Comprehensibility of Classification Trees – Survey Design Validation

Rok Piltaver[1], **Mitja Luštrek**, **Matjaž Gams**[1]
Department of Intelligent Systems
Jozef Stefan Institute
Jamova cesta 39, 1000 Ljubljana, Slovenia
{rok.piltaver, mitja.lustrek, matjaz.gams}@ijs.si

Sanda Martinčić - Ipšić
Department of Informatics
University of Rijeka
Radmile Matejčić 2, 51000 Rijeka, Croatia
smarti@inf.uniri.hr

Abstract: *Classifier comprehensibility is a decisive factor for practical classifier applications; however it is ill-defined and hence difficult to measure. Most algorithms use comprehensibility metrics based on classifier complexity – such as the number of leaves in a classification tree – despite evidence that they do not correspond to comprehensibility well. A classifier comprehensibility survey was therefore designed in order to derive exhaustive comprehensibility metrics better reflecting the human sense of classifier comprehensibility. This paper presents an implementation of a classification-tree comprehensibility survey based on the suggested design and empirically verifies the assumptions on which the survey design is based: the chosen respondent performance metrics measured while solving the chosen tasks can be used to indirectly but objectively measure the influence of chosen tree properties on their comprehensibility.*

Keywords*: classification tree, comprehensibility, understandability, survey*

1 Introduction

In data mining the comprehensibility is the ability to understand the output of an induction algorithm [11]. It is also referred to as interpretability [14] or understandability [1] and has been recognized as an important property since the early days of machine learning research [16, 18]. Although research in the last three decades is more focused on improving predictive performance of learned classifiers, comprehensibility is reported as the decisive factor when machine learning approaches are applied in industry [10]. Examples of application areas in which comprehensibility is emphasized are medicine, credit scoring, churn prediction, bioinformatics, etc.[5].

A comprehensibility metric is needed in order to compare performance of learning systems and as a (part of) heuristic function used by a learning algorithm [6, 20]. However, comprehensibility is ill-defined [10] and subjective [1, 8, 13], therefore it is difficult to measure. Instead of measuring comprehensibility directly, most algorithms

[1] The authors are also affiliated with Jozef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia

use model complexity instead; e.g. number of leaves in a tree [13], number of conjunctions in a rule set [19], number of connections in an artificial neural network [9, 12]. Although model complexity is related to comprehensibility [9], empirical user studies [1] reveal that comprehensibility measures based solely on model complexity are over-simplistic and produce incomprehensible models [5].

In order to derive exhaustive comprehensibility metrics better reflecting the human sense of classifier comprehensibility, a classifier comprehensibility survey has been designed recently [17]. The user-survey design is based on the related work and follows the observation that the comprehensibility is in the eye of the beholder [15]. The ultimate goal of the survey is to obtained insights into respondents' judgments about classifier comprehensibility and to define a good comprehensibility metric.

This paper presents an implementation of classification-tree comprehensibility survey according to the suggested survey design and empirically verifies the assumptions, which the survey design is based on: the performance of respondent solving the survey tasks depends on the classification tree comprehensibility; the observed classification tree properties influence comprehensibility and the range of classification trees and questions used in the survey is broad enough to measure the influence. Finally, the survey design is based on the assumption that the objectively measured respondent performance parameters (time to answer, probability of correct answer) are related to the subjective perception of classifier comprehensibility. The list of tasks (i.e. parts of the survey) designed to measure comprehensibility includes activities: classify instance, explain classification, validate classifier, and discover new knowledge from classifier. The considered properties of classification trees are: number of leaves, depth of the tree, depth of leaves relevant to a given survey question, branching factor, and tree presentation style. In addition, the paper considers a few minor survey design choices: the order of tasks and the range and explanations of scales used to collect subjective opinions.

The paper is organized as follows. Section 2 explains the survey implementation in detail: the chosen dataset and the list of used classification trees are described in Section 2.1, each task and the set of questions for the task are described in Section 2.2. Section 3 describes the group of survey respondents, verifies the survey design assumptions and suggests improvements of the survey. Section 4 concludes the paper with discussion of the results. Appendix contains figures of all classification trees used in the survey.

2 Survey implementation

The classification-tree comprehensibility survey is implemented as an online survey in order to facilitate accurate measurements of respondent performance, remote participation, automatic checking of the correctness of answers and saving them in a database. Each question of the survey corresponds to a web page, which is dynamically generated using PHP scripts. We designed the survey around six tasks, each composed of several questions of the same type but related to different trees or parts of a tree.

The first four tasks measure the performance of respondents asked to answer questions about given classification trees. The difficulty of the questions in each task depends on comprehensibility of the classification tree – an approach advocated by some researchers [1, 8]. Each of the first four tasks is based on [3], which reports that comprehensibility is required to explain individual instance classifications, validate the classifier, and discover new knowledge. The second part of the survey measures subjective opinion about comprehensibility of classification trees rated on the scales suggested in [17].

2.1 Dataset and classification trees

All the survey questions are related to the Zoo domain from the UCI Machine Learning Repository [2]. The domain was chosen because it meets all the requirements stated in the survey design: it is familiar and interesting to the general and heterogeneous population of respondents, but still broad and rich enough to enable learning a range of classifiers with various properties. The Zoo domain requires only elementary knowledge about animal properties expressed with 16 (mostly binary) attributes: hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, venomous, fins, legs (numeric), tail, domestic, and catsize. The attribute *animal name* from the original dataset is not used in the survey because it is a unique identifier of an instance. The seven classes given as numeric attribute in the original Zoo domain are referred to using the following names instead: mammals (41 instances), birds (20), fish (13), mollusc (10), insect (8), reptile (5), and amphibian (4).

The classification trees used in the survey (Figures 3-20) are learned and visualized using the Orange tool [4] as suggested by [17]. The basic tree shown in Figure 4 is learned using the Classification Tree widget in Orange with the following parameters: gini index as the attribute selection criterion, pruning with m-estimate where $m = 2$, and minimum of 4 instances in a leaf. It contains 7 leaves, 6 binary attributes (resulting in branching factor 2), and depth 5. Choosing other attribute selection criterion would only change the order of the same set of attributes in the tree.

