The Art of Building Decision Trees

Špela Hleb Babič,[1] Peter Kokol,[1] Vili Podgorelec,[1] Milan Zorman,[1] Matej Šprogar,[1] and Milojka Molan Štiglic[2]

Decision support systems that help physicians are becoming a very important part of medical decision making. They are based on different models and the best of them are providing an explanation together with an accurate, reliable, and quick response. One of the most viable among models are decision trees, already successfully used for many medical decision-making purposes. Although effective and reliable, the traditional decision tree construction approach still contains several deficiencies. Therefore we decided to develop and compare several decision support models using four different approaches. We took statistical analysis, a MtDeciT, in our laboratory developed tool for building decision trees with a classical method, the well-known C5.0 tool and a self-adapting evolutionary decision support model that uses evolutionary principles for the induction of decision trees. Several solutions were evolved for the classification of metabolic and respiratory acidosis (MRA). A comparison between developed models and obtained results has shown that our approach can be considered as a good choice for different kinds of real-world medical decision making.

Art (from Latin ***ars*** *meaning* ***skill****) is the skill in doing or performing that is attained by study, practice, or observation*

Microsoft Bookshelf. 1999 Edition

KEY WORDS: decision support systems; art; decision trees.

INTRODUCTION

As in many other areas, decisions also play an important role in medicine, especially in medical diagnostic processes. Decision support systems (DSS) helping physicians are becoming a very important part in medical decision making, particularly in those areas where decisions must be made effectively and reliably.(1,2) Since conceptual simple decision making models with the possibility of automatic learning should be considered for performing such tasks,(3) according to recent reviews(1,4,5)

[1] Faculty of Electrical Engineering and Computer Science, University of Maribor, Smetanova 17, SI 2000 Maribor, Slovenia, Email: {Spela.Hleb, Kokol}@uni-mb.si.

[2] Maribor Teaching Hospital, Department of Pediatric Surgery, Ljubljanska 2, SI 2000 Maribor, Slovenia.

decision trees are a very suitable candidate. Due to many various methods for decision tree construction proposed during the last few years, the building of decision trees can be regarded as a kind of art, especially in the real world situations where the number of cases is limited and the data gathered are noisy. In such a situation the selection of an appropriate construction method is crucial for the accuracy and efficiency of the decision tree being generated and the decision process resulting from the generated tree.

Decision trees have been already successfully used in DSS for many decision-making purposes, but some problems still persist in the traditional way of induction, which has not changed much since introduction. Considering the advantages of the evolutionary methods in performing complex tasks, we decided to overcome these problems by inducing decision trees also with genetic algorithms. By comparing the results and the effectiveness of decision trees constructed with classical and evolutionary approaches we want to show and discuss some interesting findings.

THE APPROACH AND AIMS OF THE RESEARCH

In our case we measured 24 attributes of 90 children that have undergone medical surgery. Most attributes were measured before and after surgery. All measurements were performed very carefully, in order to prevent mistakes; there were no missing attribute values.

At first, we applied a classical statistical method, ANOVA, and discriminant analysis both using SPSS software package. Then we built several decision trees, using the same data. We used following approaches: MtDeciT, a laboratory developed tool for building decision trees with classical method, the well-known C5.0, tool and a self-adapting evolutionary decision support model, that uses evolutionary principles for the induction of decision trees.

By comparing all the results, we would be able to see first, if and how compatible the decision trees are and the results of statistic analysis, and if decision trees really can replace some statistical methods in the process of searching for the most significant attributes. Second, how different methods influence the accuracy and preductiveness of decision trees.

In this paper the used methods will be introduced first. The classical approaches will be introduced briefly while the evolutionary approach to the induction of decision trees with emphasis on the self-adaptation mechanism will be explained more in detail. Next the MRA problem will be introduced and their the results obtained with the described models for the MRA prediction will be presented. A discussion about important findings, problems, and analysis of the results will follow after, and we will conclude with a short summary of achievements and several directives for further research.

MAIN CONTRIBUTIONS OF THE PAPER

There are three essential contributions in this paper. The first is a precisely made comparison and assessment of the effectiveness of different approaches for building decision trees that was obtained by using constructed decision models with the same set of real-world data. We showed that discretization of numerical attributes deserve special attention and we made some suggestions for dividing the numerical attributes' domain into subsets. The second is our new approach to the induction of decision trees with the use of genetic algorithms. We tried to unite the effectiveness and robustness of evolutionary methods with the simplicity and popularity of decision trees, and in this manner improve the quality of the obtained solutions. The third is the introduction of the self-adapting mechanism in the evolutionary process which adapts the evaluation function. In this way the effort and time spent on the definition of a proper evaluation function can be minimized, and the quality of the solutions is increased.

