

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/261343348>

# How to Compare and Interpret Two Learnt Decision Trees from the Same Domain?

Conference Paper · March 2013

DOI: 10.1109/WAINA.2013.201

---

CITATIONS

2

---

READS

1,045

1 author:



Petra Perner

Institute of Computer Vision and applied Computer Sciences IBal

203 PUBLICATIONS 1,766 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



Worldcongress The Frontier of Intelligent Data and Signal Analysis DSA 2018 [View project](#)



Flaw detection in radiographic images [View project](#)

## How to compare and interpret two learnt Decision Trees from the same Domain?

Petra Perner

Institute of Computer Vision and Applied Computer Sciences, IBal

Postbox 30 11 14, 04251 Leipzig

[pperner@ibai-institut.de](mailto:pperner@ibai-institut.de), [www.ibai-institut.de](http://www.ibai-institut.de)

### Abstract

*Data mining methods are widely used across many disciplines to identify patterns, rules or associations among huge volumes of data. Decision tree induction such as C4.5 is the most preferred method for classification since it works well on average regardless of the data set being used. The resulting decision tree has explanation capability but problems arise if the data set has been collected at different times or is enlarging and the decision tree induction process has been repeated. The resulting tree will change and the expert is questioning the trustworthiness of the result. That brings us to the problem of comparing two decision trees in accordance with its explanation power. In this paper, we present a method how to compare two decision trees and how to interpret the change of the structure and the attributes in the decision tree.*

### 1. Introduction

Data mining methods are widely used across many disciplines to identify patterns, rules or associations among huge volumes of data. Different methods can be applied to accomplish this. While in the past mostly black box methods such as neural nets and support vector machines (SVM) have been heavily used in technical domains, methods that have explanation capability have been particularly used in medical domains since a physician likes to understand the outcome of a classifier and map it to his domain knowledge; otherwise, the level of acceptance of an automatic system is low. Nowadays, data mining methods with explanation capability are also used for technical domains after more work on advantages and disadvantages of the methods has been done.

The most preferred method among the methods with explanation capability is decision tree induction method [1]. This method can easily learn a decision tree without heavy user interaction while in neural nets a lot of time is

spent on training the net. Cross-validation methods can be applied to decision tree induction methods; these methods ensure that the calculated error rate comes close to the true error rate. A large number of decision tree methods exist but the method that works well on average on all kinds of data sets is still the C4.5 decision tree method [2] and some of its variants. Although the user can easily apply this method to his data set thanks to all the different tools that are available and set up in such a way that none computer-science specialists, can use them without any problem, the user is still faced with the problem of how to interpret the result of a decision tree induction method. This problem especially arises when two different data sets for one problem are available or when the data set is collected in temporal sequence. Then the data set grows over time and the results might change.

The aim of this paper is to give an overview of the problems that arise when interpreting decision trees.

In Section 2, we explain the data collection problem. In Section 3, we review the explanation capability of a decision tree., before we present in Section 4 our method and the similarity measure for comparing two decision trees. Finally we give conclusions in Section 5.

### 2. The Problem

Many factors influence the result of the decision tree induction process. The data collection problem is a tricky pit fall. The data might become very noisy due to some subjective or system-dependent problems during the data collection process

Newcomers in data mining go into data mining step by step. First, they will acquire a small data set that allows them to test what can be achieved by data mining methods. Then, they will enlarge the data set hoping that a larger data set will result in better data mining results. But often this is not the case.

Others may have big data collections that have been collected in their daily practice such as in marketing and finance. To a certain point, they want to analyze these

## How to Compare and Interpret Two Learnt Decision Trees from the Same Domain?

Petra Perner, In: Leonard Barolli, Fatos Xhafa, Makoto Takizawa, Tomoya Enokido,

Hui-Huang Hsu (Eds.), IEEE 27th International Conference on Advanced Information Networking and Applications Workshops (WAINA), 25-28 March 2013, Barcelona, Catalonia, Spain, pp 318 - 322 E-ISBN 978-0-7695-4952-1, Print ISBN 978-1-4673-6239-9

data with data mining methods. If they do this based on all data they might be faced with a lot of noise in the data since customer behavior might have changed over time due to some external factors such as economic factors, climate condition changes in a certain area and so on.

Web data can change severely over time. People from different geographic areas and different nations can access a website and leave a distinct track dependent on the geographic area they are from and the nation they belong to.

If the user has to label the data, then it might be apparent that the subjective decision about the class the data set belongs to might result in some noise. Depending on the form of the day of the expert or on his experience level, he will label the data properly or not as well as he should. Oracle-based classification methods [3] or similarity-based methods [4] might help the user to overcome such subjective factors.

