TASK QUARTERLY 8 No 2 (2004), 243–247

A NEW ALGORITHM FOR GENERATION OF DECISION TREES

JERZY W. GRZYMAŁA-BUSSE[1,2], ZDZISŁAW S. HIPPE[2], MAKSYMILIAN KNAP[2] AND TERESA MROCZEK[2]

[1] *Department of Electrical Engineering and Computer Science, Kansas University, Lawrence (KS) USA*
jerzy@eecs.ku.edu

[2] *Department of Expert Systems and Artificial Intelligence, University of Information Technology and Management, Sucharskiego 2, 35-225 Rzeszow, Poland*
{*zhippe, mknap, tmroczek*}*@wenus.wsiz.rzeszow.pl*

(Received 12 September 2003; revised manuscript received 30 January 2004)

Abstract: A new algorithm for development of quasi-optimal decision trees, based on the Bayes theorem, has been created and tested. The algorithm generates a decision tree on the basis of Bayesian belief networks, created prior to the formation of the decision tree. The efficiency of this new algorithm was compared with three other known algorithms used to develop decision trees. The data set used for the experiments was a set of cases of skin lesions, histopatolgically verified.

Keywords: artificial intelligence, supervised machine learning, decision trees, Bayes networks

1. Introduction

The main goal of our research was developing a computer-assisted methodology of early and noninvasive diagnosis of one of the most dangerous human diseases, skin cancer [1]. Research described in this paper involved unsupervised machine learning in classification and identification of melanocytic skin lesions. To discover the knowledge hidden in medical datasets a computer program suite was created. Such datasets are frequently uncertain, *e.g.* conflicting. Until now, four computer programs have been created: AffinitySEEKER® (using a minimal-distance method to find similarity between the investigated objects, in our case, melanocytic skin lesions) [2], BeliefSEEKER® (generating stochastic belief networks) [3], TreeSEEKER® (generating quasi-optimal decision trees) [4], and PlaneSEEKER® (using optimized algorithms of linear machine learning to identify multicategory objects using a binary recurrent classification engine) [5]. Main features of these information systems, developed at the Kansas University in Lawrence, KS USA are (i) a uniform format of input data, compatible with the format used by the LERS system (which generates learning models in the form of sets of rules) [6], and (ii) the ability to generate twofold learning models:

a certain model (for data sets without conflicting cases) and a possible model (in case of data sets with conflicting cases). In this paper we concentrate on algorithms to generate decision trees. In our research [7] we have concluded that this type of learning model provides the most promising results in diagnosis of melanocytic skin lesions. It was necessary to equip the TreeSEEKER® system with additional, new algorithms for creating decision trees and checking their effectiveness in comparison with the earlier algorithms implemented by Czerwiński [8] and Quinlan [9]. Our experience in generating belief networks [10] suggests to use these algorithms for selection of attributes, the most significant part of the process of classification of a data set.

2. Algorithms for generation of decision trees

In our research we have used the following algorithms to generate decision trees: (i) Czerwiński's algorithm [8], (ii) our own implementation of the classical Quinlan algorithm [9], *i.e.* C4.5 using information entropy for attribute selection, (iii) the TVR algorithm (creating decision trees from fragments, which are sequences of paths from selected attributes to the decision attribute) [11] and (iv) VDP, a new algorithm searching for the most significant set of attributes, required for the correct classification of the training data. This algorithm has been the main subject of our research. The VDP algorithm is based on generating belief networks with varied Dirichlet's parameter [12]. Let us note that the descriptive attribute which has the greatest marginal influence [13] on the decision attribute is placed in the root of the decision tree.

3. Research methodology

Four algorithms for generating decision trees (Czerwiński's, Quinlan's, TVR and VDP) were compared using the following assumptions. All four algorithms were used for the NEVI dataset, which is presented in detail in [14]. In the initial stage of research, 250 cases of melanocytic skin lesions were randomly divided into two groups of 167 and 83 cases. Then, decision trees were generated for the 167 cases using all four algorithms. Quality of these trees was estimated on the basis of classification results of the 83 unseen cases. Obviously, the best algorithm is the algorithm that generates a decision tree with the lowest error rate.

