Medical Data Analysis Using Self-Organizing Data Mining Technologies

Frank Lemke[1], Johann-Adolf Müller[2]

[1] Script Software, Germany
frank@knowledgeminer.net

[2] HTW Dresden, Fachbereich Informatik/Mathematik
muellerj@informatik.htw-dresden.de

Abstract. "KnowledgeMiner" was designed to support the knowledge extraction process on a highly automated level. Implemented are 3 different GMDH-type self-organizing modeling algorithms to make knowledge extraction systematically, fast, successful and easy-to-use even for large and complex system such as one of the most complex systems: the human. Self-organizing data mining technologies in medical data analysis have to select automatically useful knowledge for medical decisions, such as diagnosis of heart disease.

1 Self-Organizing Data Mining

Today, there is an increased need to discover information - contextual data - non obvious and valuable for decision making from a large collection of data efficiently. This is an interactive and iterative process of various subtasks and decisions and is called Knowledge Discovery from Data. The engine of Knowledge Discovery - where data is transformed into knowledge for decision making - is Data Mining.

In the past years there has been much publicity about the ability of Artificial Neural Networks to learn and to generalize despite important problems with design, development and application of Neural Networks. In contrast to Neural Networks that use

- Genetic Algorithms as an external procedure to optimize the network architecture and
- several pruning techniques to counteract overtraining,

self-organizing data mining [1] introduces principles of evolution - inheritance, mutation and selection - for generating a network structure systematically enabling automatic model structure synthesis and model validation. Models are generated adaptively from data in form of networks of active neurons in an evolutionary fashion of repetitive generation of populations of competing models of growing complexity, their validation and selection until an optimal complex model - not too simple and not too complex - have been created. That is, growing a tree-like network out of seed information (input and output variables' data) in an evolutionary fashion of pairwise combination and survival-of-the-fittest selection from a simple single individual (neuron) to a desired final, not overspecialized behavior (model). Neither, the number of neurons and the number of layers in the

network, nor the actual behavior of each created neuron is predefined. All this is adjusting during the process of self-organization, and therefore, is called self-organizing data mining [1].

KnowledgeMiner (http://www.knowledgeminer.net) is a powerful and easy-to-use modeling tool which was designed to support the knowledge extraction process on a highly automated level and which has implemented three advanced self-organizing modeling technologies at present (table 1): GMDH, Analog Complexing (AC) and Fuzzy rule induction using GMDH (FRI) [1].

Table 1. Algorithms for self-organizing modeling

Data Mining functions	Algorithm
classification	GMDH, FRI, AC
clustering	AC
modeling	GMDH, FRI
time series forecasting	AC, GMDH, FRI
sequential patterns	AC

2 Diagnosis of Heart Disease

2.1 Problem

Diagnosis of diseases is an important and difficult task in medicine. Detecting a disease from several factors or symptoms is a many-layered problem that also may lead to false assumptions with often unpredictable effects. Therefore, the attempt of using the knowledge and experience of many specialists collected in databases to support the diagnosis process seems reasonable. The goal of this example application refers to the presence of heart disease in the patient. The target variable distinguishes between five levels of heart disease: 0 - no presence and 1, 2, 3, 4 - presence at a gradually increased level. So the problem is twofold: 1. a binary classification problem (0/1) on attempting to distinguish presence (level 1-4) from absence (level 0) and 2. attempting to find a model that classifies the five levels of disease most accurately (0-4).

We have used the Long Beach data set (http://www.ics.uci.edu). It contains 200 cases and 76 attributes, but all published experiments refer to using a selected subset of 14 attributes.

2.2 Application of a Nonlinear Model

We started with the reduced Long Beach data set of 14 attributes using GMDH, and then we created a GMDH model from the complete data set (76 attributes) to figure out how the other attributes would contribute to the model. We divided the data set into 180 cases for learning and 20 cases for evaluating how the models do on new data.

LB14: In KnowledgeMiner the three best nonlinear GMDH models are stored in a model base for comparison and synthesis purposes. Any GMDH model is described in analytical form by a regression equation. The LB14 data set is described by this model, for example:

Y= + 9.939e-2z21 + 3.971e-1z22 - 8.173e-1z21z22 + 5.750e-1z21z21 + 7.450e-1
z21= + 1.000e+0z11
z11= + 1.676e-1X10 + 1.723e-2X10X10 - 1.645e-1
z22= + 7.838e-1z11 + 7.859e-1z12
z11= + 2.564e-2X1 + 3.329e-1X3 - 2.689e+0
z12= + 5.218e-1X9 - 6.124e-2X6X9 + 5.602e-2X9X9 - 2.974e-1 ,

where X1 - age, X3 - cp, X6 - fbs, X9 - exang, X10 - oldpeak, and Y - binary (has/has not) target variable. Table 2 shows the classification results on both learning and testing data and figure 1 plots the Receiver Operating Characteristics (ROC) curve of the three best models M1-M3 and of a synthesized model for the binary classification task. The results show that these models have rather poor classification power.

