MEDINFO 98
B. Cesnik et al. (Eds)
Amsterdam: IOS Press

The Limitations of Decision Trees and Automatic Learning in Real World Medical Decision Making

Peter Kokol[a], Milan Zorman[a], Milojka Molan Štiglic[b], Ivan Malèiæ[c]

[a]Center of Medical Informatics, University of Maribor, FERI, Maribor, Slovenia
[b]Maribor Teaching Hospital, Department of Pediatric Surgery, Maribor, Slovenia
[c]Clinical Hospital Center - Rebro, Department of Pediatrics, Zagreb, Croatia

Abstract

The decision tree approach is one of the most common approaches in automatic learning and decision making. It is popular for its simplicity in constructing, efficient use in decision making and for simple representation, which is easily understood by humans.

The automatic learning of decision trees and their use usually show very good results in various "theoretical" environments. The training sets are usually large enough for learning algorithm to construct a hypothesis consistent with the underlying concept. But in real life it is often impossible to find the desired number of training objects for various reasons. The lack of possibilities to measure attribute values, high cost and complexity of such measurements, unavailability of all attributes at the same time are the typical representatives. There are different ways to deal with some of these problems, but in a delicate field of medical decision making, we cannot allow ourselves to make any inaccurate decisions.

We have measured the values of 24 attributes before and after the 82 operations of children in age between 2 and 10 years. The aim was to find the dependencies between attribute values and a child's predisposition to acidemia – the decrease of blood's pH. Our main interest was in discovering predisposition to two forms of acidosis, the metabolic acidosis and the respiratory acidosis, which can both have serious effects on child's health.

We decided to construct different decision trees from a set of training objects, which was complete (there were no missing attribute values), but on the other hand not large enough to avoid the effect of overfitting. A common approach to evaluation of a decision tree is the use of a test set. In our case we decided that instead of using a test set, we ask medical experts to take a closer look at the generated trees. They examined and evaluated the decision trees branch by branch. Their comments on the generated trees can be found in this paper.

The comments show, that trees generated from available training set mainly have surprisingly good branches, but on the other hand some are very "stupid" and no medical explanation could be found. Thereafter we can conclude, that the decision tree concept and automatic learning can be successfully used in real world situations, constrained with the real world limitations, but they should be used only with the guidelines of appropriate medical experts.

Keywords

Decision Trees; Automatic Learning; Decision Making; Acidosis; Medical Informatics

Introduction

By applying the decision tree learning algorithm to a set of data, we expect to gain more general knowledge for solving the problem. That knowledge can be used in future application for making new decisions [1, 2, 3, 4]. We usually take for granted that an appropriate set of training data is available to the automatic learning tool. Unfortunately, the most real world cases are not ideal. We have to cope with missing attribute values, noisy data and the lack of training objects. All these factors have a negative effect on the accuracy of knowledge, accumulated in the decision tree.

There are ways to reduce the negative effect, but they cannot exclude it completely [5, 6, 7].

Aim of the Paper

The aim of the paper is to present MtDeciT decision tool, which we developed for it's use in real world decision making. The main advantages of MtDeciT are:

- multi class decision making
- efficient pruning algorithm
- user friendly interface.

Materials and Methods

The Approach and Aims of the Research

In our case we measured 24 attributes of 82 children, that have undergone a medical surgery. Most attributes were measured before and then also after the surgery. All measurements were performed very carefully, in order to prevent mistakes. There were no missing attribute values.

So our only problem was the lack of training objects, in our case patients, that underwent a surgery.

Though we were facing a typical problem of overfitting, the decision was taken to go through the process of generating different decision trees for two cases of acidosis, called metabolic and respiratory acidosis. For the purpose of testing the efficiency of the decision trees, a test set is usually used. If the data from the training set is used instead of the test set, the tree will perform extraordinary well, but that would be a fake success. Due to the general lack of training objects, the decision has been taken to let experts examine and evaluate the decision trees. With that, another lack of using a test set has been avoided. It is almost impossible to pick such set of test data, that will enter every branch in the decision tree (except in using a training set). The medical experts, on the other hand, checked and evaluated every single branch in the tree. The analysis of generated trees by medical experts should reveal how "good" the accumulated knowledge is.