The survey is based on trees with three different sizes: small trees with 3 or 4 leaves, the basic tree with 7 leaves, and big trees with 9 or 10 leaves. The big trees (Figure 5) were learned on the original dataset using modified pruning parameters. The two small trees are shown in Figure 3. They were learned using a reduced set of attributes because such trees are more natural and comprehensible then the pruned versions of the basic tree; which is caused by the uneven class distribution in the dataset. In this way an unnatural classifiers – a possible survey design error [1] – was avoided. The sets of attributes used to learn the small trees were chosen so that the leaves of the learned trees correspond to clusters obtained using hierarchical clustering (by Euclidian distance and complete linkage criterion) on the original dataset. As a result, each leaf in the learned trees contains animals from a single class or classes of animals with similar properties.

In addition, the survey analyses influence of tree branching factor on comprehensibility, therefore the trees described above were modified to obtain trees with branching factor 3 (Figures 6-8) and 4 (Figures 9-11). This was achieved by adding new aggregate attributes, which were computed as natural combinations of original attributes selected so that the structure of the trees with higher branching factor is as similar as possible to the structure of the trees with branching factor 2. Note that two version of the tree with 9 leaves (Figure 8) and branching factor 3 were obtained, yet they differ in the maximal depth of the tree.

In addition to the trees described above, their modified representations (Figures 12-17) are used in the last task as discussed in the description of the *compare* task. The trees shown in Figures 18-20 are used only in the *discover* task and were obtained using the Orange Interactive Tree Builder as discussed in the paragraph about the task.

2.2 Tasks and questions

Each of the tasks starts with an instructions page. It includes an explanation of the task on an example: figures and explanations showing how to solve the task step by step. The instruction page includes an example question in exactly the same format as the questions in the task but on a different domain. The respondents are allowed to start answering the questions only after correctly answering the test question. The test

question was added to the survey because significantly longer times of answering the first question compared to the subsequent questions were observed in the initial testing.

The first task – **classify** – asks a respondent to classify an instance according to a given classification tree – the same task type was used in [1, 8]. When a webpage with a survey question (Figure 1) is opened, only the instructions and footer of the page are shown. The instruction for the first task says: "*Classify the instance described in the table on the left using the classification tree on the right.*" The footer contains respondent's name, name of the task and current question number, link to a help page, and support e-mail address. After reading the instructions, the respondent clicks the "*Start solving*" button. This calls a JavaScript function that starts the timer and displays the question data and the answer form. The question data consists of a table with ten alphabetically sorted attribute-value pairs shown on the left and an image of a classification tree in SVG format shown on the right (Figure 1). The answer form is displayed below the question data; the label says: "*The instance belongs to class*:" and is followed by a drop-down menu offering the names of the seven classes as an answer to the question. Whenever the respondent changes a value of an answer form field, the time and action type are recorded. When the respondent clicks the "*Submit answer*" button, the answer fields are disabled, the timer is stopped and the time needed to answer the question is calculated.

In addition, the respondent is asked to give the subjective judgment of questions difficulty on the scale with five levels. Each label of the scale is accompanied with an explanation in order to prevent variation in subjective interpretations of the scale:

- Very easy – I answered without any problems in less than 5 seconds.
- Easy – I found the answer quite quickly and without major problems.
- Medium.
- Difficult – I had to think hard and am not sure if I answered correctly.
- Very difficult – Despite thinking very hard my answer is likely to be wrong.

After rating the question's difficulty, the respondent clicks the "*Next question*" button, which calls a JavaScript function that assigns the calculated performance values to the hidden form fields in order to pass them to the PHP script that stores the data in the database and displays the next question. One question per each leaf depth was asked for each tree shown in Figures 1-11, which amounts to 30 questions. The number of questions in other tasks was reduced because the first group of respondents reported that the number of questions should not be any higher in order to prevent them from becoming tired or bored while answering the survey.

The second task – **explain** – asks a respondent to answer which attributes' values must be changed or retained in order for the tree to classify the given instance into another class. This corresponds to explaining an individual instance classification. For example, which habits (values of attributes) would a patient with high probability of getting cancer (class) have to change in order to stay healthy? The web pages for questions in the second task are similar to the ones in the first task with the following differences. The instruction for the task says for example: "*What should be the values of the attributes for the example (listed in the table on the left) that is classified as fish so that the classification tree (shown on the right) would classify it as amphibian? If the value of an attribute is not relevant for the classification into the new class, maintain the default choice (i.e. "irrelevant"); otherwise change it to "unchanged" if the value must remain as it is, or "different" if it needs to be changed.*" The table of attribute-value pairs includes an additional column named "*new value*" with drop-down menus with three choices: irrelevant (default value), different, and unchanged.

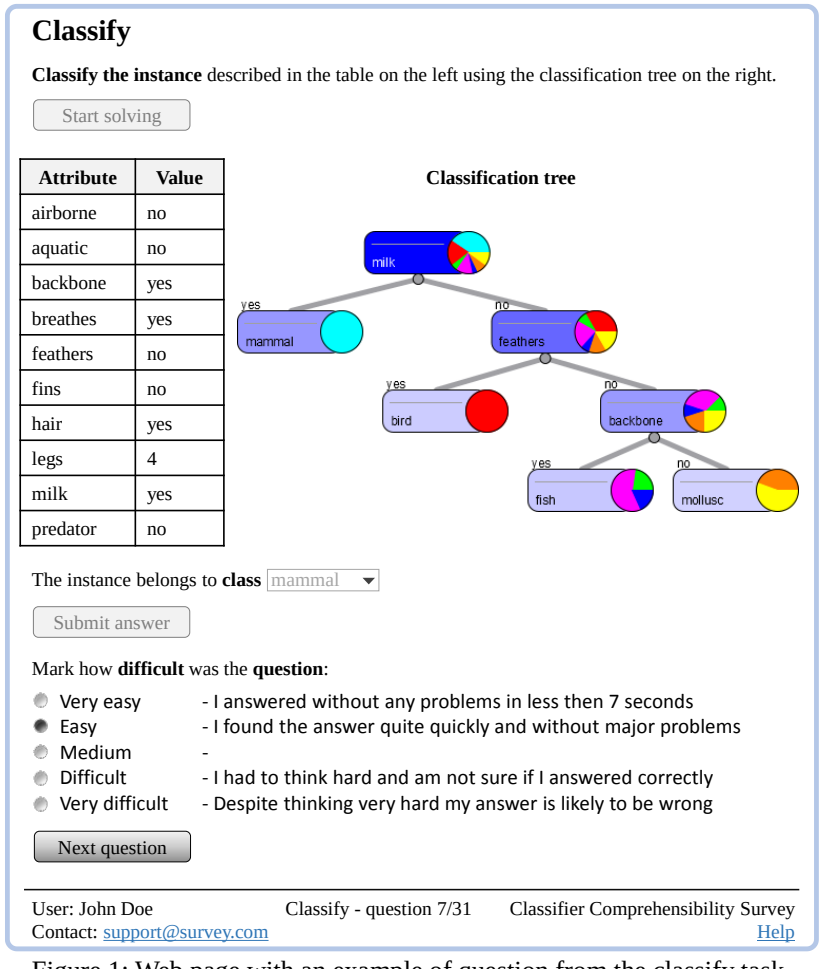

Figure 1: Web page with an example of question from the classify task.