If the data have been collected over an extended period of time, there might be some concept drift. In case of a web-based shop the customers frequenting the shop might have changed because the products now attract other groups of people. In a medical application the data might change because the medical treatment protocol has been changed. This has to be taken into consideration when using the data.

It is also possible that the data are collected in time intervals. The data in time period  $_1$  might have other characteristics than the data collected in time period  $_2$ . In agriculture this might be true because the weather conditions have changed. If this is the case, the data cannot make up a single data set. The data must be kept separate with a tag indicating that they were collected under different weather conditions.

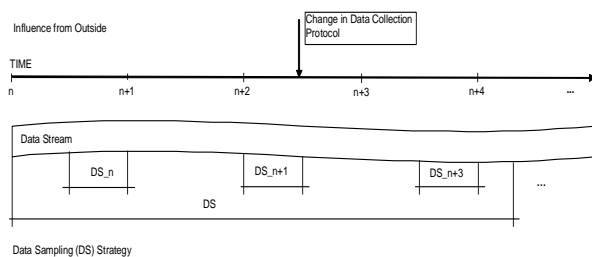


Fig. 1 The Data Collection Problem

## 3 Explanation Capability of the Decision Tree Model

Suppose we have a data set  $X$  with  $n$  samples. The outcome of the data mining process is a decision tree that represents the model in a hierarchical rule-based fashion. One path from the top to the leaf of a tree can be de-

scribed by a rule that combines the decisions of each node by a logical AND. The closer the decision is to the leaf, the more noise is contained in the decision since the entire data set is subsequently split into two parts from the top to the bottom and in the end only a few samples are contained in the two data sets. Pruning is performed to avoid that the model overfits the data. Pruning provides a more compact tree and often a better model in terms of accuracy.

The pruning algorithm is based on an assumption regarding the distribution of the data. If this assumption does not fit the data, the pruned model does not have better accuracy. Then it is better to stay with the unpruned tree.

When users feel confident about the data mining process, they are often keen on getting more data. Then they apply the updated data set that is combined of the data set  $X$  and the new data set  $X'$  containing  $n+t$  samples ( $n < n+t$ ) to the decision tree induction. If the resulting model only changes in nodes close to the leaf of a decision tree, the user understands why this is so.

There will be confusion when the whole structure of the decision tree has been changed especially, when the attribute in the root node changes. The root node decision should be the most confident decision. The reason for a change can be that there were always two competing attributes having slightly different values for the attribute selection criteria. Now, based on the data, the attribute ranked second in the former procedure is now ranked first. When this happens the whole structure of the tree will change since a different attribute in the first node will result in a different first split of the entire data set.

It is important that this situation is visually presented to the user so that he can judge what happened. Often the user has already some domain knowledge and prefers a certain attribute to be ranked first. A way to enable such a preference is to allow the user to actively pick the attribute for the node.

Another way to judge this situation is to look for the variance of the accuracy. If the variance is high, this means that the model is not stable yet. The data do not give enough confidence in regard to the decision.

The described situation can indicate that something is wrong with the data. It often helps to talk to the user and figure out how the new data set has been obtained. To give you an example: A data base contains information about the mortality rate of patients that have been treated for breast cancer. Information about the patients, such as age, size, weight, measurements taken during the treatment, and finally the success or failure, is reported. In the time period T1, treatment with a certain cocktail of medicine, radioactive treatment and physiotherapy has taken place; the kind of treatment is called a protocol. In the time period T2, the physicians changed the protocol since

## How to Compare and Interpret Two Learnt Decision Trees from the Same Domain?

Petra Perner, In: Leonard Barolli, Fatos Xhafa, Makoto Takizawa, Tomoya Enokido, Hui-Huang Hsu (Eds.), IEEE 27th International Conference on Advanced Information Networking and Applications Workshops (WAINA), 25-28 March 2013, Barcelona, Catalonia, Spain, pp 318 - 322 E-ISBN 978-0-7695-4952-1, Print ISBN 978-1-4673-6239-9

other medicine is available or other treatment procedures have been reported in the medical literature as being more successful. The physicians know about the change in protocol but they did not inform you accordingly. Then the whole tree might change and as a result the decision rules are changing and the physicians cannot confirm the resulting knowledge since it does not fit their knowledge about the disease as established in the meantime. The resulting tree has to be discussed with the physicians; the outcome may be that in the end the new protocol is simply not good.

Noisy data might be caused by wrong labels applied by the expert to the data. A review of the data with an expert is necessary to ensure that the data labels are correct. Therefore, the data are classified by the learnt model. All data sets that are misclassified are reviewed by the domain expert. If the expert is of the opinion that the data set needs another label, then the data set is relabeled. The tree is learnt again based on the newly labeled data set.