Table 2. Classification results for the Long Beach data set

		LB 14		**LB 76**		
		GM DH		*GM DH*		*FRI*
		0 / 1	0 - 4	0 / 1	0 - 4	0 / 1
train	false classified	37	116	2	24	3
	Accuracy [%]	79,44	35,56	98,89	86,67	98,33
test	false classified	7	16	2	5	2
	Accuracy [%]	65,00	20,00	90,00	75,00	90,00

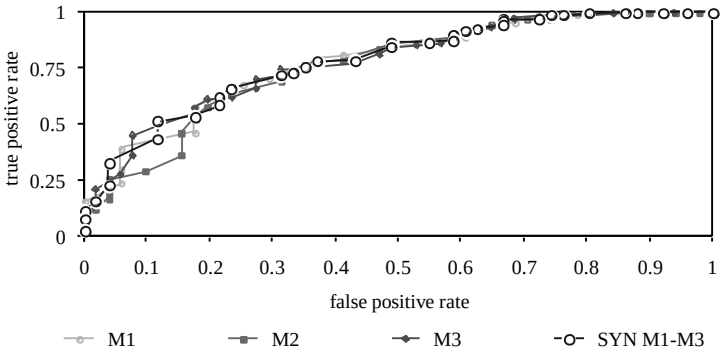

Figure 1. ROC curves of LB14 GMDH models

LB76: The complete data set of 76 attributes contains some factors that are stated as unused so that 64 variables were used effectively, and a created model describing the two classes is composed of these input factors, e.g.: X4 - painexer (1 = provoked by exertion; 0 = otherwise), X17 - ekgmo (month of exercise ECG reading), X49 - ladprox, X50 - laddist, X51 - diag, X52 - cxmain, X54 - om1, X56 - rcaprox,. The data source information indicates that the majority of the input factors (X49 - X56) selected in this model (and the other models as well) are vessels. This seems surprising, because these factors were not included as important in the LB14 data set a priori. Additionally, this model shows an

significantly increased classification capability (table 2), also compared with other published results of about 80% accuracy for the binary classification task. The corresponding ROC curves for the set of LB76 models (figure 2) underlines the classification power.

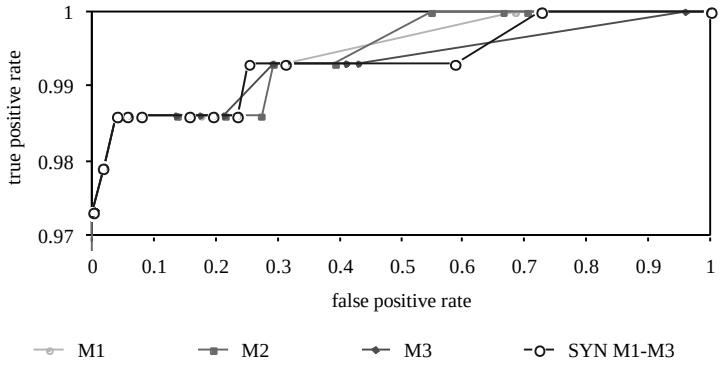

Figure 2. ROC curves of LB76 GMDH models

2.3 Application of Fuzzy Rules

For FRI each attribute was fuzzified into 5 linguistic variables. To get a sharper difference between presence and absence of disease, we transformed the target variable into -1 for absence, and 1-4 for presence correspondingly. From Fuzzy Rule Induction using the LB76 data set these two models were generated for the binary classification task (table 2):

IF PB_om1 **OR** PB_ladd **OR** PB_cxmain **OR** PB_ladp **OR** PB_rcap **THEN** Presence

IF NB_junk **OR** PB_ladd **OR** PB_cxmain **OR** PB_ladp **OR** PB_rcap **THEN** Presence

The ROC analysis results are displayed in figure 3.

When modeling all five levels separately, this model was obtained for absence of disease:

IF NB_om1 **&** NB_cxmain **&** NB_rcap **&** NB_ladp **&** NB_cxmain **&** NB_ladd **&** NB_rcap **THEN** Absence

and the following rules are created for the different levels of presence:

IF NS_slop **OR** PB_slop **&** PS_cyr **OR** NB_ladp **&** ZO_slop **&** PB_cyr **OR** NB_cyr **OR** NS_cyr **&** NB_cxmain **OR** NB_ladp **&** ZO_slop **&** PB_cyr **THEN** NS_class (class 1)

IF PM_old **&** PB_om1 **OR** ZO_slop **&** PM_cyr **OR** PM_old **&** PB_om1 **OR** NS_cyr **&** PB_om1 **THEN** ZO_class (class 2)

F NB_old **&** PB_om1 **OR** PB_ladd **&** PB_rcap **OR** PB_ladp **&** PB_cxmain **&** PB_cxmain **&** PB_rcap **OR** ZO_cyr **&** PB_om1 **OR** PB_ladd **&** PB_rcap **OR** PB_cxmain **&** PB_ladp **&** PB_rcap **THEN** PS_class (class 3)

IF PM_old **&** NB_slope **OR** PB_old **&** NB_ladp **THEN** PB_class (class 4)