The third task – **validate** – asks a respondent to check whether a statement about the domain is confirmed or rejected by the presented tree – this corresponds to validating a part of the classification tree. Similar questions were also asked in [8]. The question web pages for the third task are similar to the ones in the first task with the following differences. The instruction for the task says: "*Does the classification tree agree with the statement below it?*" The answer form is replaced with a statement, e.g.: "*For animals from class reptile it holds that aquatic = yes and feathers = no*", followed by a drop-down menu with two possible answers: yes or no. Each statement is composed of two conditions (for two different attributes), except when knowledge about animal class belonging to a leaf at depth one is verified, e.g. questions about mammals for the trees in Figure 3. The number of conditions to be validated is limited to two in order to observe the relative difficulty of validating domain knowledge corresponding to leaves at various depths regardless of the number of attributes relevant for classification of the instances belonging to those leaves – this is already considered in the classify and explain tasks.

The fourth task – **discover** – asks the respondent to find a property (attribute-value pair) that is unusual for instances from one class, which corresponds to discovering new knowledge from the classification tree. Rather than rediscovering known relations

between attributes and classes, the questions ask to find an unusual property for a class of animals – a common property of outliers, e.g. it is unusual for a mammal to lay eggs. Therefore special trees offering the information about the outliers were constructed. The outliers in the dataset were first identified using the Outliers widget in Orange. After that the trees that misclassify the outliers were constructed using the Interactive Tree Builder widget in Orange. Some parts of the tree were selected manually in order to place the attribute that splits the outliers from the normal instances (the attribute expected as the answer) at a desired depth in the tree. The remaining parts of the tree were built using the automatic tree learning widget function. In this way the four trees shown in Figures 18-20 were constructed. In contrast with the trees used in other tasks, each of their nodes includes a number of instances of a given class belonging to the node – this is used by respondents to identify the common and rare properties of animals belonging in to a class. A question for each node containing outlier instances observable in the trees shown in Figures 18-20 was asked amounting to 8 questions.

The fifth task – **rate** – requests the user to give the subjective opinion about the classification trees on a scale with five levels:

- Very easy to comprehend – I can use the knowledge represented by the classification tree as soon as I read it for the first time; I can easily remember it; I can explain it to another person without looking at the figure.
- Easy to comprehend – I can use the knowledge represented by the classification tree after studying it for some time; I can remember it with some effort; I can explain it to another person without looking at the figure, but I am not sure that the person will fully understand it.
- Comprehensible – Without long study I can use most of the knowledge represented by the classification tree, but need the figure for some details; it would be difficult to remember it; I can explain it to another person if I see the figure while the other person does not see it, but that person is unlikely to fully understand it.
- Difficult to comprehend – Without long study I can use only some of the knowledge represented by the classification tree, and need the figure for the rest; It would be very difficult to remember it; I can explain the outline to another person if I see the figure while the other person does not see it.
- Very difficult to comprehend – I need the figure to use the knowledge represented by the classification tree; it would be extremely difficult to remember it; I can explain it to another person only if we both see the figure.

Each label of the scale is accompanied with an explanation in order to prevent variation in subjective interpretations of the scale. The web pages of questions in the fourth task are again similar to the ones in the first task with some differences. Namely, the instruction for the task says: "*How comprehensible is the tree shown below?*" Additionally, the answer form is replaced with the table containing the comprehensibility scale and a radio button for each of the comprehensibility levels. No attribute-value table is shown in this task. The respondents were asked to rate the comprehensibility of each tree 12 trees shown in Figures 3-11.

Task six – **compare** – asks the respondents to rate which of the two classification trees shown side by side is more comprehensible on the scale with four levels. The instructions say: "*The following question type measures the subjective opinion about the tree comprehensibility; there are no correct and wrong answers and time needed to answer each question is not important at all. Compare the classification trees in the pictures and choose the answer that best fits your opinion*." Clicking on a tree opens a new window showing full-screen picture of the selected tree and a back button. This is needed because text in nodes of some of the bigger trees becomes difficult to read when

the trees are scaled to half of the screen width. The answer is given by clicking a radio button in front of one of the following answers:

- The tree on the left is much more comprehensible: answering questions similar to the ones in this survey about the other tree would be significantly more difficult and would definitely take more time.
- The tree on the left is more comprehensible: answering questions similar to the ones in this survey about the other tree would be more difficult or take more time.
- The tree on the left is slightly more comprehensible: although one tree is slightly more comprehensible, answering questions similar to the ones in this survey about the other tree would not be more difficult.
- The trees are equally comprehensible: I don't notice any difference in comprehensibility between the two trees.

The three answers preferring the tree on the right are also offered, because the trees in each question are positioned to the left or the right side randomly. One of the trees in this task is already used in the previous five tasks – serving as a known frame of reference – while the other one is a previously unseen tree with the same content but represented in different style. Figure 12 shows a version of the tree without pie charts that represent learning dataset class distribution in the nodes of the tree. The pie-charts enable the user to find leaves with the same class (same prevalent colour of the pie-chart) quickly and provide additional information in easy-to-read graphical form. Figure 13 shows a version of a tree with meaningless attribute and attribute value names, which makes it more difficult to comprehend the tree, because domain knowledge cannot be used and because remembering numeric attribute names and values is more difficult than remembering known and meaningful semantic names. Figure 14 shows a version of a tree as drawn by Weka data mining program [7]; it is compared to the Orange representation of the same trees. Figure 15 shows a version of a tree in plain text output as obtained from Weka. The layout of the tree nodes in the plain text version is in general more difficult to read than the Orange output. Figure 16 shows a version of the tree with different layout of the tree branches than the rest of the trees; the other trees used in the survey place shallow subtrees to the left and deep subtrees to the right, which corresponds to the natural reading order in the western culture – from top to bottom and from left to right. The subtrees in the tree in Figure 16, on the other hand, are scrambled regardless of their depth. The tree in Figure 17 uses the default Orange format, but is learned using a subset of attributes with less evident relation with the class, e.g. without milk (which is true only for mammals) and feathers (which is true only for birds) attributes. Nevertheless, the classification accuracy of the tree is similar to classification accuracy of the tree with the same number of leaves learned on the entire dataset. In addition, three questions comparing comprehensibility of the trees with the same representation were asked: comparing the trees with the same number of nodes but different branching factor (Figure 5 vs. bottom tree in Figure 8), comparing the trees with the same branching factor but different number of nodes (Figure 5 vs. the right tree in Figure 3), and comparing the trees with the same branching factor and number of nodes but different structure resulting in different depth of the tree (trees in Figure 8).

