}, \ldots, f_{i_m}$ have numbers $a_1, \ldots, a_m$, respectively.

Let $\mathbb{R}$ be the set of real numbers and $\mathbb{R}_+$ be the set of all nonnegative real numbers. An *uncertainty measure* is a function $U : \mathcal{T} \to \mathbb{R}$ such that $U(T) \geq 0$ for any $T \in \mathcal{T}$, and $U(T) = 0$ if and only if T is a degenerate table. The following functions (we assume that, for any empty table, the value of each of the considered functions is equal to 0) are uncertainty measures:

- Misclassification error $me(T) = N(T) - N_{mcd(T)}(T)$.
- Relative misclassification error $rme(T) = (N(T) - N_{mcd(T)}(T))/N(T)$.
- Entropy $ent(T) = -\sum_{t \in D(T)} (N_t(T)/N(T)) \log_2(N_t(T)/N(T))$.
- Gini index $gini(T) = 1 - \sum_{t \in D(T)} (N_t(T)/N(T))^2$.
- Function R where $R(T)$ is the number of unordered pairs of rows of T labeled with different decisions (note that $R(T) = N(T)^2 gini(T)/2$).

2.2 Decision Trees and Cost Functions

Let T be a decision table with n conditional attributes $f_1, \ldots, f_n$.

A decision tree over T is a finite directed tree with root in which nonterminal nodes are labeled with attributes from the set $\{f_1, \ldots, f_n\}$, terminal nodes are labeled with numbers from ω, and, for each nonterminal node, edges starting from this node are labeled with pairwise different numbers from ω.

Let Γ be a decision tree over T and v be a node of Γ. We define now a subtable $T(v) = T_\Gamma(v)$ of the table T. If v is the root of Γ then $T(v) = T$. Let v be not the root of Γ and $v_1, e_1, \ldots, v_m, e_m, v_{m+1} = v$ be the directed path from the root of Γ to v in which nodes $v_1, \ldots, v_m$ are labeled with attributes $f_{i_1}, \ldots, f_{i_m}$ and edges $e_1, \ldots, e_m$ are labeled with numbers $a_1, \ldots, a_m$, respectively. Then $T(v) = T(f_{i_1}, a_1) \ldots (f_{i_m}, a_m)$.

Let U be an uncertainty measure and α be a real number such that $0 \leq \alpha \leq 1$.

A decision tree Γ over T is called a (U, α)*-decision tree for* T if, for any node v of Γ,

- If $U(T(v)) \leq \alpha U(T)$ then v is a terminal node which is labeled with the number $mcd(T(v))$.
- If $U(T(v)) > \alpha U(T)$ then v is a nonterminal node labeled with an attribute $f_i \in E(T(v))$, and if $E(T(v), f_i) = \{a_1, \ldots, a_t\}$ then t edges start from the node v which are labeled with $a_1, \ldots, a_t$, respectively.

Let α, β be real numbers such that $0 \leq \alpha < \beta \leq 1$ and Γ be a (U, α)-decision tree for T. Then it is possible that Γ is not a (U, β)-decision tree for T.

A cost function for decision trees is a function $\psi(T, \Gamma)$ which is defined on pairs of decision table T and a decision tree Γ for T, and has values from $\mathbb{R}_+$.

We now consider examples of cost functions for decision trees:

- *Depth* $h(T, \Gamma) = h(\Gamma)$ of a decision tree Γ for a decision table T is the maximum length of a path in Γ from the root to a terminal node.
- *Average depth* $h_{avg}(T, \Gamma)$ of a decision tree Γ for a decision table T. Let $Row(T)$ be the set of rows of T and, for any row r of T, let $l_\Gamma(r)$ be the length of a path in Γ from the root to a terminal node v such that the row r belongs to $T_\Gamma(v)$. Then $h_{avg}(T, \Gamma) = (\sum_{r \in Row(T)} l_\Gamma(r))/N(T)$.
- *Number of nodes* $L(T, \Gamma) = L(\Gamma)$ of a decision tree Γ for a decision table T.

2.3 Totally Optimal Decision Trees

Let U be an uncertainty measure and α be a real number such that $0 \leq \alpha \leq 1$.

For a cost function ψ, we denote by $\psi^{U,\alpha}(T)$ the minimum cost of a (U, α)-decision tree for T relative to the cost function ψ. Let ψ_1, ψ_2 be cost functions. A (U, α)-decision tree Γ for T is called a *totally optimal* (U, α)*-decision tree for* T *relative to the cost functions* ψ_1, ψ_2 if $\psi_1(T, \Gamma) = \psi_1^{U,\alpha}(T), \psi_2(T, \Gamma) = \psi_2^{U,\alpha}(T)$, i.e., Γ is optimal relative to ψ_1, ψ_2 simultaneously.

