Identification of Cancer Diagnosis Estimation Models Using Evolutionary Algorithms - A Case Study for Breast Cancer, Melanoma, and Cancer in the Respiratory System

Stephan M. Winkler
Upper Austria University of Applied Sciences
Department of Bioinformatics
Softwarepark 11
4232 Hagenberg, Austria
stephan.winkler@fh-hagenberg.at

Michael Affenzeller
Upper Austria University of Applied Sciences
Department of Software Engineering
Softwarepark 11
4232 Hagenberg, Austria
michael.affenzeller@fh-hagenberg.at

Witold Jacak
Upper Austria University of Applied Sciences
Department of Software Engineering
Softwarepark 11
4232 Hagenberg, Austria
witold.jacak@fh-hagenberg.at

Herbert Stekel
General Hospital Linz
Central Laboratory
Krankenhausstraße 9
4021 Linz, Austria
herbert.stekel@akh.linz.at

ABSTRACT

In this paper we present results of empirical research work done on the data based identification of estimation models for cancer diagnoses: Based on patients' data records including standard blood parameters, tumor markers, and information about the diagnosis of tumors we have trained mathematical models for estimating cancer diagnoses.

Several data based modeling approaches implemented in HeuristicLab have been applied for identifying estimators for selected cancer diagnoses: Linear regression, k-nearest neighbor learning, artificial neural networks, and support vector machines (all optimized using evolutionary algorithms) as well as genetic programming. The investigated diagnoses of breast cancer, melanoma, and respiratory system cancer can be estimated correctly in up to 81%, 74%, and 91% of the analyzed test cases, respectively; without tumor markers up to 75%, 74%, and 87% of the test samples are correctly estimated, respectively.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining; I.2.8 [**Artificial Intelligence**]: Heuristic methods; J.3 [**Life and Medical Sciences**]: Medical Information Systems

GECCO'11, July 12–16, 2011, Dublin, Ireland.
Copyright 2011 ACM 978-1-4503-0690-4/11/07 ...$10.00.

General Terms

Algorithms, Reliability, Experimentation, Standardization

Keywords

Cancer Diagnosis Estimation, Tumor Marker Data, Data Mining, Machine Learning, Statistical Analysis

1. INTRODUCTION

In this paper we present research results achieved within the research center *Heureka!*[1]: Data of thousands of patients of the General Hospital (AKH) Linz, Austria, have been analyzed in order to identify mathematical models for cancer diagnoses. We have used a medical database compiled at the central laboratory of AKH in the years 2005 – 2008: 28 routinely measured blood values of thousands of patients are available as well as several tumor markers (substances found in humans that can be used as indicators for certain types of cancer). Not all values are measured for all patients, especially tumor marker values are determined and documented only if there are indications for the presence of cancer. The results of empirical research work done on the data based identification of estimation models for cancer diagnoses are presented in this paper: Based on patients' data records including standard blood parameters, tumor markers, and information about the diagnosis of tumors we have trained mathematical models for estimating cancer diagnoses.

The following data based modeling methods (implemented in HeuristicLab [29]) have been used for producing classifiers: Linear regression, k-nearest neighbor classification,

[1] Josef Ressel Center for Heuristic Optimization; `http://heureka.heuristiclab.com/`

neural networks, support vector machines, and genetic programming.

In the following section (Section 2) we describe the database we have used for our research work as well as the tumor markers for which we have developed classifiers; we also describe the data preprocessing steps. For each tumor for which we have developed classifiers we define the sets of input variables used in this research project. In Section 3 we describe the modeling methods used in this research project as well as the parameter settings applied, and in Section 4 we summarize and analyze the modeling results we have achieved. The conclusion of this paper is given in Section 5, followed by references and an appendix in which we summarize optimized modeling details.

2. DATABASE

2.1 Available Patient Data

The blood data measured at the AKH in the years 2005–2008 have been compiled in a database storing each set of measurements (belonging to one patient): Each sample in this database contains an unique ID number of the respective patient, the date of the measurement series, the ID number of the measurement, and a set of parameters summarized in Table 11; standard blood parameters are stored as well as tumor marker values and cancer diagnosis information. Patients personal data were at no time available to the authors except the head of the laboratory.

In total, information about 20,819 patients is stored in 48,580 samples. Please note that of course not all values are available in all samples; there are many missing values simply because not all blood values are measured during each examination. Further details about the data set can for example be found in [33].

2.1.1 Standard Parameters

Information about the blood parameters stored in the AKH database (which are listed in the upper part of Table 11 at the end of this paper) can be found in [20] and [31], e.g.

2.1.2 Tumor Markers

In general, tumor markers are substances found in humans (especially in the blood or in body tissues) that can be used as indicators for certain types of cancer. There are several different tumor markers which are used in oncology to help detect the presence of cancer; elevated tumor marker values can indicate the presence of cancer, but there can also be other causes. As a matter of fact, elevated tumor marker values themselves are not diagnostic, but rather suggestive; tumor markers can be used to monitor the result of a treatment (as for example chemotherapy).

Literature discussing tumor markers, their identification, their use, and the application of data mining methods for describing the relationship between markers and the diagnosis of certain cancer types can be found for example in [16] (where an overview of clinical laboratory tests is given and different kinds of such test application scenarios as well as the reason of their production are described), [27], [36], [6], and [37].

Information about the tumor markers stored in the AKH database are listed in the lower part of Table 11.