3 Survey design verification

In order to verify the survey design that is based on the related work, the survey was filled in by group of 18 students and teaching assistants from Department of Informatics – University of Rijeka and the collected answers were analysed.

Figure 2 shows the range of trees used in the survey according to the comprehensibility rated by the respondents (the rate task). The percentages of answers

for each of the five comprehensibility levels and for each tree used in tasks classify, explain, validate, and compare are shown. The average rating for the most comprehensible tree on a scale from 1 (very easy to comprehend) to 5 (very difficult to comprehend) is 1.44 and the average rating for the least comprehensible tree is 3.83. The tree comprehensibility is well spread from very easy to medium or difficult to comprehend trees; therefore the range of trees is validated as appropriate. The survey implementation lacks a tree that is very difficult to comprehend; however, it is not possible to learn a trees that is very difficult to comprehend in the Zoo domain. Using a more complex domain would allow the construction of such trees, but might not be able to provide trees that are very easy to comprehend and not over-simplistic at the same time. In addition, classification trees are known as one of the most comprehensible classification models and might not be very difficult to comprehend even in the demanding domains. Figure 2 shows that respondent's ratings of the tree comprehensibility agree well, therefore the scale is confirmed as appropriate.

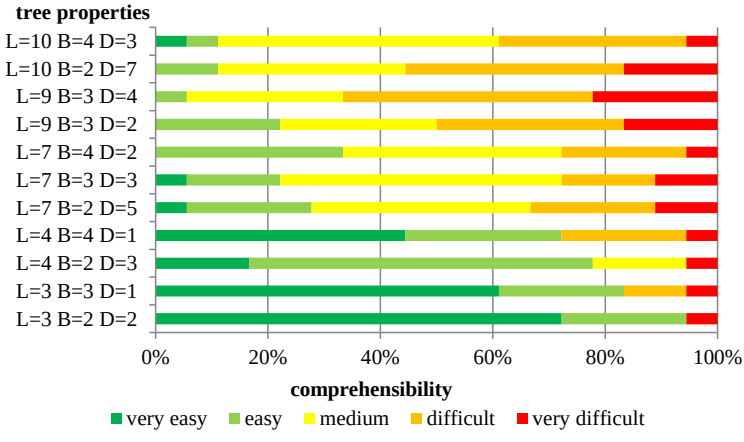

Figure 2: Subjective opinion about comprehensibility of various classification trees (L is the number of leaves in a tree, B is the branching factor, and D is the depth of a tree).

The data about the four tasks (classify, explain, verify, and discover) for which the respondent performance was measured is in Table 1. The first task (classify) was confirmed as the easiest: 98.1 % answers in the task were correct and the difficulty of individual questions in the classify task were rated from minimum 1.33 (very easy) to maximum 2.06 (easy) with an average of 1.63 on the scale from 1 (very easy) to 5 (very difficult). The fourth task (discover) was confirmed as the most difficult: 62.5 % of the questions were answered correctly and the difficulty of questions was rated from 2.00 (easy) to 2.82 (medium). The difficulty of the second and third tasks (explain and verify) are positioned between the difficulty of classify and discover tasks, as expected and incorporated into the survey design.

According to percent of the correct answers, the explain task is easier; however according to the rated question difficulty the verify task is slightly easier. The times of solving the questions from the explain tasks are longer then in verify and classify tasks, however this is partially caused by more mouse clicks required to answer a question. Based on the additional statistics it was estimated that the respondents needed about 1 to 2.5 seconds to select an answer from a drop-down menu. If this time is subtracted from the measured total time to answer a question, the time needed to reason about the most difficult question in the explain task (6 drop-down menu selections) is between 35 and 45 seconds (17 to 21 seconds per question on average). This suggests that difficulty of

the explain task is similar to the difficulty of the verify task in terms of the time needed to answer the question as well. The above observations verify that the tasks are ordered from the easiest to the most difficult ones (the order of explain and verify tasks could be switched as well).

The range of questions in each task according to their difficulty is broad: there are substantial differences between minimum and maximum time needed to solve a question in each task (see Table 1). The difficulty of questions ranges for about one level on the difficulty scale within each task. The range is even greater if individual respondents' rates are considered instead of the average rate over all the respondents.

Table 1: Overall statistics for the tasks and questions used in the survey.

task	time-difficulty correlation	correct-difficulty correlation	correct answers (%)	question time (ms)			question difficulty		
				min	avr	max	min	avr	max
classify	0.772	0.094	98.1	8.6	16.7	31.4	1.33	1.63	2.06
explain	0.957	-0.432	92.0	8.4	24.7	50.8	1.50	2.02	2.61
verify	0.720	-0.658	96.4	7.6	14.9	22.1	1.50	1.95	2.33
discover	0.901	0.548	62.5	12.7	28.6	44.6	2.00	2.53	2.82

The correlation between the rated difficulty of question and the two objective measure of question difficulty – the time needed to answer a question and the percent of correct answers – were calculated in order to verify whether they can be used to estimate the question difficulty and the tree comprehensibility objectively. The time to answer a question (averaged over all respondents that answered correctly) is clearly correlated with the rated question difficulty: the correlation ranges from 0.720 to 0.957 across the four tasks. The correlation of the percent of correctly answered questions and the rated question difficulty is almost zero for the classify task, because almost all questions were answered correctly; the task is so easy that the percent of correct answers does not change over the trees used in the survey. For the explain and the verify tasks the correlation is -0.432 and -0.658 respectively – this means that respondents correctly answered fewer questions that they rated as more difficult compared to the questions rated as easier. Interestingly, the correlation is positive in the discover task in which only 62.5 % of questions were answered correctly. If the respondents who did not know how to solve the task (rated all the questions as very difficult and answered most of them incorrectly) and the respondents who did not understand the task (rated questions as very easy or easy but answered almost all of them incorrectly) are removed, the correlation drops to 0.135. The few remaining respondents are mostly data mining experts, therefore they rated all the questions in the discover task as easy or of medium difficulty. This suggests that the survey should be slightly modified in order to be more understandable for the non-expert population.