The latter was the main reason, why we picked decision trees out of all forms of supervised learning [8, 9, 10]. Unlike to some other approaches, the representations of the decision trees can easily be understood by humans. All tests in internal nodes of the tree can be determined, so different rules can be obtained from the tree.

Metabolic and Respiratory Acidosis

Through the help of classical medical research it has been established, that surgeries under the general anaesthesia cause in organism a tendency to dropping the blood's pH value, also known as predisposition to acidemia. We used the results of blood's gases analysis, serum electrolytes analysis, blood count and values of length of the operation, blood pressure, pulse, temperature, age, weight, height, sex, duration of the surgery and transfusion volume of 82 children as an input to a decision tree generator. With the model of a decision tree we wanted to establish the sequence of attributes, that have the largest influence on the sequence of physiological changes that lead into the state of either metabolic or respiratory acidosis. We would also be able to see, if and how compatible are the decision tree decisions with the medical explanations.

Decision Trees

A decision tree is induced on a training set, which consists of training objects (instances). Each training object is completely described by a set of attributes and a class label (category, outcome). Classes are mutually exclusive, what means that the training object can belong to only one class. Attributes can be continuous (numeric) or discrete. Further, we distinguish between ordered and unordered discrete attributes. Continuous attributes are not suitable for learning a tree, so they must be mapped into discrete space. A decision tree contains nodes and edges (links). There are two types of nodes. Each internal node (non-terminal node) has a split, which tests the value of the attributes, that have come into this node and with that splits the training set. Each internal node has at least two child nodes. External nodes, also called leaves or terminal nodes, are labelled with outcomes. Nodes (internal and external) are connected with edges. Edges are labelled with different outcomes of test, performed in the source node. Number of edges that come out of the node depends on the number of possible outcomes of the test.

The procedure of constructing a decision tree from a set of training objects is called induction of the tree. In induction of a decision tree, we start with an empty tree and an entire training set. At every step an 'unused' attribute is chosen with the help of a heuristic evaluation function and the training set is divided, according to the values of the chosen attribute (divide-and-conquer). This procedure is repeated over and over, until we end up with an empty training set and a decision tree. If the number of training objects is not large enough and if data involve some noise, it comes to overfitting. The algorithm tries to fit every singe attribute, including noisy and meaningless ones. This results in very large trees, which have many meaningless branches.

Heuristic Evaluation Function

As mentioned before, candidate attributes that may possibly describe the concept are outlined, using a heuristic evaluation function. Our modified heuristic function is also suitable for multi-class decision tree generation. Here we will describe, how this function operates.

$o_1, o_2, o_3, \ldots, o_m$ - number of training objects with outcomes 1 to m.

m - number of possible outcomes

$$I(o_1,o_2,\ldots,o_m)=-\frac{o_1}{\sum_{j=1} o_j}*\log_m\frac{o_1}{\sum_{j=1} o_j}-\frac{o_2}{\sum_{j=1} o_j}*\log_m\frac{o_2}{\sum_{j=1} o_j}-\ldots-\frac{o_m}{\sum_{j=1} o_j}*\log_m\frac{o_m}{\sum_{j=1} o_j} \tag{10}$$

Function I, as shown in Equation (1), measures the information content. The return value is in interval [0..1] for all cases, no matter how many outcomes are possible. Therefor the return value cannot be measured in bits (except in case, where m=2), but in some more abstract information content measure.

Let us presume, we have a set of attributes $A_1, A_2, \ldots, A_n$ with possible values (or intervals) for each A_i marked as $A_{i1}, A_{i2}, \ldots, A_{iv}$.

Then attribute A_i divides training set C into subsets $C_{i1}, C_{i2}, \ldots, C_{iv}$, where C_{i1} contains training objects with A_i attribute values equal to A_{i1}.

The heuristic function $E(C,A_i)$ for evaluation of attribute A_i is defined in Equation (2), where o_{ijk} is a number of training objects in subset C_{ij} with values A_{ik}.

$$E(C,A_i)=\sum_{j=1}^{v}\frac{\sum_{k=1}^{m} o_{ijk}}{\sum_{k=1}^{m} o_k}*I(o_{ij1},o_{ij2},\ldots,o_{ijm}) \tag{11}$$

A decision tree construction tool uses outlined attributes to formulate the appropriate decision tree, which identifies all class instances of the underlying concept, according to objects with known classification.