METHODS

Decision Trees

Inductive inference is the process of moving from concrete examples to general models, where the goal is to learn how to classify objects by analyzing a set of instances (already solved cases) whose classes are known. Instances are typically represented attribute-value vectors. Learning input consists of a set of such vectors, each belonging to a known class, and the output consists of a mapping from attribute values to classes. This mapping should accurately classify both given instances and other unseen instances.

A decision tree[(7–9)] is a formalism for expressing such mappings and consists of tests or attribute nodes linked to two or more sub-trees and leafs or decision nodes labeled with a class which means the decision. A test node computes some outcome based on the attribute values of an instance, where each possible outcome is associated with one of the sub-trees. An instance is classified by starting at the root node of the tree. If this node is a test, the outcome for the instance is determined and the process continues using the appropriate sub-tree. When a leaf is eventually encountered, its label gives the predicted class of the instance.

Building Decision Trees Using Traditional Method: C5.0 and MtDeciT

MtDeciT, a tool for building decision trees, was developed in our laboratory. The techniques it uses are: discretization of numerical attributes can be done by hand (for attributes like body temperature, where we know how to divide interval to subintervals), or divided to quartiles or octiles. The heuristic based on information theory in combination with attribute priorities. Pre-pruning and error reduction based post-pruning. C5.0 is the latest evolution of C4.5 that includes discretization of numerical attributes using information theory based functions, boosting, pre- and post-pruning and some other state-of-the-art options for building decision trees.

Construction of Decision Trees Using Evolutionary Method

While building decision trees using evolutionary methods is our original contribution, let us explain more precisely as other widely used approaches.

Genetic algorithms are generally used for very complex optimization tasks,[12] for which no efficient heuristic method is developed. Construction of decision trees is a complex task, but an exact heuristic method exists that usually works efficiently and reliably.[8,9] Nevertheless, there are some reasons justifying our evolutionary approach. Genetic algorithms provide a very general concept, that can be used in all kinds of decision-making problems. Because of their robustness they can also be used on incomplete, noisy data (which often happens in medicine because of measurement errors, unavailability of proper instruments, risk to the patient, etc.) and which are not very successfully solvable by traditional techniques of decision tree construction. Furthermore, genetic algorithms use evolutionary principles to evolve solutions, therefore solutions can be found that can be easily overlooked. Another important advantage of genetic algorithms is the possibility of optimizing the decision tree's topology and the adaptation of class intervals for numeric attributes, simultaneously with the evolution process. One further advantage of the evolutionary approach should not be overlooked: not only one, but several equally qualitative solutions are obtained for the same problem (in most cases). In this way an expert can decide which of the given solutions will be used. And last but not least, by weighting different parameters, searching can be directed to the situation that best applies to current needs, particularly in multi-class decision-making processes, where we have to decide for which decision the reliability should be maximized.

Evolutionary Process

When defining the internal representation of individuals within the population, together with the appropriate genetic operators that will work upon the population, it is important to assure the feasibility of all solutions during the whole evolution process. Therefore we decided to present individuals directly as decision trees. This approach has some important features: all intermediate solutions are feasible, no information is lost because of conversion between internal representation and the decision tree, the evaluation function can be straightforward, etc. The problem with direct coding of solution may bring some problems in defining of genetic operators. As decision trees may be seen as a kind of simple computer program (with attribute nodes being conditional clauses and decision nodes being assignments) we decided to define genetic operators similar to those used in genetic programing where individuals are computer program trees.[14]

For the selection, a slightly modified linear ranking selection was used and the evaluation function was presented as a weighted sum of wrong classifications for learning objects.

Crossover works on two selected individuals as an exchange of two randomly selected sub-trees. Mutation consists of three parts: the first part randomly replaces one attribute node with another, randomly chosen from the whole list; the second part randomly replaces a selected decision with another; and the third part constructs a new sub-tree and connects it to a selected node in the tree.

As the evolution repeats, more qualitative solutions are obtained regarding the chosen evaluation function. The evolution stops when an optimal or at least an acceptable solution is found or if the fitness score of the best individual does not change for a predefined number of generations.

