EXPLORIS

The general architecture of the modeling process for developing a memetic algorithm for risk assessment of patients with suspected CAD

The models used in the project for diagnostic of coronary artery disease (CAD) have been built with the AI-X-ENGINE which is an advanced data mining and predictive modeling software package based on enhanced methods from the area of machine learning and AI. The AI-X-ENGINE is based on a multi-level modeling architecture, which delivers the highest possible quality of classification and prediction by combining, evaluating, and optimizing the use of various methods in an automated self-learning process based on an evolutionary optimization procedure.[1,2] The imitation of evolutionary principles in pattern recognition and variable selection enables a sophisticated combination and nesting of different methods from the field of artificial intelligence. Models are built using a combination of associative mathematical structures like an Artificial Neuronal Network (ANN), particularly Self-Organizing Maps (SOM), or Decision Tree Ensembles and other Ensemble Methods with an Artificial Evolutionary Procedure (AEVOP)[3,4]. The methods used is best conceived as a heterogeneous adaptive system (HAS)[5], employing competitive learning driven by artificial evolution as a basis, and using statistical models for selecting the finally desired model. The ensemble methods and SOM are able to detect autonomously hidden correlational structures, even if they are non-linear. The approach reaches this target using multi-level re-sampling and cross validation where each pattern goes through a number of stability and performance checks competing with other patterns for survival and only the fittest survive in a final model.

An outstanding characteristic of the approach is that the whole modeling process has been highly automated and standardized. No pre-evaluation of data attributes by experts is needed – the software detects hidden relationship in a given data sample (e.g., a clinical study) by its own and evaluates the relevant prognostic factors.

The main advantages of the process used for building the diagnostic models are that

a. The process is able to find and describe small, multivariate patterns in the data and to combine them together into a high-performance model.

b. The process is highly automated. In the AI-X-ENGINE the manual try and error method is replaced with a directed search for the best possible methods and attributes combinations which are performed automatically by the system. As a result, a highly effective model is delivered much faster and without large efforts from the modeling person.

c. The process is designed for working with data of large dimensionality. The feature selection and dimensionality tasks are performed automatically. The process provides specialized methods (for example, based on classifiers voting procedure, data noising and random records generation) specifically for small data samples.

EXPLORIS

d. Several classification models, optimized for different alpha/beta-errors costs-ratio or built of different data samples, can be combined into a single nested model.

e. The AI-X-engine provides advanced tools for data cleaning, preparation and transformation including more than 80 methods for data transformation.

MPA calibration according ESC guidelines

For even better prediction of the CAD risk the MPA model thresholds were calibrated on the base of the new ESC-guidelines and a priori information from the Luric paper without using any information from the low-intermediate risk cohort.

CAD risk classes	MPA model original	MPA model calibrated	MPA-model ESC adjusted
Class 1 Very low risk	0·0% (16·5%)	0·0%(16·7%)	4·2% (67·7%)
Class 2 Low risk	3·3% (17·2%)	5·6% (51·0%)	21·6% (16·7%)
Class 3 Medium risk	7·1% (34·5%)	21·6% (16·7%)	
Class 4 High risk	20·6% (9·1%)	46·0% (7·2%)	46 ·0% (7·2%)
Class 5 Very high risk	50·0% (22·7%)	76·3% (8·5%)	76·3% (8·5%)

Results in prevalence of CAD percentage in a class with in parenthesis is the percentage of the population in this class. Prevalence of CAD in the total CVC population is 16%. CAD: coronary artery disease; MPA: memetic pattern-based; Green: effective risk for CAD <5%, excluding CAD without further testing; yellow: effective risk for CAD 5-70%, requiring further non-invasive testing; orange: effective risk of CAD >70%, requiring direct invasive angiography. No model provided a group with sufficient prevalence to make the diagnosis of CAD (i.e. >85%).

[1] W. E. Hart, N. Krasnogor and J. E. Smith, Memetic Evolutionary Algorithms, Studies in Fuzziness and Soft Computing, 2005, Volume 166/2005, 3-27

EXPLORIS

[2] Merz P. (2001). On the Performance of Memetic Algorithms in Combinatorial Optimization. 2nd Workshop on Memetic Algorithms, GECCO, San Francisco, CA, USA.

[3] Stanley K.O., Miikkulainen R. (2002). Efficient Evolution of Neural Network Topologies. Proceedings of the 2002 Congress on Evolutionary Computation (CEC '02). Honolulu, Hawaii: IEEE

[4] Sexton, R. S., Dorsey, R. E., and Johnson, J. D. (1999). Optimization of neural networks: A comparative analysis of the genetic algorithm and simulated annealing, European Journal of Operational Research, 114: 589-601

[5] Duch W., Grabczewski K. (2002). Heterogeneous adaptive systems. World Congress of Computational Intelligence, Honolulu, May 2002