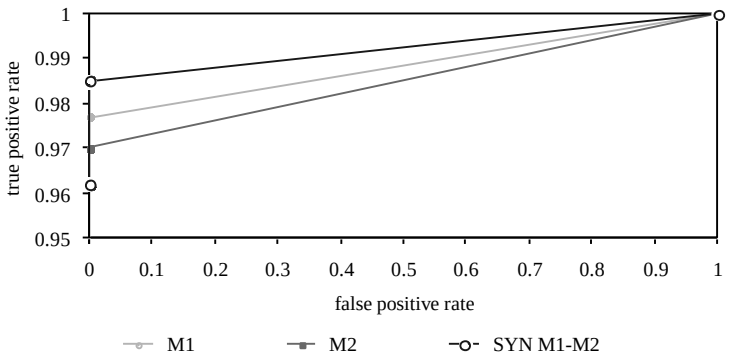

Figure 3. ROC curves of LB76 Fuzzy models

2.4 Classification Using Analog Complexing

"KnowledgeMiner" provides an Analog Complexing [1] algorithm for prediction, clustering and classification. It can be considered as a sequential pattern recognition method. Using the LB14 and LB76 data sets classification results are generated, which are shown in table 3 and figure 4, 5. By means of clustering in the sample space can be estimated clusters of variables that similarly influence the objects (nucleus). Classifications based on estimated nuclei (*LB14:* 11 variables excluding X2 (sex), X7 (restecg) and X12 (ca); *LB76:* X1 (age in years), X49 (ladprox), X50 (laddist), X52 (cxmain), X54 (om1)) are presented in table 3 and figure 4,5.

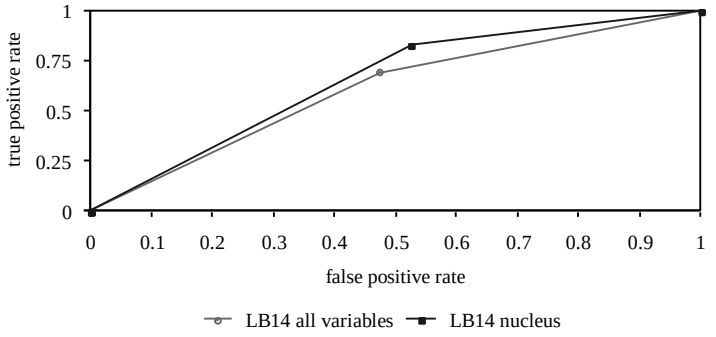

Figure 4. ROC analysis of LB14 Analog Complexing classification models

Table 3. Analog Complexing classification results

	LB 14		**LB 76**	
	Has/Has	*Not*	*Has/Has*	*Not*
classified	all variables	nucleus	all variables	nucleus
false	81	69	64	0
Accuracy [%]	59,5	65,5	68	100

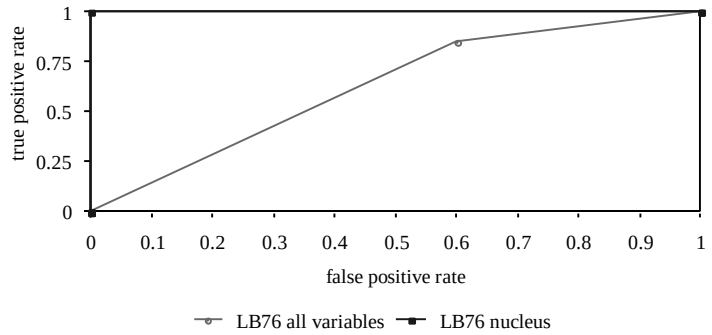

Figure 5. ROC analysis of LB76 Analog Complexing classification models

3 Conclusions

Self-organizing data mining technologies have shown their power to extract valuable information for classification purposes. Based on their strong advantages, speed, self-selection of relevant input variables and generation of an analytical model for interpretation, the complete data set of 76 attributes was applicable and has proven to significantly increase classification accuracy using a small subset of attributes only.

This overall result was confirmed on several other medical/bio-chemical modeling and classification tasks like the carcinogenicity prediction of aromatic compounds from a set of molecular descriptors (COMET project, ENV4-CT97-0508, funded by the European Commission).

With the extracted knowledge, detection of disease may also get more efficient by reducing both time and costs for the corresponding procedure. Besides classification accuracy, the efforts (time, costs) for creating/applying a classification model are quite important also. Since self-organizing data mining selects a subset of attributes necessary to obtain a certain classification quality, some cost saving effects may appear from this perspective also.

Table 4 examplarily expresses this advantage compared with a model that would use all 13 provided variables (like many Neural Networks do, e.g.).

Table 4. Cost reduction from using a GMDH model (LB14)

	all variables	GMDH model
Costs per patient [CAN$]	600,57	181,80
Costs saved [%]		69,73

References

[1] Müller, J.-A., F. Lemke: Self-Organizing Data Mining. Libri, Hamburg 2000, ISBN 3-89811-861-4