4. Results

The generated decision trees are presented in Figures 1–4. Results of classification of unseen objects are presented in Table 1. For an idea of a mean number of questions see [15].

Table 1. Quality of the tested algorithms

Tested algorithm	Mean number of questions	Number of nodes	Error of classification [%]
Czerwiński's	2.70	7	2.41
Quinlan's	2.43	7	2.41
TVR	2.00	3	1.20
VDP	2.32	5	2.27

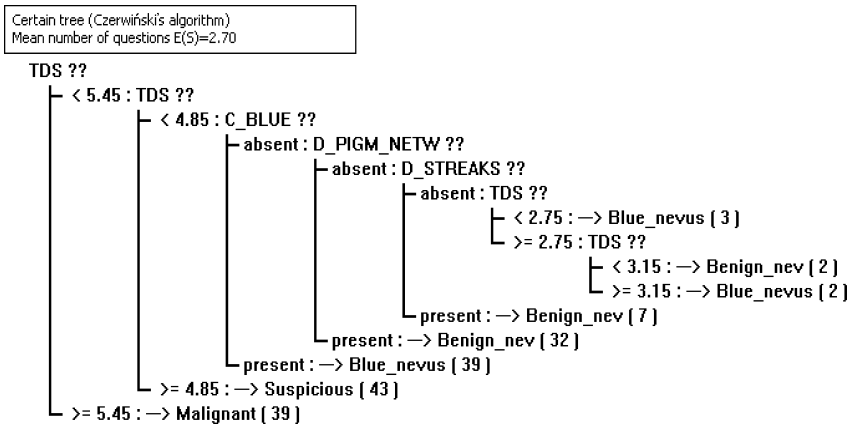

Figure 1. Decision tree generated by Czerwiński's algorithm

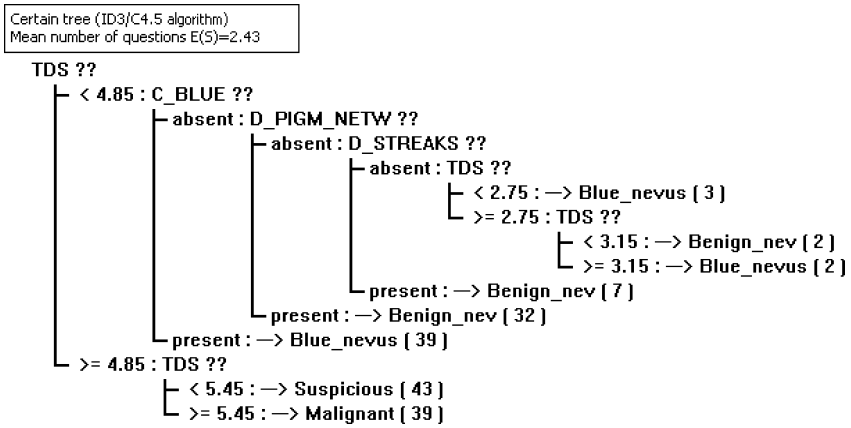

Figure 2. Decision tree generated by Quinlan's algorithm

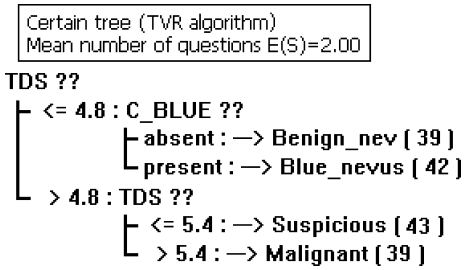

Figure 3. Decision tree generated by the TVR algorithm

5. Conclusions

Let us note that the error rates while using all four algorithms were low compared with *e.g.* a set of rules [1]. The TVR algorithm generates a decision tree with an unusually low error of classification of unseen objects (1.20%) requires verification. This algorithm pruned the decision tree by eliminating 4 cases from the source data set. However, using belief networks to improve the process of decision tree generation, we obtained surprisingly good results, better than the results obtained by means