Correlations of tree properties and the respondents' performance as well as their subjective opinions about comprehensibility of various trees is analysed in order to validate the importance and interestingness of the observed tree parameters. Correlation between the number of leaves in a tree and the rated comprehensibility of the tree is 0.951, the correlation of the number of leaves in a tree and the rated question difficulty ranges from 0.452 to 0.802 across the first four tasks, and the correlation between the number of leaves in a tree and the time needed to correctly answer the question ranges from 0.418 to 0.663 across the first four tasks. In addition the respondents rated a tree with 4 leaves as more comprehensible then tree with 10 leaves with rate 2.89 on a scale from 1 (same comprehensibility) to 4 (much more comprehensible). This supports the hypothesis that increasing number of leaves in a tree decreases its comprehensibility. Similar conclusion can be drawn for the depth of the tree; however the correlations are

lower for this parameter. Interestingly, the correlations with the branching factor are negative and close to zero, although trees with high branching factor were expected to be less comprehensible. There are two possible reasons for this. First, the difference between branching factor 2 and 4 used in the survey is not very important, therefore trees with higher branching factor should be considered in the survey. Second, increasing the branching factor does not decrease the comprehensibility because it decreases the depth of the tree and the leaves with the question answers at the same time. This explanation is supported by [19], which advocates deep model theories, which correspond to the aggregate attributes used in the survey to produce the trees with branching factor higher than two. In any case, the influence of the branching factor on the comprehensibility of classification trees should be investigated further. Another interesting conclusion of the analysis is that the depth of the question is even more correlated to respondents' performance than the number of leaves in the tree: the correlations range from 0.606 to 0.943. Therefore more emphasis should be given to this parameter in further studies. The related work [1, 8] considers only the comprehensibility of a classification tree as a whole instead of emphasizing the parts of the tree that classify more instances, are used more often, or contain instances whose correct classification and explanation is more important for the user. Finally, the presentation of classification tree clearly influences their comprehensibility and should be studied in detail. As expected, the respondents rated a simple text representation of a tree much more difficult to comprehend than a tree output produced by Orange (rate 3.72 on scale from 1 to 4). The second most important presentation property turns out to be the meaningfulness of class names and attribute values and names (rate 2.44). The arrangement of branches and Weka output versus the Orange output (rates 1.89 and 1.83) influences the comprehensibility ratings as well.

4 Conclusion

The paper presents an implementation of tree comprehensibility survey according to classifier comprehensibility survey design [17]. The analysis of the data obtained form 18 respondents (mainly CS students and some DM experts) supports that the design of tasks (classify instance, explain classification, validate classifier, and discover new knowledge from classifier) and questions is appropriate to measure the classification-tree comprehensibility. The investigated properties of classification trees (quantitative measures: number of leaves, depth of the tree and depth of relevant leaves, branching factor; and tree presentation style) show to be relevant for tree comprehensibility evaluation as well. Furthermore, the obtained results supports that objectively measured respondent performance parameters (time to answer, probability of correct answer) can be used to estimate the comprehensibility of classification trees. In addition, a few minor survey design choices are confirmed to be correct. The data provided by the 18 respondents suffices to validate the survey design choices; however, the implementation of the survey with few minor improvements (e.g. clearer instructions in the discover task, additional questions related to tree presentation in the compare task, questions related to additional trees: less comprehensible and with branching factor more than 4) should be performed with more respondents in order to obtain enough data to perform statistical analysis about the influence of various classification tree parameters on tree comprehensibility. Finally, in the future work, data mining methods will be employed in comprehensibility evaluation task, since they might prove useful in explaining the interplay of various parameters, and deriving a formal model of classification-tree comprehensibility.

9 References

[1] Allahyari, H.; Lavesson, N. User-oriented Assessment of Classification Model Understandability, 11th Scandinavian Conference on AI, pages 11-19, 2011.

[2] Bache, K.; Lichman, M. UCI Machine Learning Repository, http://archive.ics.uci.edu/ml, downloaded: July 2014.

[3] Craven, M. W.; Shavlik, J. W. Extracting Comprehensible Concept Representations from Trained Neural Networks. Working Notes on the IJCAI'95 WS on Comprehensibility in ML, pages 61-75, 1995.

[4] Demšar, J.; Curk,T.; Erjavec, A. Orange: Data Mining Toolbox in Python. Journal of Machine Learning Research, 14(Aug):2349−2353, 2013.

[5] Freitas, A. A. Comprehensible classification models – a position paper. ACM SIGKDD Explorations, 15 (1): 1-10, 2013.

[6] Giraud-Carrier, C. Beyond predictive accuracy: what? In Proceedings of the ECML-98 Workshop on Upgrading Learning to Meta-Level: Model Selection and Data Transformation, pages 78-85, 1998.

[7] Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA Data Mining Software: An Update. SIGKDD Explorations, 11(1):10-18 2009.

[8] Huysmans, J.; Dejaeger, K.; Mues, C.; Vanthienen, J.; Baesens, B. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. Decision Support Systems, 51(1):141-154, 2011.

[9] Jin, Y. Pareto-Based Multiobjective Machine Learning – An Overview and Case Studies, IEE transactions on systems, man, and cybernetics-part c: applications and reviews, 28(3):397-415, 2008.

[10] Kodratoff, Y. The comprehensibility manifesto, KDD Nuggets, 94:9, 1994.

[11] Kohavi, R. Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid. Proceedings of the 2nd Int. Conf. on KD and DM, pages. 202-207, 1996.

[12] Liu, G. P.; Kadirkamanathan, V. Learning with multi-objective criteria. In Proceedings of IEE Conference on Artificial Neural Networks, pages 53-58, 1995.

[13] Maimon, O.; Rokach, L. Data Mining and Knowledge Discovery Handbook. Springer, 2005.