Self-adaptation by Information Spreading in Multi-population Model

The quality and reliability of the obtained results depend a great deal on the selected evaluation function, therefore it is reasonable to expect that automatic adaptation of an evaluation function would result in better results. One basic question has led us to the development of a multi-population model for the construction of decision trees: how to provide a method for automatic adaptation of the evaluation function that would still assure the quality of evolved solutions? Namely, if the modification of the evaluation function is unsupervised, the evaluation function can easily become inappropriate which of course gives bad solutions. Our idea was that the evaluation function should evolve together with the decision trees, where solutions are regulated in that they are being compared with another set of decision trees, where the evaluation function is predefined and tested. This brings us to our multi-population model (Fig. 1).

The model consists of three independent populations: one main population and two competing ones. First population is a single decision tree, initially induced in traditional way with C4.5 algorithm,[9] and the decision trees in other two populations evolve in accordance with the evolutionary algorithm. The best individual from the main population, which serves as the general solution is competing with the best individuals from other populations in solving a given problem upon a learning set. When it becomes dominant over the others (in the sense of the accuracy of the classification of learning objects) it spreads its "knowledge" to them. In this way the other solutions become equally successful and the main population has to improve further to beat them. One global evaluation function (different from those used in the evolutionary process to evaluate the fitness of individuals) is used to

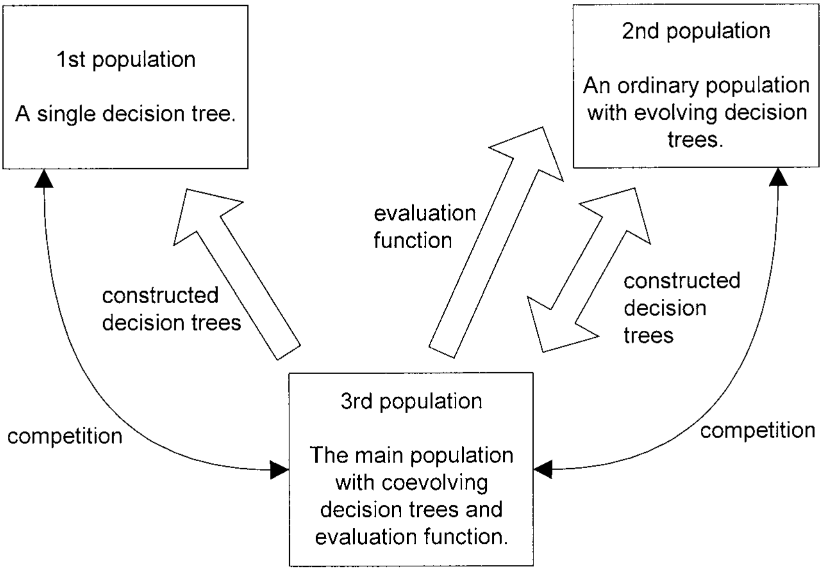
1st population
A single decision tree.
2nd population
An ordinary population with evolving decision trees.
evaluation function
constructed decision trees
constructed decision trees
competition
competition
3rd population
The main population with coevolving decision trees and evaluation function.

Fig. 1. The multi-population decision support model with information spreading.

determine the dominance of one population over another. It is predetermined as a weighted sum of accuracy, sensitivity and specificity (look at the results in the next section) of the given solution.

One of the recent and very promising aspect of genetic algorithms that is still unexploited is using the natural phenomenon of co-evolution. In nature it often happens that two or more species have to evolve with respect to each other to become more successful fighting for resources in the environment as a whole. We used the principle of co-evolution to solve the simultaneous evolution of decision trees and the adaptation of evaluation function in the main population of our multipopulation model. By competing with other populations the adaptation of evolving evaluation function is supervised, and in this way the problem of automatic adaptation is solved.

APPLICATION AND RESULTS

Metabolic and Respiratory Acidosis

Through the help of classical medical research it has been established, that surgeries under the general anesthesia 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.

Predicting Power of Various Decision Trees Methods

All 82 patients were divided into a learning and a testing set, 70 patients were selected for the learning and the rest, 12, for the testing set. The instances were described by 24 attributes and they were classified into two classes; NO—no predisposition to metabolic acidosis and YES—a predisposition to metabolic acidosis.