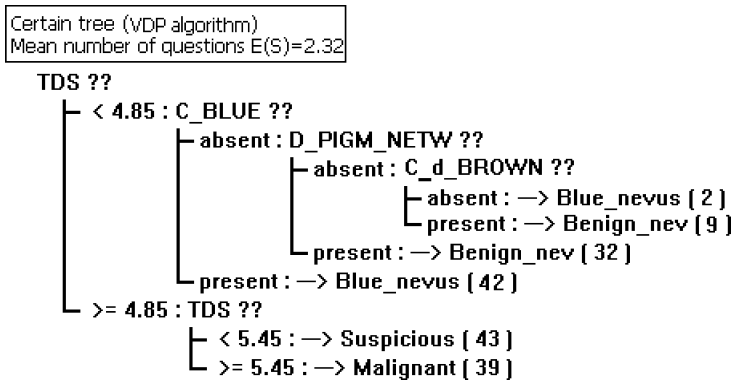

Figure 4. Decision tree generated by the VDP algorithm

of Czerwiński's algorithm and/or the classic Quinlan algorithm. In our research we observed that all four algorithms selected the ⟨TDS⟩ attribute for the root of the decision tree. In every case this attribute had the same range of value after the discretization process. Similarly, most of the used algorithms selected the ⟨C_BLUE⟩ attribute as the next test in the process of diagnosis of melanocytic skin lesions. The further selected attributes were different. The mean number of questions and the number of nodes with tests show a very similar character. In conclusion, we can say that the proposed new algorithm to generate quasi-optimal decision trees, applying Bayesian belief networks, yielded promising results. The algorithm requires further verification, especially in relation to decision tables containing attributes with mixed (numeric and symbolic) values.

Acknowledgements

Financial support of our research project No 7 T11E 030 21 obtained from the State Committee for Scientific Research (Warsaw) is gratefully acknowledged.

References

[1] Grzymała-Busse J W and Hippe Z S 2000 *Advances in Soft Computing (Intelligent Information Systems)* (Kłopotek M, Michalewicz M and Wierzchoń S T, Eds.), Physica-Verlag, Heidelberg, pp. 27–34

[2] Hippe Z S and Błajdo P 2002 *Methods of Artificial Intelligence* (Burczyński T, Cholewa W and Moczulski W, Eds.), Silesian University of Technology Edit. Office, Gliwice, pp. 181–185

[3] Lauria E J M and Tayi G K 2003 *Data Mining: Opportunities and Challenges* (Wang J, Ed.), Idea Group Publishing, Hershey (PA), pp. 260–277

[4] Hippe Z S, Knap M and Paja W 2002 *Methods of Artificial Intelligence* (Burczyński T, Cholewa W and Moczulski W, Eds.), Silesian University of Technology Edit. Office, Gliwice, pp. 177–180

[5] Hippe Z S and Wrzesień M 2002 *Methods of Artificial Intelligence* (Burczyński T, Cholewa W and Moczulski W, Eds.), Silesian University of Technology Edit. Office, Gliwice, pp. 185–189

[6] Grzymała-Busse J W 1997 *Fundameta Informaticae* **31** 27

[7] Grzymała-Busse J W, Hippe Z S, Knap M and Mroczek T 2003 *Knowledge Engineering and Expert Systems* (Bubnicki Z and Grzech A, Eds.), Wroclaw University of Technology Edit. Office, Wroclaw **1**, pp. 239–247 (in Polish)

[8] Czerwiński Z 1970 *Przegląd Statystyczny* **1970/2**, PWN, Warsaw (in Polish)

[9] Quinlan J R 1993 *C4.5 Programs for Machine Learning*, Morgan Kaufmann, San Mateo (CA)

[10] Hippe Z S and Mroczek T 2003 *Computer Recognition Systems* (Kurzyński M, Puchała E and Woźniak M, Eds.), Wroclaw University of Technology Edit. Office, Wroclaw, pp. 337–342
[11] Hippe Z S and Knap M 2003 *Algorithm for Generation of Decision Trees via Rules*, Internal Report, Department of Expert Systems and Artificial Intelligence, WSIZ, Rzeszow (in Polish)
[12] Jensen F V 2001 *Bayesian Networks and Decision Graphs*, Springer-Verlag, Heidelberg
[13] Heckerman D 1999 *A Tutorial on Learning with Bayesian Networks*, Technical Report MSR-TR-95-06
[14] Hippe Z S 1999 *TASK Quart.* **4** 483
[15] Dąbrowski A 1974 *On the Information Theory*, WSiP, Warsaw (in Polish)