RULES FOR MELANOMA SKIN CANCER DIAGNOSIS

Włodzisław Duch, Krzysztof Grąbczewski, Rafał Adamczak, Karol Grudziński

Department of Computer Methods, Nicholas Copernicus University, Grudziądzka 5, 87-100 Toruń, Poland; http://www.phys.uni.torun.pl/kmk

Zdzisław S. Hippe

Department of Expert Systems and Artificial Intelligence, University of Information Technology and Management, 35-225 Rzeszów, Poland

ABSTRACT

Analysis of medical data requires not only classification of patterns but also some data understanding. Several systems for extraction of logical rules from data have been applied to analysis of the melanoma skin cancer data. These systems include neural, decision tree and inductive algorithms for rule extraction and minimal-distance methods used for comparison. Very simple and accurate classification rules for the 4 types of melanoma are given.

INTRODUCTION

Expert systems for medical diagnosis support require logical rules for pattern classification. Machine learning methods allow for automatic acquisition of knowledge, extracting logical rules from the data. Early diagnosing of skin cancer is a very important practical problem. In this paper we continue [1, 2] experiments to find the optimal (accurate and comprehensible) set of decision rules suitable for distinguishing among 4 types of melanoma: Benign nevus, Blue nevus, Suspicious nevus, and Melanoma malignant. The data is described in the next section.

So far two types of rule extraction systems have been applied to this data. The LERS (Learning from Examples based on Rough Sets) system provides weighted rules [3, 4]. The GTS (General-To-Specific) algorithm [5, 6] is an inductive system starting from the most general rules and developing more specific decision rules in the learning process. A simple version of the nearest neighbor algorithm was also applied to this data [1]. Unfortunately statistical reliability of the rules has not been estimated in these studies since the test set was quite small. New rule extraction methods based on SSV (Separability Split Value) decision tree

[7], and neural methods based on constraint multilayer perceptron [8] and probability density estimation [9], have been used recently for medical data. Simple rules were discovered giving accurate results [10]. Selection of reference vectors in the similarity-based methods allow for prototype-based understanding of data [11]. These methods are applied here to the melanoma data.

In the next section the melanoma data used in our study is described, the third section contains brief description of the methods used, and the fourth section compares these results. The paper is finished with a short discussion.

DESCRIPTION OF DATA SETS USED

The data describing melanoma was collected in the Outpatient Center of Dermatology in Rzeszów, Poland [12]. The main database called NEVI-414 has been reviewed in details in [13]. Each of the cases belongs to one of the four types of Melanoma: benign, blue, suspicious, or malignant. Currently the data set NEVI-414 contains 250 cases, with almost equal distribution of all four classes of Nevi Pigmentosi.

The data contains a collection of descriptive attributes required for calculation of the TDS (Total Dermatoscopy Score) index [14], an important tool used in diagnosing of melanoma. Calculation of the TDS index is based on the 4 variables: *Asymmetry, Border, Color* and *Diversity* of the skin cancer mark's structure. Variable *Asymmetry* has three different values: symmetric-spot, 1-axial asymmetry, and 2-axial asymmetry. *Border* is a numerical attribute, with values from 0 to 8. *Color* has six possible values: white, blue, black, red, light brown, and dark brown, each coded as binary variables (several 1's are possible at the same time). *Diversity* has five values: pigment globules, pigment dots, branched strikes, structureless areas and pigment network, also coded as binary variables. In all of these 11 binary attributes for *Color* and *Diversity* the values are 0 or 1, 0 meaning lack of the corresponding property and 1 meaning the occurrence of the property.

Thus, every record in the database has 13 attributes. The TDS index is computed from these attributes using the following formula (known as the ABCD formula):

$$\mathrm{TDS} = 1.3\, \mathit{Asymmetry} + 0.1\, \mathit{Border} + 0.5 \sum \mathit{Colors} + 0.5 \sum \mathit{Diversities}$$

where for *Asymmetry* the value symmetric-spot counts as 0, 1-axial symmetry counts as 1, and 2-axial symmetry counts as 2, $\sum$ *Colors* is the sum of all values of the six color attributes and $\sum$ *Diversities* is the sum of all values of the five diversity attributes.

The second data set, NEVI-T14, contains collection of 26 newer, "unseen" cases used for testing the quality of the developed learning model represented by a set of decision rules, also with almost equal distribution in the four classes. All records in the testing set have the same structure as records in the training set. The data may be analyzed with or without the TDS coefficients.

METHODS USED

Since detailed description of the methods used in our study has already been published only the aspects relevant to evaluation of the results are presented here.