[14] Maimon, O.; Rokach, L. Decomposition Methodology for Knowledge Discovery and Data Mining: Theory and Applications, World Scientific Publishing Company, 2005.

[15] Martens, D.; Vanthienen, J.; Verbeke, W.; Baesens, B. Performance of classification models from a user perspective. Decision Support Systems, 51(4):782-793, 2011.

[16] Michalski, R. A theory and methodology of inductive learning. Artificial Intelligence 20:111-161, 1983.

[17] Piltaver, R.; Luštrek, M.; Gams, M.; Martinčić - Ipšić, S. Comprehensibility of classification trees – survey design. In Proceedings of 17th International multiconference Information Society, pages 70-73, Ljubljana, Slovenia, 2014.

[18] Quinlan, J.R. Some elements of machine learning. In Proceedings of 16th International Conference on ML (ICML-99), pages 523-525, Bled, Slovenia, 1999.

[19] Sommer, E. An approach to quantifying the quality of induced theories. In Proceedings of the IJCAI Workshop on ML and Comprehensibility, 1995.

[20] Zhou, Z. H. Comprehensibility of data mining algorithms. Encyclopaedia of Data Warehousing and Mining. Idea Group Publishing, USA, 2005, pages 190-195.

Appendix: classification trees used in the survey

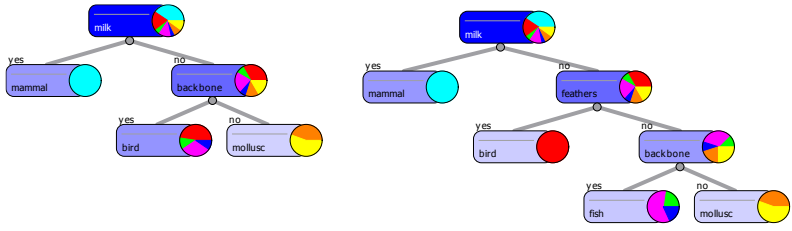

Figure 3: trees with 3 or 4 leaves and branching factor 2.

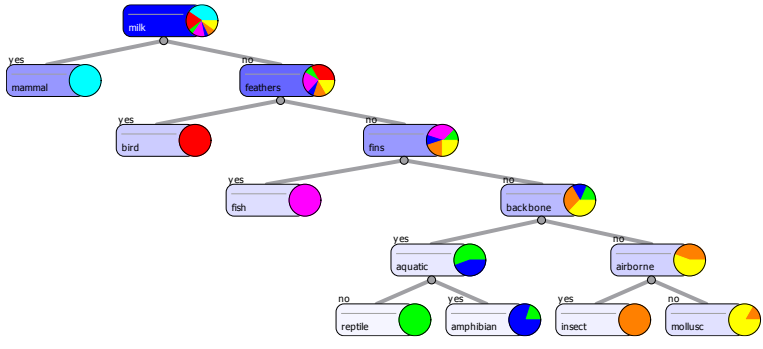

Figure 4: tree with 7 leaves and branching factor 2 – learned using the default parameters of the Classification Tree widget in Orange.

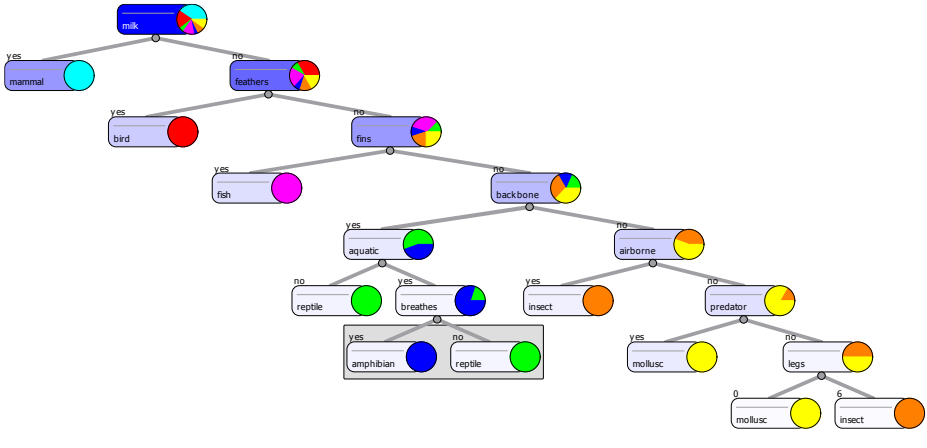

Figure 5: trees with 9 (without the subtree marked with grey rectangle) or 10 leaves and branching factor 2 – unpruned version of the tree in Figure 4.

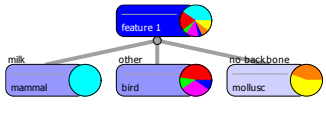

Figure 6: version of the left tree in Figure 3 with branching factor 3.

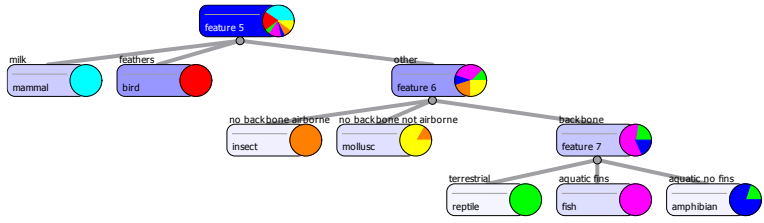

Figure 7: version of the tree in Figure 4 with branching factor 3.

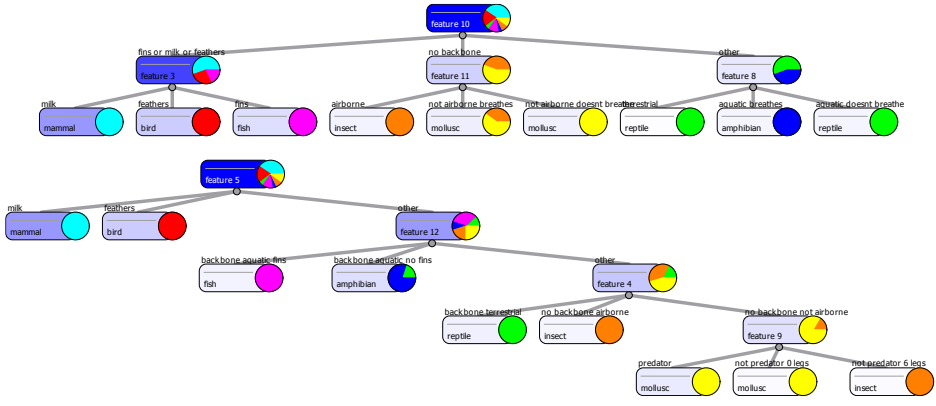

Figure 8: two version of the tree in Figure 5 with branching factor 3.