## 5 Comparison of two Decision Trees

Two data sets of the same domain that might be taken at different times, might result in two different decision trees. Then the question arises how similar these two decision trees are. If the models are not similar then something significant has changed in the data set.

A measure that allows us to judge the similarity between two trees would be useful.

The path from the top of a decision tree to the leaf is described by a rule like "IF attribute  $A \leq x$  and attribute  $B \leq y$  and attribute  $C \leq z$  and ... THEN Class\_1". The transformation of a decision tree in a rule-like representation can be easily done. The location of an attribute is fixed by the structure of the decision tree.

Comparison of rule sets is known from rule induction methods in different domains [5]. Here the induced rules are usually compared to the human-built rules. Often this is done manually and should give a measure about how good the constructed rule set is.

These kinds of rules can also be automatically compared by substructure mining. The following questions can be asked: a) How many rules are identical? b) How many of them are identical compared to all rules? c) What rules contain part structures of the decision tree?

We propose a first similarity measure for the differences of the two models as follows:

1. Transform two decision trees  $d_1$  and  $d_2$  into a rule set.
2. Order the rules of two decision trees according to the number  $n$  of attributes in a rule.
3. Then build substructures of all  $l$  rules by decomposing the rules into their Substructures.

4. Compare two rules  $i$  and  $j$  of two decision trees  $d_1$  and  $d_2$  for each of the  $n_j$  and  $n_i$  substructures with  $s$  attributes.
5. Build similarity measure  $SIM_{ij}$  according to formula 1-5.

The similarity measure is:

$$SIM_{ij} = \frac{1}{n} (Sim_1 + Sim_2 + \dots + Sim_k + \dots + Sim_n) \quad (1)$$

with  $n = \max\{n_i, n_j\}$  and

$$Sim_k = \begin{cases} 1 & \text{if substructure identity} \\ 0 & \text{if otherwise} \end{cases} \quad (2)$$

If the rule contains a numerical attribute  $A \leq k_1$  and  $A' \leq k_2 = k_1 + x$  then the similarity measure is

$$Sim_{num} = 1 - \frac{A - A'}{t} = 1 - \frac{k_1 - k_1 - |x|}{t} = 1 - \frac{|x|}{t} \quad (3)$$

for  $x < t$  and  $Sim_k = 0$  for  $x \geq t$

with  $t$  a user chosen value that allows  $x$  to be in a tolerance range of  $s\%$  (e.g. 10%) of  $k_1$ . That means as long as the cut-point  $k_1$  is within the tolerance range we consider the term as similar, outside the tolerance range it is dissimilar. Small changes around the first cut-point are allowed while a cut-point far from the first cut-point means that something seriously has happened with the data.

The similarity measure for the whole substructure is:

$$Sim_k = \frac{1}{s} \sum_{z=1}^s \begin{cases} Sim_{num} \\ 1 & \text{for } A = A' \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

The overall similarity between two decision trees  $d_1$  and  $d_2$  is

$$Sim_{d_1, d_2} = \frac{1}{l} \sum_{i=1}^l \max_{\forall j} Sim_{ij} \quad (5)$$

## How to Compare and Interpret Two Learnt Decision Trees from the Same Domain?

Petra Perner, In: Leonard Barolli, Fatos Xhafa, Makoto Takizawa, Tomoya Enokido,

Hui-Huang Hsu (Eds.), IEEE 27th International Conference on Advanced Information Networking and Applications Workshops (WAINA), 25-28 March 2013, Barcelona, Catalonia, Spain, pp 318 - 322 E-ISBN 978-0-7695-4952-1, Print ISBN 978-1-4673-6239-9

for comparing the rules  $i$  of decision  $d_1$  with rules  $j$  of decision  $d_2$ . Note that the similarity  $Sim_{d_1,d_2}$  must not be the same.

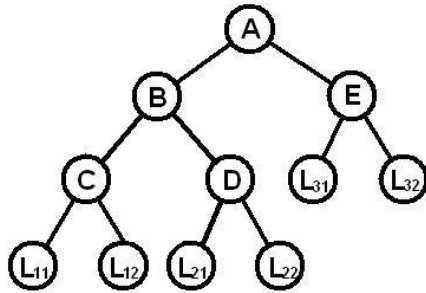


Fig. 2. Decision\_Tree\_1,  $Sim_{d1,d1}=1$

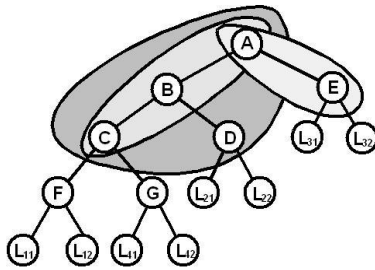


Fig. 3. Substructures of Decision Tree\_1 to Decision Tree\_2;  $Sim_{d1,d2}=0.9166$