First we constructed two decision trees traditionally with the help of MtDeciT tool. For ordinal grouping of attributes we used quartiles for the first (MtDeciT(4)) and octiles for the second decision tree (MtDeciT(8)). Then we built a decision tree using the C5.0 tool (C5.0) and finally a few decision trees were evolved through our new evolutionary approach (SAEDT). All decision models were built upon the same learning and testing sets. A comparison of results obtained from all constructed decision trees has been made (Table I). We got the same results from MtDeciT(8), C5.0, and SAEDT methods (correctly classified all negative instances and missclassified 1 positive instance) and MtDeciT(4) and discriminant analysis were a little bit worse.

Table I. The Results of Metabolic Acidosis Classification by Various Methods

	MtDeciT (4)		MtDecit (8)		C5.0		SAEDT		Discr. analysis	
	NO[a]	YES[a]	NO	YES	NO	YES	NO	YES	NO	YES
NO	9	1	10	0	10	0	10	0	9	1
YES	1	1	1	1	1	1	1	1	1	1

[a] NO—Classified as no predisposition to MRA, YES—classified as predisposition to MRA.

Determination Importance of Attributes

From comparison of importance of attributes yield from used methods we can observe that similar attributes were chosen as the most important (the root of a decision tree), but the less important ones (deeper in tree, near to decision attributes) were quite different.

Table II shows the important attributes for each method.

Since the first four models differ in the way of discretization of numerical attributes it obviously plays a great role in defining hierarchy of decision trees and therefore also in making decisions as correct as possible.

Effectiveness of Used Methods

A comparison of the effectiveness of different diagnostic methods in the field of machine learning is usually described by accuracy and in the field of medicine

Table II. The Comparison of Importancy of Attributes Obtained by Use of Different Approaches

	MtDecit (4)	MtDecit (8)	C5.0	SAEDT	ANOVA	Discr. analysis
Leucocytes			X (level 3)			
Erythrocytes						
Hemoglobin						X (4)
Hematocrit	X (level 2)	X (level 2)			X	X (5)
Thrombocytes					X	
PH			X (level 2)		X	
B.E.	X (level 1)	X (level 1)	X (level 1)	X (level 1)	X	X (1)
HCO3				X (level 3)	X	X (2)
PCO2						X (3)
PO2						
HbO2						
Na		X (level 2)		X (level 2)	X	
K				X (level 3)		
Cl					X	
Temperature						
Pulse						
RR1					X	
RR2						
Sex	X (level 3)	X (level 3)		X (level 2)		
Age	X (level 3)					

Table III. A Comparison of Effectiveness for Different Methods

	MtDeciT (4)	MtDecit (8)	C5.0	SAEDT	Discr. analysis
Accuracy	83,33%	91,67%	91,67%	91,67%	83,33%
Sensitivity	50,00%	50,00%	50,00%	50,00%	50,00%
Specificity	90,00%	100%	100%	100%	90,00%

by sensitivity and specificity (Table III). The accuracy of a diagnostic method is calculated as the relation between the correctly classified and all testing objects. Sensitivity is based on ratio of correctly classified positive patients (patients having predisposition to MRA in our case) compared to all positive patients, and specificity is based on ratio of correctly classified negative patients (without predisposition to MRA) compared to all negative patients. All values are given as percentages. From Table III we can see that we got the best results from MtDeciT(8), C5.0 and SAEDT methods and MtDeciT(4) and discriminant analysis were a little bit worse.

The next important characteristic, maybe even more important than the accuracy of suggested decisions, is the topology and size of the constructed decision trees (Table IV).

The results present in Table IV support our hypothesis about the optimization of tree topology and complexity and is very important because it enables a physician to predict the presence of MRA with fewer examinations than before—a result that really made the medical experts enthusiastic.

DISCUSSION

Several interesting discoveries have been made while comparing four different decision trees and statistical analysis results, some of them expected and others surprising and worth thinking over and doing further research. Note the small size of training and testing sets; it may influence the quality of the obtained solutions. Since the real problem of MRA is rare and it is not easy to collect data, we have done our best with the available set of instances.

The most significant observation described in the work is definitely the influence of discretization of numerical attributes. The results were better while building the decision tree with octiles than with quartiles. The C5.0 method uses division of intervals with determination of threshold (only two subintervals) while the SAEDT uses quartiles as well. The last method improves the accuracy of decisions with a

Table IV. A Comparison of the Complexity between Traditionally and Self-Adapting Evolutionary Constructed Decision Trees

	MtDeciT (4)	MtDeciT (8)	C5.0	SAEDT
Num. of attribute nodes	5	4	4	7
Num. of decision nodes	7	4	5	16
Maximum depth	4	3	4	4
Average depth	3,08	2,11	2,8	2,81

self-adapting algorithm that yields in better effectiveness as MtDeciT(4) which use quartiles too.