Assume that $\psi_1 \in \{h_{avg}, L\}$ and $\psi_2 \in \{h, h_{avg}, L\}$. We now describe how to recognize the existence of a (U, α)-decision tree for T which is a totally optimal (U, α)-decision tree for T relative to the cost functions ψ_1, ψ_2.

First, we apply to T the procedure of optimization relative to ψ_2 [8]. As a result, we obtain the number $\psi_2^{U,\alpha}(T)$. Next, we sequentially apply to T the procedures of optimization relative to the cost functions ψ_1, ψ_2 [8]. As a result, we obtain the number $\psi_{1,2}^{U,\alpha}(T)$ which is the minimum value of the cost function ψ_2 among all (U, α)-decision trees for T with the minimum value of the cost function ψ_1 (see [8]). One can show that a totally optimal (U, α)-decision tree for T relative to the cost functions ψ_1, ψ_2 exists if and only if $\psi_{1,2}^{U,\alpha}(T) = \psi_2^{U,\alpha}(T)$.

Let us consider an example of a totally optimal decision tree. The disjunction $x_1 \vee x_2 \vee x_3$ can be represented as a decision table with three conditional attributes x_1, x_2, x_3 and eight rows corresponding to tuples from $\{0,1\}^3$ that are labeled by decisions – values of $x_1 \vee x_2 \vee x_3$ on these tuples. One can show that the decision tree depicted in Fig. 1 is a totally optimal $(U, 0)$-decision tree relative to h, h_{avg}, and L for the considered decision table. This is a totally optimal decision tree relative to h, h_{avg}, and L computing disjunction $x_1 \vee x_2 \vee x_3$.

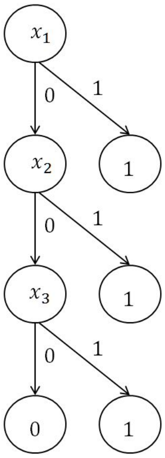

Fig. 1. Totally optimal decision tree relative to h, h_{avg}, and L computing disjunction $x_1 \vee x_2 \vee x_3$

3 Experimental Results

In this section, we consider results of computer experiments with nine decision tables from [6]. Before the experimental work, some preprocessing procedures are performed. A conditional attribute is removed if it has unique value for each row. The missing value for an attribute is filled up with the most common value for this attribute. In some tables, there are equal rows with, possibly, different decisions. In this case, each group of equal rows is replaced with a single row from the group with the most common decision for this group. As a result, we obtain consistent decision tables without missing values (a decision table is called consistent if it has no equal rows with different decisions).

The nine decision tables from [6] used in experiments are described in Table 1. The first column 'Decision table' refers to the name of the decision table from [6], the second column 'Rows' refers to the number of rows, and the last column 'Attributes' refers to the number of conditional attributes.

Table 1. Decision tables used in experiments

Decision table	Rows	Attributes
BALANCE-SCALE	625	5
BREAST-CANCER	266	10
CARS	1728	7
HAYES-ROTH-DATA	69	5
HOUSE-VOTES-84	279	17
LYMPHOGRAPHY	148	19
SOYBEAN-SMALL	47	36
TIC-TAC-TOE	958	10
ZOO-DATA	59	17

For each decision table T, each uncertainty measure $U \in \{ent, gini, me, rme, R\}$, each $\alpha \in \{0.00, 0.01, 0.02, \ldots, 1.00\}$, and each pair of different cost functions $\psi_1, \psi_2 \in \{h, h_{avg}, L\}$, we check if there exists a totally optimal (U, α)-decision tree for T relative to the cost functions ψ_1, ψ_2.

The results of experiments can be found in Figs. 2, 3, 4, 5, and 6 for uncertainty measures *ent*, *gini*, *me*, *rme*, and R, respectively.

In each figure,

- The graph 'havg_h' shows, for the uncertainty measure U considered in this figure and each $\alpha \in \{0.00, 0.01, 0.02, \ldots, 1.00\}$, the number of decision tables T for each of which there exists a totally optimal (U, α)-decision tree for T relative to the cost functions h_{avg}, h.
- The graph 'havg_L' shows, for the uncertainty measure U considered in this figure and each $\alpha \in \{0.00, 0.01, 0.02, \ldots, 1.00\}$, the number of decision tables T for each of which there exists a totally optimal (U, α)-decision tree for T relative to the cost functions h_{avg}, L.
- The graph 'L_h' shows, for the uncertainty measure U considered in this figure and each $\alpha \in \{0.00, 0.01, 0.02, \ldots, 1.00\}$, the number of decision tables T for each of which there exists a totally optimal (U, α)-decision tree for T relative to the cost functions L, h.