The GTS algorithm [5, 6] is a new type of covering algorithm, enhanced by the recursive reduction of the number of decision rules induced for a given information system. In its recent version the user may interactively guide the development of the learning model. The number of attributes used for classification is decreased in a controlled way; this may be supported by application of specially designed computer program system VVT [15]. Selection of the proper combination of attributes in the process of generating a learning model is based on two parameters, Frequency and Ranking. GTS allows to create many different sets of rules, depending on the user's choices, therefore in a complex situation it may be rather difficult to use.

SSV is a decision tree [7] based on a separability criterion, maximizing an index of separability for a given split value for continuous attribute or a subset of discrete values. Decision tree is easily converted into a set of crisp logical rules, with the number of correctly classified cases divided by the number of total cases that the rule classifies giving a measure of confidence in diagnosis. Various pruning techniques are used to ensure the simplest set of rules that will generalize well. Since the method is fully automatic and very efficient it can be used with crossvalidation tests to provide estimations of statistical accuracy of the extracted rules.

MLP2LN [8] allows to convert standard MLP neural network into a network performing logical function. This function is then written as a set of crisp logical rules, although fuzzy rules with "soft-trapezoid" membership functions may also be extracted. The procedure is almost fully automatic, giving the user a choice between the simplest possible description of the data and perhaps more accurate, but more complex, description.

FSM [9], Feature-Space Mapping, is a neural network estimating probability density of data using separable transfer functions. Each component of a transfer function may be interpreted as a context-dependent membership function. Using rectangular functions crisp logic rules are derived, while trapezoidal, triangular or Gaussian functions allow to derive fuzzy logic rules. The learning algorithm includes feature selection and network pruning.

SBM [11, 17], Similarity-Based-Methods includes many classification models based on evaluation of similarity. The software implementation of SBM, the Similarity-Based-Learner (SBL), is a very general program that includes various extensions of the k-nearest neighbor algorithms. Although crisp logical rules are not provided the program is capable of finding small number of prototype cases that are used to explain the data class structure.

COMPARISON OF RESULTS

Since the training data contains only 26 cases for 4 classes meaningful comparison of the statistical accuracy is not possible. A reasonable estimation of accuracy may be found by using all training (250) cases and test cases and performing 10-fold crossvalidation. Unfortunately such results are easy to obtain only for fully automatic and very efficient methods. Although we have not used risk matrices in our cost functions according to medical doctors identification of Blue nevus instead Benign nevus is acceptable, as both type of skin spots may be treated as belonging to a broader group, generally to Benign nevus.

GTS: After some experimentation among 14 descriptive attributes only a few were found to be important: TDS, color C-blue, and diversity D-structureless-areas. Various sets of decision rules were generated, selecting the following combinations of attributes: TDS & C-blue & Asymmetry & Border (4 attributes, based on the experience of medical doctors consulting our research), TDS & C-blue & D-structureless-areas) (3 attributes), TDS & C-Blue (2 attributes), and exclusively TDS (1 attribute). Learning models with 2-4 attributes had accuracy roughly on the level of 81-85% of error rate. On the other hand, the learning model based exclusively on the attribute TDS has lower effectiveness.

GTS generated a large number (198) of rules and only after combination and generalization of these rules a simplified version of 4 rules has been found [16]. These rules make only 6 errors on the training and no errors on the test set. TDS appears to be the most important attribute for correct distinction between Benign nevus, Suspicious nevus, and Melanoma malignant classes.

SSV: Our decision tree [7] finds the two relevant attributes (TDS, C-BLUE) and rules very efficiently, the only user-defined parameter being the degree of pruning. The default setting (based on maximization of crossvalidation accuracy on the training set) gives 98% accuracy (5 errors) on the training set and 100% accuracy on the test set. The rules are:

R1: IF TDS $\leq$ 4.85 AND C-BLUE IS absent THEN MELANOMA IS Benign-nevus
R2: IF TDS $\leq$ 4.85 AND C-BLUE IS present THEN MELANOMA IS Blue-nevus
R3: IF TDS $>$ 5.45 THEN MELANOMA IS Malignant
R4: IF TDS $>$ 4.85 AND TDS $<$ 5.45 THEN MELANOMA IS Suspicious

In the 10-fold crossvalidation on the training set rules derived from SSV (in each crossvalidation slightly different rules may be created) may give an average accuracy of 97.5%. Calculations were repeated 10 times giving the standard deviation of 0.30%. The result is quite close to 98% of accuracy for these rules on the training set, as should be expected. With lower degree of pruning 7 rules with 5 attributes are created, giving 99.2% accuracy (2 errors) on the training set and 92.3% (2 errors) on the test set. This set of rules is already too complex, showing some overfitting. Rule R1 is responsible for all 5 errors, covering 5 cases of the Blue-nevus class; as we have mentioned this type of error is rather innocent. The confusion matrix is given below:

Original class	Benign-nev	Blue-nevus	Malignant	Suspicious
Calculated				
Benign-nev	62	5	0	0
Blue-nevus	0	59	0	0
Malignant	0	0	62	0
Suspicious	0	0	0	62

The case “one-axial-symmetry, 5, 0, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 5.3, suspicious” that both LERS and kNN classified erroneously is correctly handled by these rules. Calculations with only 2 attributes (TDS and C-Blue) cannot be perfect since in the training set there are 5 pairs of vectors from different classes that are identical. This situation may create a problem for some learning algorithms.