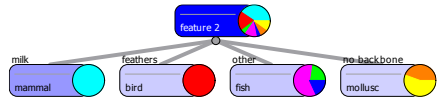

Figure 9: version of the right tree in Figure 3 with branching factor 4.

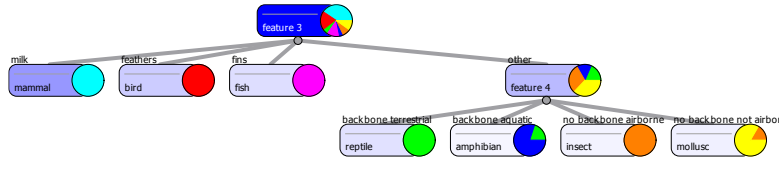

Figure 10: version of the tree in Figure 4 with branching factor 4.

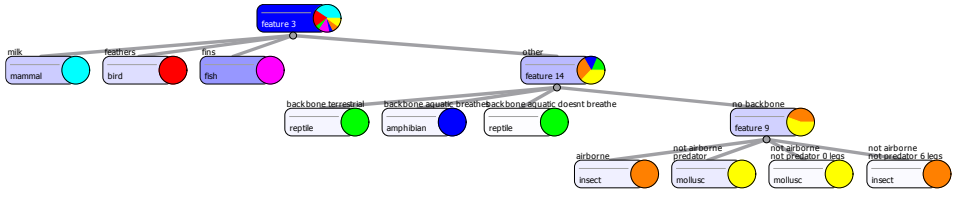

Figure 11: version of the tree in Figure 5 with branching factor 4.

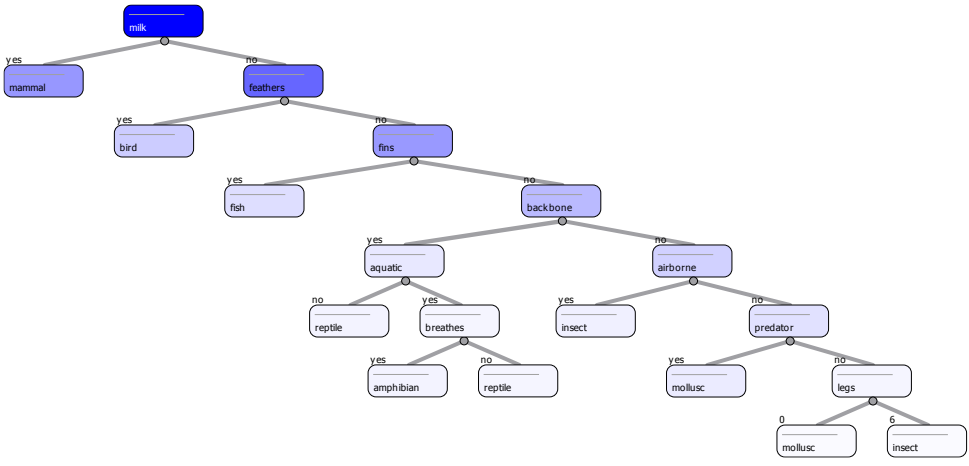

Figure 12: A version of the tree shown in Figure 5 without pie charts.

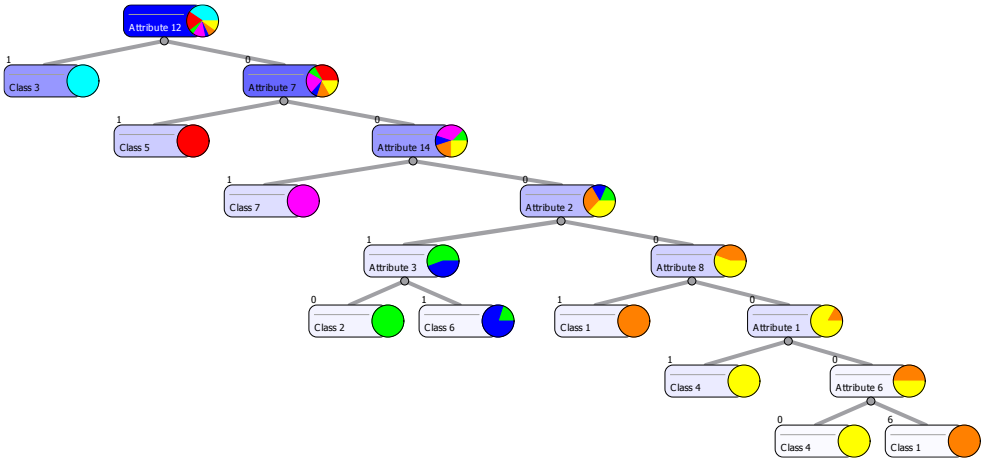

Figure 13: A version of the tree shown in Figure 5 with meaningless names of attributes and attribute values.

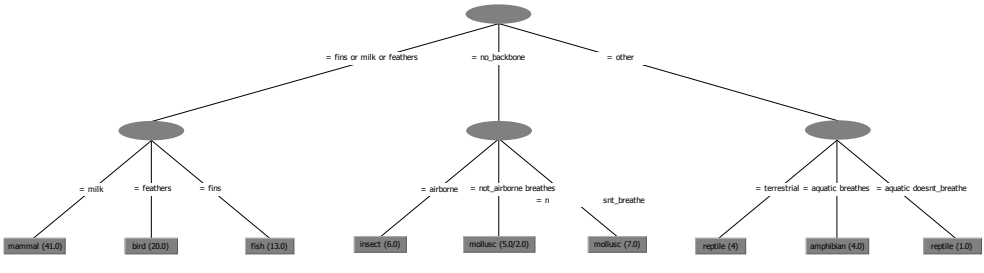

Figure 14: A version of the top tree shown in Figure 8 in Weka representation.

```
milk = no
|   feathers = no
|   |   fins = no
|   |   |   backbone = yes
|   |   |   |   aquatic = no: reptile (4.0)
|   |   |   |   aquatic = yes: amphibian (5.0/
|   |   |   backbone = no
|   |   |   |   airborne = no: mollusc (12.0/2
|   |   |   |   airborne = yes: insect (6.0)
|   |   fins = yes: fish (13.0)
|   feathers = yes: bird (20.0)
milk = yes: mammal (41.0)
```

Figure 15: A version of the tree shown in Figure 4 in plain-text format.