The obtained experimental results show that totally optimal decision trees exist in many cases. There are some additional observations:

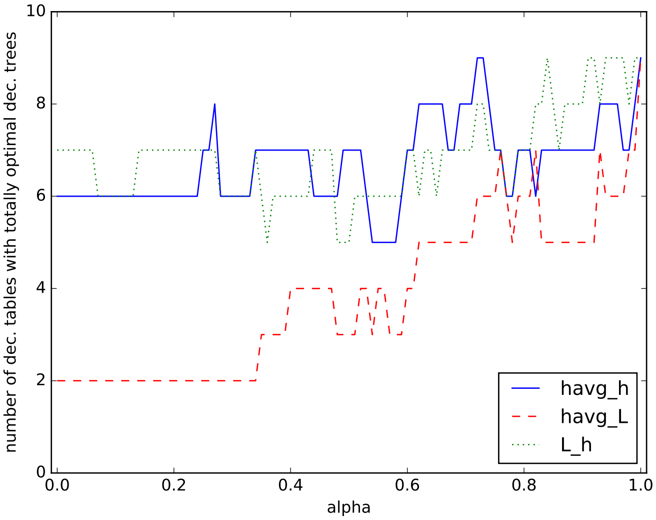

Fig. 2. Number of decision tables with totally optimal decision trees for uncertainty measure *ent*

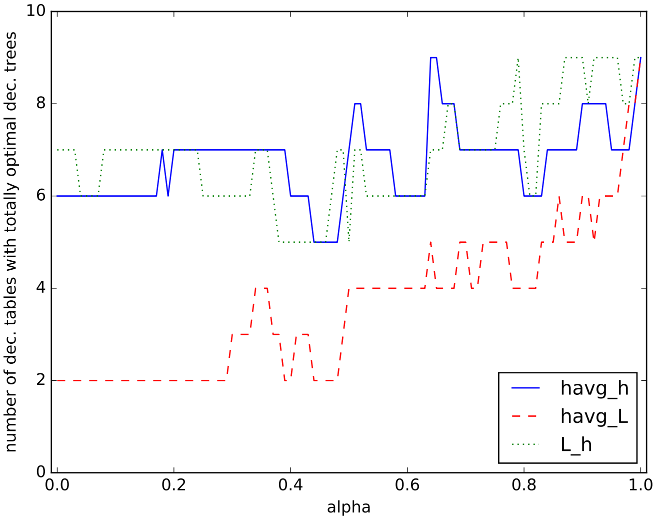

Fig. 3. Number of decision tables with totally optimal decision trees for uncertainty measure *gini*

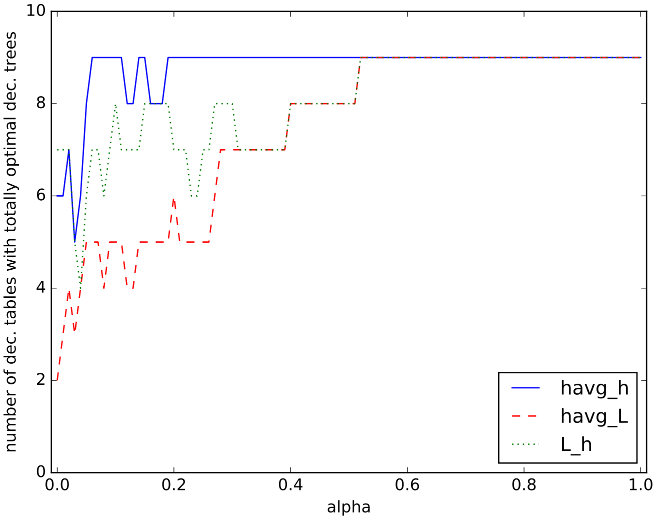

Fig. 4. Number of decision tables with totally optimal decision trees for uncertainty measure *me*

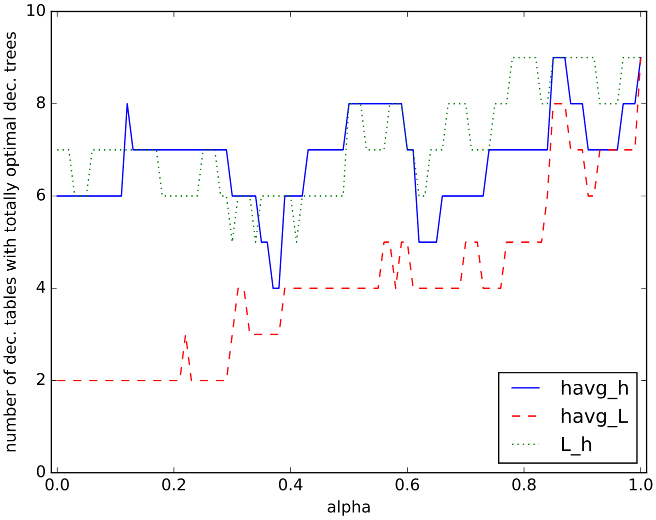

Fig. 5. Number of decision tables with totally optimal decision trees for uncertainty measure *rme*