MLP2LN constructive constrained multilayer perceptron algorithm [8] gave essentially the same rules as the SSV decision tree, increasing our confidence that this is indeed the simplest solution. Slight difference is that it labels the same 5 vectors that were wrongly assigned by the SSV tree as “unclassified".

FSM: This density estimation network [9] performs selection of features during learning. Rectangular functions were used to obtain logical rules. 7 such nodes are created on average, giving in 10-fold crossvalidation 95.5$\pm$ 1.0% and 100% accuracy on the test set. Creating a committee of 20 FSM networks does not change the training results (95.5$\pm$1.1% in CV) decreasing the test set accuracy (92.6%). In addition, the use of Gaussian functions decreases accuracy leading to more complex solutions (equivalent to 15 fuzzy rules), showing that for this datasets simplest solutions should strongly be preferred.

SBM: kNN results obtained previously [2] used $k = 1$ with Euclidean distance function. We have used our SBL program performing 10-fold crossvalidation calculations on the training set to select the type of the distance function – Manhattan was the best choice – and feature selection – TDS and C-Blue features were left as the most important. Perfect accuracy is obtained on the test set and the estimation of accuracy on the training set was similar to the accuracy of the SSV decision tree, 97.4$\pm$0.3%. Finally, we have used the prototype selection to reduce the number of reference vectors in each class. The selection procedure [17] left only 13 prototype vectors (7 for the first class and 2 for every other class) still giving 100% accuracy on the test set and 6 errors on the training set (237=250-13 vectors), corresponding to 97.5%.

Reducing the number of prototypes further to 7 (2 or 1 prototype per class) decreases the training set accuracy to 91.4% but makes no errors on the test set. This proves that one cannot judge the expected accuracy of classification methods using results obtained on small test set only. The training set accuracy may be improved by shifting the position of the prototype vectors. This approach provides an alternative to traditional rules that use intervals or subsets of values for single attributes. Rules here have the form: if the case X is similar

Table 1: Comparison of results on the NEVI-414 datasets. Accuracy on the training set estimated using the 10-fold crossvalidation method for SSV was 97.5±0.3%; in the table training set accuracy was obtained by applying the rules to the training set. For kNN the number of reference vectors is given instead of the number of rules.

Method	Rules	Training %	Test%
FSM, rectangular functions	7	95.5±1.0	100±0.0
FSM, Gaussian functions	15	93.7±1.0	95±3.6
GTS - initial, 2 features [16]	198	85	84.6
GTS - simplified	4	97.6	100
LERS, weighted rules	21	–	96.2
MLP2LN, crisp rules	4	98.0	100
SSV Tree, crisp rules	4	98.0	100
k-NN k=1 [2], all features	250	–	96.2
k-NN Manhattan, 2 features	250	97.4±0.3	100
k-NN Manhattan, 2 features	13	97.5	100

to the prototype *P* than it is of the same class as *P*. Logical rules partition the input space in hyperrectangles, while prototype-based rules use similarity function that measures distance from *P* providing spherical (for Euclidean) or other types of decision borders. For melanoma database only 2 features are used in the distance function, binary C-Blue and continuous TDS. Therefore in this case it is easy to write logical rules based on similarity to the prototype cases. Comparison of the number of rules and accuracy on the training/test sets is given in Table 1.

CONCLUSIONS AND FURTHER WORK

Several rule extraction systems have been applied to the skin cancer data. Out of 14 original input features contained in the database only 2 seem to be most important. Expected statistical accuracy of the rules derived with the decision tree (one rule per class) is around 97.5% and the errors came from assignment of Blue-nevus to the Benign-nevus classes, belonging to the same broader group. It remains to be seen how confident the experts are in the diagnosis and the data for these 5 cases. Before the rules presented here will be used for diagnosis support we should increase the database, including also non-cancer marks that should be distinguished from melanoma.

The same database was also analyzed using the LERS algorithm based on the rough set approach. The LEM2 algorithm implemented in LERS [4] found 21 weighted rules with a total of 81 conditions. These rules make only 1 error on the test set (suspicious melanoma is diagnosed as malignant), but without crossvalidation tests on the training set it is hard to

judge their accuracy.