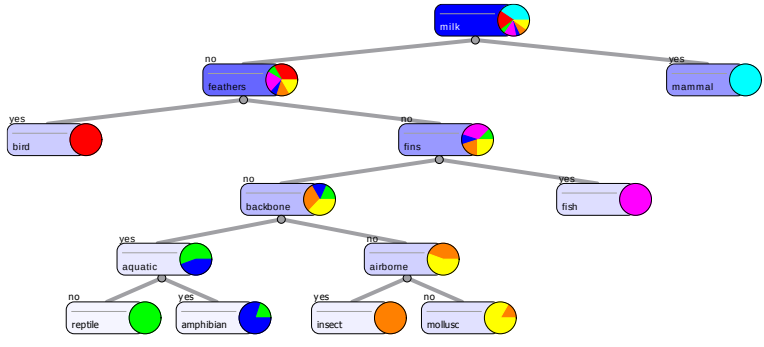

Figure 16: A version of the tree shown in Figure 4 with different layout of tree branches.

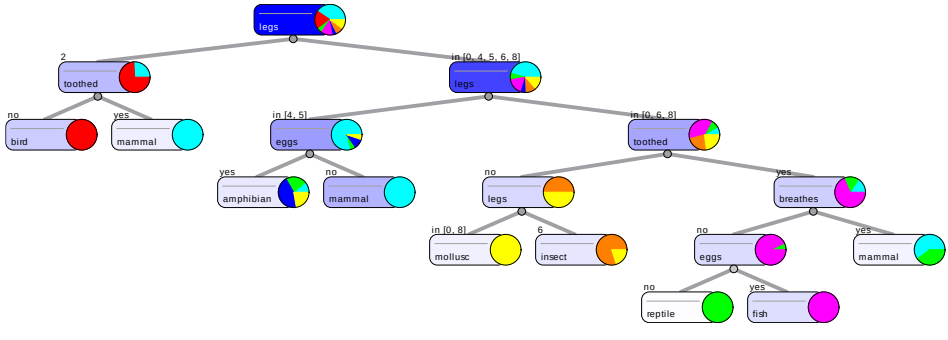

Figure 17: A tree with 9 leaves and branching factor 2 (same as the tree in Figure 5) learned using a subset of attributes with less evident relation with the class.

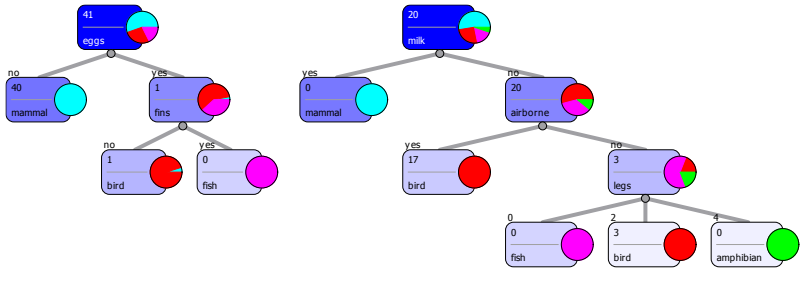

Figure 18: Small trees showing an unusual property of mammals (left) and birds (right).

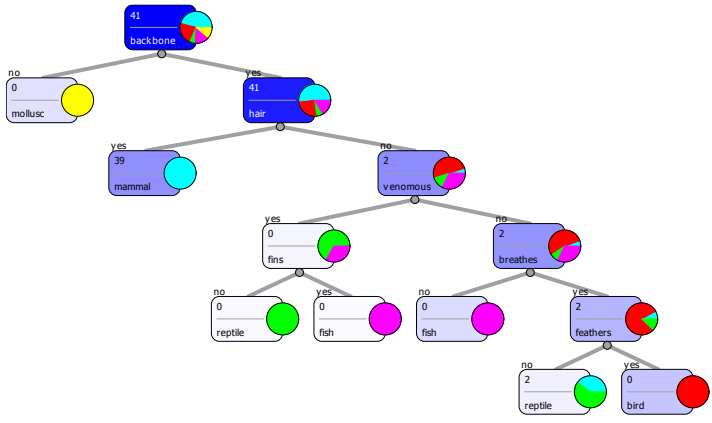

Figure 19: Medium tree showing unusual properties of mammals and fish (numbers in nodes correspond to the number of mammals in each node).

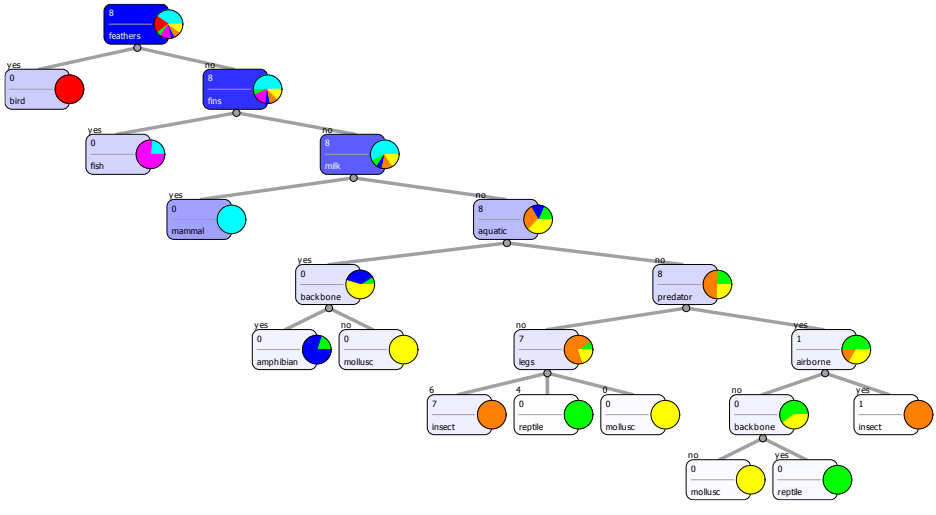

Figure 20: Big tree showing unusual properties of mammals, insects, and reptiles (numbers in nodes correspond to the number of insects in each node).