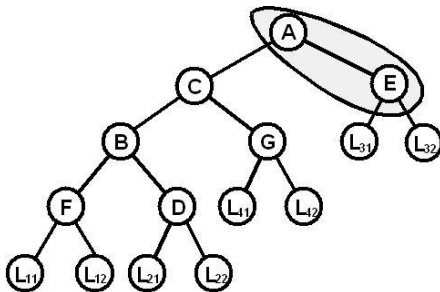


Fig. 4. Substructures of Decision Tree\_1 to Decision Tree\_3;  $Sim_{d1,d3}=0.375$ ;  $Sim_{d2,d3}=0.375$

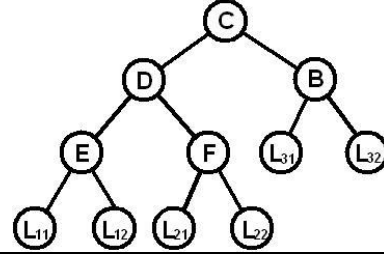


Fig. 5. Decision Tree\_4 dissimilar to all other Decision Trees,  $Sim_{d1,d4}=0$

The comparison of decision tree\_1 in Figure 2 with decision tree\_2 in Figure 3 gives a similarity value of 0.75 based on the above described measure. The upper structure of decision tree\_2 is similar to decision tree\_1 but decision tree\_2 has a few more lower leaves. The decision tree\_3 in Figure 4 is similar to decision tree\_1 and decision tree\_2 by a similarity value of 0.125. Decision tree\_3 in Figure 5 has no similarity at all compared to all other trees. The similarity value is zero.

Such a similarity measure can help an expert to understand the developed model and also help to compare two models that have been built based on two data set, wherein one contains  $N$  examples and the other one contains  $N+L$  samples.

There are other options for constructing the similarity measure. It is left for our further work.

## 6 Conclusions

The aim of this paper is to discuss how to deal with the result of data mining methods such as decision tree induction. This paper has been prompted by the fact that domain experts are able to use the tools for decision tree induction but have a hard time interpreting the results. A lot of factors have to be taken into consideration. The quantitative measures of the quality of a decision tree give a good overview in regard to the quality of the learnt model. But computer science experts claim that decision trees have explanation capabilities and that, compared to neural nets and SVM, the user can understand the decision. This is only partially true. Of course, the user can follow the path of a decision from the top to the leaves and this provides him with a rule where the decisions in the node are combined by logical ANDs. But often this is tricky. A user likes rules that fit his domain knowledge and make sense in some way. Often this is not the case

## How to Compare and Interpret Two Learnt Decision Trees from the Same Domain?

Petra Perner, In: Leonard Barolli, Fatos Xhafa, Makoto Takizawa, Tomoya Enokido, Hui-Huang Hsu (Eds.), IEEE 27th International Conference on Advanced Information Networking and Applications Workshops (WAINA), 25-28 March 2013, Barcelona, Catalonia, Spain, pp 318 - 322 E-ISBN 978-0-7695-4952-1, Print ISBN 978-1-4673-6239-9

since the most favored attributes of the user do not appear at a high position.

That makes the interpretation of a decision trees difficult. The user's domain knowledge, even if it is only limited, is an indicator whether he accepts the tree or not. The decision tree induction algorithm should allow the user to interact with the induction algorithm. If two attributes are ranked more or less the same the user should be able to choose which one of the attributes to pick.

The comparison of trees is an important issue that a user needs in order to understand what has changed. It can also indicate what has been gone wrong in the data. Therefore, proper similarity measures are needed that give a measure of goodness.

This paper described such a similarity measure and showed the usefulness for the interpretation of a decision tree.

## References

- [1] Perner, P.: Data Mining on Multimedia Data. LNCS, vol. 2558, Springer Verlag (2002)
- [2] Quinlan, JR: Decision trees and multivalued attributes. In: Hayes, J.E, Michie, D, Richards, J. (eds.) Machine Intelligence 11, Oxford University Press(1988),
- [3] Muggleton, S.: Duce - An Oracle-based Approach to Constructive Induction. Proceeding of the Tenth International Join Conference on Artificial Intelligence (IJCAI'87), pp. 287-292 (1987)
- [4] Perner, P.: Prototype-Based Classification. Applied Intelligence 28(3), 238-246 (2008)
- [5] Georg, G., Séroussi, B., Bouaud, J.: Does GEM-Encoding Clinical Practice Guidelines Improve the Quality of Knowledge Bases? A Study with the Rule-Based Formalism. AMIA Annu Symp Proc. 2003, pp. 254-258 (2003)