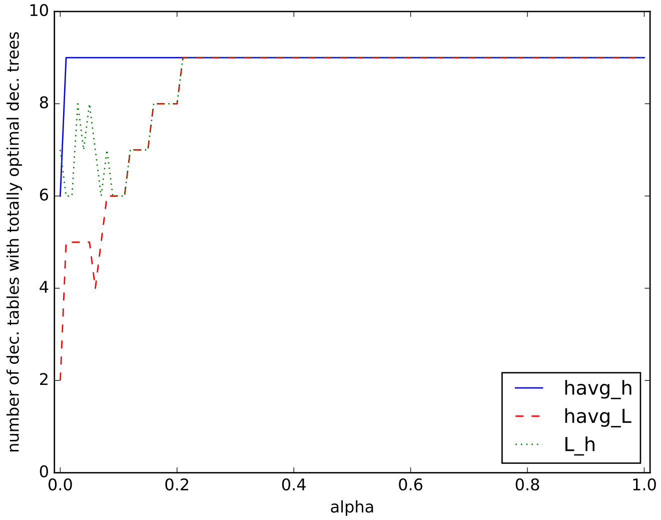

Fig. 6. Number of decision tables with totally optimal decision trees for uncertainty measure R

- In Figs. 2-6, the behavior of lines is mainly irregular. The only exception is the line 'havg_h' for the uncertainty measure R.
- Some lines have upward trend: lines 'havg_h' and 'L_h' for uncertainty measures me and R; lines 'havg_L' for each of the considered uncertainty measures.
- In almost all cases, the lines 'havg_L' are below the lines 'havg_h' and 'L_h'. This result is consistent with the results obtained in [5] for totally optimal decision trees computing Boolean functions.

4 Conclusions

In this paper, we experimentally studied the existence of totally optimal decision trees (exact and approximate) for nine decision tables from UCI Machine Learning Repository [6]. Totally optimal decision trees exist in many cases. Lines describing how the number of decision tables with totally optimal decision trees grows with the decreasing of tree accuracy have often upward trend but the behavior of these lines is mainly irregular. In the future, we are planning to continue the study of totally optimal trees since they can be useful in various applications related to problem solving and knowledge representation.

Acknowledgments

Research reported in this publication was supported by King Abdullah University of Science and Technology (KAUST).

The authors are indebted to the anonymous reviewers for useful suggestions.

References

1. Alkhalid, A., Amin, T., Chikalov, I., Hussain, S., Moshkov, M., Zielosko, B.: Dagger: a tool for analysis and optimization of decision trees and rules. In: Ficarra, F.V.C., Kratky, A., Veltman, K.H., Ficarra, M.C., Nicol, E., Brie, M. (eds.) Computational Informatics, Social Factors and New Information Technologies: Hypermedia Perspectives and Avant-Garde Experiencies in the Era of Communicability Expansion, pp. 29–39. Blue Herons (2011)
2. Alkhalid, A., Amin, T., Chikalov, I., Hussain, S., Moshkov, M., Zielosko, B.: Optimization and analysis of decision trees and rules: dynamic programming approach. International Journal of General Systems 42(6), 614–634 (2013)
3. Azad, M., Chikalov, I., Hussain, S., Moshkov, M.: Multi-pruning of decision trees for knowledge representation and classification. In: 3rd IAPR Asian Conference on Pattern Recognition, ACPR 2015, Kuala Lumpur, Malaysia, November 3-6, 2015. pp. 604–608. IEEE (2015)
4. Azad, M., Moshkov, M.: Multi-stage optimization of decision and inhibitory trees for decision tables with many-valued decisions. European Journal of Operational Research 263(3), 910–921 (2017)
5. Chikalov, I., Hussain, S., Moshkov, M.: Totally optimal decision trees for Boolean functions. Discrete Applied Mathematics 215, 1–13 (2016)
6. Lichman, M.: UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences (2013), http://archive.ics.uci.edu/ml
7. Moshkov, M.: Time complexity of decision trees. In: Peters, J.F., Skowron, A. (eds.) Trans. Rough Sets III, Lecture Notes in Computer Science, vol. 3400, pp. 244–459. Springer (2005)
8. Moshkov, M., Chikalov, I.: Consecutive optimization of decision trees concerning various complexity measures. Fundam. Inform. 61(2), 87–96 (2004)
9. Pawlak, Z.: Rough Sets – Theoretical Aspect of Reasoning About Data. Kluwer Academic Publishers, Dordrecht (1991)
10. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Information Sciences 177(1), 3–27 (2007)
11. Rokach, L., Maimon, O.: Data Mining with Decision Trees: Theory and Applications. World Scientific Publishing, River Edge, NJ (2008)