The results proved that building decision trees really is an art. We can construct several different decision trees upon the same set of data, where different attributes have different importance and contributions to the final decision. We can observe that especially the evolutionary approach results in various equivalent decision trees with different attributes or a different attribute hierarchy. That is especially convenient for real medical cases where a particular patient has already undergone some investigation and the most relevant decision tree can be chosen. The same opinion is shared by some medical experts who have tried to predict the MRA in different ways.

The analysis of obtained results show us possible improvements of the decision-making processes mostly in more sophisticated discretization of numerical attributes. Our further research will focus on, first; human predicted division of attribute domain to proper intervals that will result in so called fuzzy decision trees; and second, evolutionary developed discretization, where the proper intervals would be obtained by a step-to-step evolution process involving all typical genetic processes and a continual evaluation of the results.

CONCLUSIONS

In the paper we present the comparison study of effectiveness of decision trees based DSS. The four different models—decision trees were constructed and the results were compared to each other and to results obtained from statistical analysis. We applied it to the prediction of MRA in children, but it can be easily used for other kinds of medical decision-making purposes.

All methods are described; classical briefly; and our original approach for building decision trees using evolutionary algorithms more in detail. The whole evolution process is described, with emphasis on the self-adaptation by information spreading. Results of MRA classification by classical and evolutionary constructed decision trees are compared with those obtained by statistical analysis.

There are three essential contributions to this paper. The first one is the precisely made comparison and assessment of the effectiveness of different approaches for building decision trees that was obtained by using the same set of real-world data for all constructed decision models. We showed that discretization of numerical attributes deserve special attention and we made some suggestions for dividing the numerical attributes' domain into subsets. The second is our new approach to the induction of decision trees with the use of genetic algorithms. The third one is the introduction of the self-adapting mechanism in the evolutionary process which adapts the evaluation function. In this way the effort and time spent on the definition of a proper evaluation function can be minimized, and the quality of the solutions is increased.

Since the present paper proved that division of numerical attributes' domain is very important and has great influence to quality in decision making, we expect great improvement from our further research and a step forward in general decision making processes.

REFERENCES

1. Kokol, P. *et al.,* Decision trees and automatic learning and their use in cardiology, *J. Med. Systems* 19(4): 1994.
2. Kokol, P., Podgorelec, V., and Malcic, I., Diagnostic process optimization with evolutionary programming, proceedings of the 11th IEEE Symposium on Computer-based Medical Systems CBMS'98, pp. 62–67, Lubbock, Texas, June 1998.
3. Kokol, P., Stiglic, B., and Zumer, V., Metaparadigm: a soft and situation oriented MIS design approach, *Int. J. Bio-Med. Comput.* 39:243–256, 1995.
4. Kokol, P. *et al., Spreadsheet Software and Decision Making in Nursing, Informatics '91* (Hovenga E.J.S. *et al.,* eds.), Springer Verlag, 1991.
5. Quinlan, J.R., Decision trees and decision making, *IEEE Trans System, Man and Cybernetics* 20(2):339–346, 1990.
6. Podgorelec, V., and Kokol, P. Evolutionary construction of medical decision trees. Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society EMBS'98, Hong Kong, 1998.
7. Quinlan, J.R., Induction of decision trees. *Machine Learning.* No.(1):81–106, 1986.
8. Quinlan, J.R., Simplifying decision trees. *Int. J. Man-Machine Studies* (27):221–234, 1987.
9. Quinlan, J.R., *C4.5: Programs for Machine Learning,* Morgan Kaufmann, 1993.
10. Bäck, T., *Evolutionary Algorithms in Theory and Practice,* Oxford University Press, Inc., 1996.
11. Forrest, S., *Genetic Algorithms, ACM Computing Surveys,* pp. 77–80, Vol. 28, No. 1, March 1996.
12. Goldberg, D.E., *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley, Reading MA, 1989.
13. Holland, J.H., *Adaptation in natural and artificial systems,* MIT Press, Cambridge MA, 1975.
14. Koza, J.R., *Genetic Programming: On the Programming of Computers by Natural Selection,* MIT Press, 1992.
15. Podgorelec, V., and Kokol, P., Genetic algorithm based system for patient scheduling in highly constrained situations. *J. Med. Systems.* 21:417–427, 1997.