This decision tree characterisation becomes next a basis for:

- forecasting in which class an object previously unseen belongs; and
- the hierarchical representation of the most important attributes of the concept being investigated.

Pruning of the Decision Trees

One way of reducing the effect of overfitting is the use of pruning algorithms. The idea of pruning is to remove those branches from the decision tree, that are result of noise included in the training data. We distinguish between the two forms of pruning. Pre-pruning is executed as a part of tree building algorithm. It is also know as a stopping criteria. On the other side, post-pruning can only be performed on the decision tree, that has already been built. Our decision support tool MtDeciT enables the user to decide, which form of pruning to choose.

MtDeciT Decision Support Tool

MtDeciT is written in Visual C++ language and runs under the Windows environment.

Supported are all phases from preparing the training data, automatic learning, pruning, to decision making. Very friendly user interface enables even the novice users to exploit all tool's capabilities. It also enables settings of various building properties and with that influences the form of extracted knowledge.

Results and Discussion

The Generated Trees and their Explanations

For the purpose of determining a predisposition to acidemia, we generated numerous trees. Just to depict the results, we will present three generated trees with their medical explanations. Also described are other properties, specific to the generated tree. All decision trees are drawn hierarchically, so that every level is indented for a few spaces. Levels are also connected by lines. Described parts of the tree are marked with numbered parenthesis.

Metabolic Acidosis – attribute values measured after the surgery

This decision tree was generated from data, measured after the surgery. Two outcomes were possible, NO_PREDISPOSITION for 72 presumably healthy individuals and PREDISP_TO_METABOLIC_ACIDOSIS for 10 children, where a predisposition to metabolic acidosis was noticed. Comments are made on marked parts of the tree.

Comments:

1. An important factor that influences the occurrence of metabolic acidosis is low hematocrit.

Explanation: Anaemia can be the cause of low oxygenation of tissues and with that accelerated anaerobic glycolytic processes, which lead into state of metabolic acidosis [11].

2. The connection of age and relatively low body weight.

```
Decision : Predisposition_to_Metabolic_Acidosis(after)
Tree termination criteria : 0%
Heuristic : Logarithm
Ordinal Grouping of Attributes : Quartiles
Alternative Grouping : Disabled
```

1. parameter – Value of Parent Attribute (placed in [])
2. parameter – Attribute or Decision

```
[] B.E.
|
|
|___[-7.5000 .. -4.9750] NO_PREDISPOSITION
|
|___[-4.9750 .. -2.4500] NO_PREDISPOSITION
|
|___[-2.4500 .. 0.0750] NO_PREDISPOSITION
|
|
|___[0.0750 .. 2.6000] hematocrit                                 -
    |                                                              |
    |___[0.3000 .. 0.3425] sex                                     |
    |   |                                                          |
    |   |___[M] NO_PREDISPOSITION                                  |
    |   |                                                          |
    |   |___[Z] PREDISPOSITION_TO_METABOLIC_ACIDOSIS               |
    |                                                              |
    |___[0.3425 .. 0.3850] age(years)                          -   |
    |   |                                                      |   |
    |   |___[2 .. 4] NO_PREDISPOSITION                         |   |
    |   |                                                      |   |
    |   |___[4 .. 6] weight(kg)                                |> 2|
    |   |   |                                                  |   |
    |   |   |___[10 .. 20.5000] NO_PREDISPOSITION              |   |
    |   |   |                                                  |   |
    |   |   |___[20.5000 .. 31] PREDISP_TO_METABOLIC_ACIDOSIS-     |
    |   |                                                          |> 1
    |   |___[6 .. 8] NO_PREDISPOSITION                             |
    |   |                                                          |
    |   |___[8 .. 10] NO_PREDISPOSITION                            |
    |                                                              |
    |___[0.3850 .. 0.4275] duration(min)                   -       |
    |   |                                                  |       |
    |   |___[15 .. 58.7500] PREDISP_TO_METABOLIC_ACIDOSIS  |       |
    |   |                                                  |       |
    |   |___[58.7500 .. 102.5000] NO_PREDISPOSITION        |       |
    |   |                                                  |       |
    |   |___[102.5000 .. 146.2500] age(years)              |       |
    |   |   |                                              |> 3    |
    |   |   |___[6 .. 8] NO_PREDISPOSITION                 |       |
    |   |   |                                              |       |
    |   |   |___[8 .. 10] PREDISP_TO_METABOLIC_ACIDOSIS    |       |
    |   |                                                  |       |
    |   |___[146.2500 .. 190] NO_PREDISPOSITION            |       |
    |                                                      |       |
    |___[0.4275 .. 0.4700] NO_PREDISPOSITION               -       -
```