An interesting aspect that has not been investigated in this paper is feature aggregation. TDS is a single parameter obtained as a linear combination of other input values, containing much more information than any single feature. A good classification method (for example, MLP neural network with 2 hidden layers) should be able to automatically discover and optimize such combination. What is the best data representation for this problem? This issue should be investigated further.

Acknowledgments: Support of the Toruń group by the Polish Committee for Scientific Research, grant no. 8 T11C 006 19, is gratefully acknowledged. We would like to thank the Outpatient Center of Dermatology in Rzeszów, Poland, directed by Dr. S. Bajcar, for support and consultations.

References

[1] Grzymała-Busse J.W., Hippe Z.S, Data mining experiments with a Melanoma training set. In: Kłopotek M., Michalewicz M., Wierzchoń S.T. (Eds.) Intelligent Information Systems, Physica-Verlag, Heidelberg 2000, pp. 27-34.

[2] Grzymała-Busse J.W., Hippe Z.S., (2001) Melanoma prediction using k-Nearest Neighbor and LEM2 algorithms. (accepted for presentation during 10th International Symposium on Intelligent Information Systems, Poland, June 18-22, 2001).

[3] Grzymała-Busse J.W , (1992) LERS - A system for learning from examples based on rough sets. In: Słowiński R. (Ed.) Intelligent Decision Support. Handbook of Applications and Advances of the Rough Sets Theory, Kluwer Academic Publishers, Norwell (MA), pp. 3-18.

[4] Grzymała-Busse J.W, LERS - A knowledge discovery system. In: Polkowski L., Skowron A. (Eds.) Rough Sets in Knowledge Discovery 2. Applications, Case Studies and Software Systems, Physica-Verlag, Heidelberg 1998, pp. 562-565.

[5] Hippe Z.S., Hippe T.M., (1997) An attempt to automatize modeling of medical data. In: Kącki E. (Ed.) Computers in Medicine, Polish Society of Medical Informatics, Łódź, pp. 24-31.

[6] Hippe Z.S., Machine Learning - A promising strategy for business information processing? In: Abramowicz W. (Ed.) Business Information Systems'97, Academy of Economy Edit. Office, Poznań 1997, pp. 603-622.

[7] Grąbczewski, K., Duch, W., A general purpose separability criterion for classification systems. 4th Conf. on Neural Networks and Their Applications, Zakopane, Poland, May 1999, pp. 203–208

[8] Duch, W., Adamczak, R., Grąbczewski, K., Methodology of extraction, optimization and application of crisp and fuzzy logical rules. IEEE Transactions on Neural Networks 12 (2001) March issue.(in print)

[9] Duch, W., Adamczak, R., Jankowski, N., New developments in the Feature Space Mapping model. 3rd Conf. on Neural Networks, Kule, Poland, Oct. 1997, pp. 65-70

[10] Duch, W., Adamczak, R., Grąbczewski, K., Żal, G., Hayashi, Y., Fuzzy and crisp logical rule extraction methods in application to medical data. In: P.S. Szczepaniak, P.J.G. Lisboa, J. Kacprzyk (eds.), Fuzzy systems in medicine. Physica - Verlag, Springer, Berlin, Heidelberg, New York 2000, 593-616

[11] Duch W., Similarity based methods: a general framework for classification, approximation and association, Control and Cybernetics 29, No. 4 (2000)

[12] Bajcar S., Grzegorczyk L., Endangerment by skin cancer among population of south-east part of Poland. Hospital No. 1, Research Report, Rzeszów 1997.

[13] Hippe Z.S., Data mining in medical diagnosis. In: Kącki E. (Ed.) Computers in Medicine, Polish Society of Medical Informatics, Łódź 1999, Vol. 1, pp. 25-34.

[14] Braun-Falco O., Stolz W., Bilek P., Merkle T., Landthaler M., Das Dermatoskop. Eine Vereinfachung der Auflichtmikroskopie von pigmentierten Hautveränderungen, Hautarzt 40 (1990) 131-135.

[15] Hippe Z.S., Iwaszek G., From research on a new method of development of quasi-optimal decision trees. In: Kłopotek M., Michalewicz M., Wierzchoń S.T. (Eds.) Intelligent Information Systems IX, Instytut Informatyki PAN, Warszawa 2000, pp. 31-35.

[16] Grzymała-Busse J.W., Hippe Z.S., Application of covering algorithm for classification of melanoid marks on the skin. Systems-Modelling-Control, Zakopane, Poland, May 2001 (in print).

[17] Grudziński K, Duch W., SBL-PM: A Simple Algorithm for Selection of Reference Instances for Similarity Based Methods, Intelligent Information Systems IIS'2000, Physica Verlag, Springer 2000, pp. 99-108