Figure 1 - Decision tree for metabolic acidosis - values measured before surgery

Explanation: Preoperative starving and the act of surgery itself cause a complex state of stress in child's organism, which also leads into state of metabolic acidosis [12].

3. The duration of the surgery.

Explanation: Predisposition to the metabolic acidosis can be noticed after relatively short surgeries and also after longer surgeries. Due to the lack of training objects, this attribute wasn't presented as an important one, even though medical experts pay a great attention to it. Different outcomes spread over the time interval prove the prediction, that causes for the state of metabolic acidosis are of a very complex nature (for instance bleeding, drop in blood pressure,) [13].

Metabolic Acidosis – the difference between attribute values measured before and after the surgery (in %)

This decision tree was generated from the difference (in percent) between data, measured before and after the surgery. Two outcomes were possible, NO_PREDISPOSITION for 72 presumably healthy individuals and PRED_TO_METAB_ACIDOSIS for 10 children, where a predisposition to metabolic acidosis was noticed.

```
Decision : Predisposition_to_Metabolic_Acidosis(difference)
Tree termination criteria : 0%
Heuristic : Logarithm
Ordinal Grouping of Attributes : Quartiles
Alternative Grouping : Disabled

1. parameter - Value of Parent Attribute (placed in [])
2. parameter - Attribute or Decision

[] B.E.
|
|__[-2800..-1275] age(years)
|  |
|  |__[2..4] NO_PREDISPOSITION
|  |
|  |__[4..6] PREDISP_TO_METABOLIC_ACIDOSIS
|
|__[-1275..250] hematocrit
|  |
|  |__[-21.4100..-11.4325] age(years)
|  |  |
|  |  |__[2..4] PRED_TO_METAB_ACIDOSIS
|  |  |
|  |  |__[4..6] NO_PREDISPOSITION
|  |  |
|  |  |__[6..8] NO_PREDISPOSITION                  |> 1
|  |  |
|  |  |__[8..10] weight(kg)
|  |     |
|  |     |__[20.5000..31] PRED_TO_METAB_ACIDOSIS
|  |     |
|  |     |__[41.5000..52] NO_PREDISPOSITION
|  |
|  |__[-11.4325..-1.4550] erythrocytes
|  |  |
|  |  |__[-19.4125..-8.6750] NO_PREDISPOSITION
|  |  |
|  |  |__[-8.6750..2.0625] HCO3
|  |  |  |
|  |  |  |__[-44.0900..-26.9275] NO_PREDISPOSITION
|  |  |  |
|  |  |  |__[-26.9275..-9.7650] NO_PREDISPOSITION
|  |  |  |
|  |  |  |__[-9.7650..7.3975] SCl
|  |  |     |
|  |  |     |__[-8.7000..-3.8650] NO_PREDISPOSITION
|  |  |     |
|  |  |     |__[-3.8650..0.9700] hemoglobin
|  |  |     |  |
|  |  |     |  |__[-11.9800..-4.5600] NO_PREDISPOSITION
|  |  |     |  |
|  |  |     |  |__[-4.5600..2.8600] RR2
|  |  |     |     |                                        |> 3
|  |  |     |     |__[-10.1675..2.7450] PRED_TO_METAB_ACIDOSIS
|  |  |     |     |
|  |  |     |     |__[2.7450..15.6575] PRED_TO_METAB_ACIDOSIS     |>2
|  |  |     |     |
|  |  |     |     |__[15.6575..28.5700] NO_PREDISPOSITION
|  |  |     |
|  |  |     |__[0.9700..5.8050] thrombocytes
|  |  |     |  |
|  |  |     |  |__[-32.6400..11.0825] NO_PREDISPOSITION
|  |  |     |  |                                           |> 4
|  |  |     |  |__[11.0825..54.8050] PRED_TO_METAB_ACIDOSIS
|  |  |     |
|  |  |     |__[5.8050..10.6400] PRED_TO_METAb_ACIDOSIS
|  |  |
|  |  |__[2.0625..12.8000] NO_PREDISPOSITION
|  |
|  |__[-1.4550..8.5225] pulse
|  |  |
|  |  |__[-3.4700..-2.3350] RR2
|  |  |  |
|  |  |  |__[-23.0800..-10.1675] NO_PREDISPOSITION
|  |  |  |
|  |  |  |__[-10.1675..2.7450] PRED_TO_METAB_ACIDOSIS
|  |  |  |
|  |  |  |__[2.7450..15.6575] NO_PREDISPOSITION
|  |  |
|  |  |__[-0.0650..1.0700] NO_PREDISPOSITION
|  |
|  |__[8.5225..18.5000] NO_PREDISPOSITIN
|
|__[250..1775] NO_PREDISPOSITION
|
|__[1775..3300] NO_PREDISPOSITION
```

Figure 2 - :Decision tree for metabolic acidosis - the difference between attribute values (in %)

Comments

1. The connection of age and relatively low body weight.

Explanation: Preoperative starving and the act of surgery itself cause a complex state of stress in child's organism, which also leads into state of metabolic acidosis.

2. Very high ranked factors in this decision tree are lowered level of hematocrit and lowered number of erythrocites [14].

Explanation: Low oxygenation of tissues. The anaerobic glycolytic processes cause the lowered bicarbonate level and changes in the level of serum chlorides.

3. The values of hemoglobin and blood pressure.

Explanation: In this sub-branch occur lowered value of hemoglobin and compensatory growth of RR2 and pulse (as expected).

4. The influence of thrombocytes.

Explanation: There is no logical medicine explanation for the influence of this factor.

Conclusion

In the paper we presented the use of decision tree approach in the real world medical decision making. Our aim was to test the ability of decision tree approach to cope with the insufficient size of training set. The results show that use of decision trees pays off even in situations, where we don't expect the success of the approach. The expected influence of overfitting was surprisingly weaker as we thought it would be. The generated trees can be of a great assistance to a user with proper background knowledge. This success encourages the use of decision trees in other areas, where traditional convictions don't support the use of automatic learning approaches. Recently, the decision trees are being used in combination with other learning strategies in order to provide us with even better performance [15].

References

[1] Kokol P, Završnik J, Zorman M, Malèiæ I, Kancler K. Participative design, decision trees, automatic learning and medical decision making. In: Brender J, ed. *Studies in health technology and informatics*. Vol.18. Amsterdam [etc.]: IOS Press; Tokyo: Ohmsha, pp. 501-505, 1996.

[2] Russel SJ, Norvig P, et al., *Artificial intelligence: a modern approach*, Englewood cliffs, Prentice-Hall, 1995.

[3] Quinlan JR, Induction of decision trees, *Machine learning*, no.1, pp. 81-106, 1986.

[4] Selby RW, Porter AA, Learning from examples: generation and evaluation of decision trees for software resource analysis, *IEEE Transactions on software engineering*, vol.14, no. 12, pp. 1743-1757, December 1988.

[5] Quinlan JR, Simplifying decision trees, *International journal of man-machine studies*, no.27, pp. 221-234, 1987.

[6] Quinlan JR, *C4.5:Programs for machine learning*, Morgan Kaufmann publishers, San Mateo, CA, 1993.

[7] Quinlan JR, Decision Trees and Instance-Based Classifiers. In: Allen B Tucker jr, ed. *CRC Computer science and engineering handbook*. CRC Press, ch. 22, pp. 521-535, 1997.

[8] Kokol P, Stiglic B, _umer V: Metaparadigm: a soft and situation oriented MIS design approach. *International Journal of Bio-Medical Computing*. vol. 39. pp. 243- 256. 1995.

[9] Baeck T: *Evolutionary Algorithms in Theory and Practice*. Oxford University Press, Inc., 1996.

[10] Holland JH: *Hidden Order, How Adaption Builds Complexity*. Addison-Wesley Publishing Company, 1995.

[11] Rose BD: *Clinical Physiology of Acid-Base and Electrolyte Disorderes*. New York: Mc Graw-Hill, 1989.