2.1.3 Cancer Diagnoses

Finally, information about cancer diagnoses is also available in the AKH database: If a patient is diagnosed with any kind of cancer, then this is also stored in the database.

Our goal in the research work described in this paper is to identify estimation models for the presence of the following types of cancer: Malignant neoplasms in the respiratory system (RSC, cancer classes C30–C39 according to the International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10)), melanoma and malignant neoplasms on the skin (Mel, C43–C44), and breast cancer (BC, C50).

2.2 Data Preprocessing

Before analyzing the data and using them for training classifiers we have preprocessed the available data:

- All variables have been linearly scaled to the interval [0;1]: For each variable v_i, the minimum value min_i is subtracted from all contained values and the result divided by the difference between min_i and the maximum plausible value $maxplau_i$; all values greater than the given maximum plausible value are replaced by 1.0.

- All samples belonging to the same patient with not more than one day difference with respect to the measurement data have been merged. This has been done in order to decrease the number of missing values in the data matrix. In rare cases, more than one value might thus be available for a certain variable; in such a case, the first value is used.

- Additionally, all measurements have been sample-wise re-arranged and clustered according to the patients' IDs. This has been done in order to prevent data of certain patients being included in the training as well as in the test data.

Before starting the modeling algorithms for training classifiers we had to compile separate data sets for each analyzed target tumor t_i: First, blood parameter measurements were joined with diagnosis results; only measurements and diagnoses with a time delta less than a month were considered. Second, all samples containing measured values for t_i are extracted. Third, all samples are removed that contain less than 15 valid values. Finally, variables with less than 10% valid values are removed from the data base.
This procedure results in a specialized data set dst_i for each tumor marker t_i. In Table 1 we summarize statistical information about all resulting data sets for the markers analyzed here; the numbers of samples belonging to each of the defined classes are also given for each resulting data set.

3. MODELING METHODS

In this section we describe the modeling methods applied for identifying estimation models for cancer diagnosis: On the one hand we apply hybrid modeling using machine learning algorithms and evolutionary algorithms for parameter optimization and feature selection (as described in Section 3.1), on the other hand apply use genetic programming (as described in Section 3.2).

Table 1: Overview of the data sets compiled for selected cancer types

Cancer Type	*Input Variables*	*Total Samples*	*Samples in*		Missing Values
			Class 0	*Class 1*	
`Breast Cancer`	`AGE, SEX, AFP, ALT, AST, BSG1, BUN, C125, C153, C199, C724, CBAA, CEA, CEOA, CH37, CHOL, CLYA, CMOA, CNEA, CRP, CYFS, FE, FER, FPSA, GT37, HB, HDL, HKT, HS, KREA, LD37, MCV, NSE, PLT, PSA, PSAQ, RBC, S100, SCC, TBIL, TF, TPS, WBC`	706	324 (45.89%)	382 (54.11%)	46.67%
`Melanoma`		905	485 (53.59%)	420 (46.41%)	47.79%
`Respiratory System Cancer`		2,363	1,367 (57.85%)	996 (42.15%)	44.76%

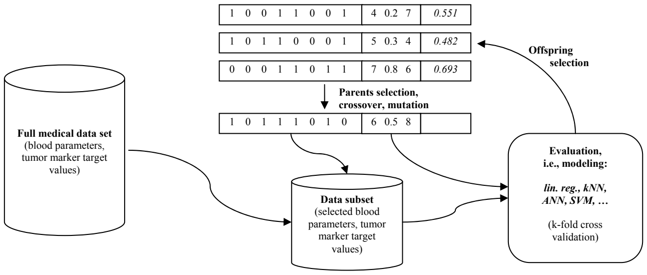
1 0 0 1 1 0 0 1 | 4 0.2 7 | 0.551
1 0 1 1 0 0 0 1 | 5 0.3 4 | 0.482
0 0 0 1 1 0 1 1 | 7 0.8 6 | 0.693
Offspring selection
Parents selection, crossover, mutation
1 0 1 1 1 0 1 0 | 6 0.5 8
Full medical data set (blood parameters, tumor marker target values)
Data subset (selected blood parameters, tumor marker target values)
Evaluation, i.e., modeling:
lin. reg., kNN, ANN, SVM, ...
(k-fold cross validation)

Figure 1: A hybrid evolutionary algorithm for feature selection and parameter optimization in data based modeling.

3.1 Hybrid Modeling Using Machine Learning Algorithms and Evolutionary Algorithms for Parameter Optimization and Features Selection

3.1.1 General Modeling Approach: Definition and Evaluation of Solution Candidates

Feature selection is often considered an essential step in data based modeling; it is used to reduce the dimensionality of the datasets and often conducts to better analyses. Given a set of n features $F = \{f_1, f_2, \ldots, f_n\}$, our goal here is to find a subset $F' \subseteq F$ that is on the one hand as small as possible and on the other hand allows modeling methods to identify models that estimate given target values as well as possible. Additionally, each data based modeling method (except plain linear regression) has several parameters that have to be set before starting the modeling process.

The fitness of feature selection F' and training parameters with respect to the chosen modeling method is calculated in the following way: We use a machine learning algorithm m (with parameters p) for estimating predicted target values $est(F', m, p)$ and compare those to the original target values $orig$; the coefficient of determination (R^2) function is used for calculating the quality of the estimated values. Additionally, we also calculate the ratio of selected features $|F'|/|F|$. Finally, using a weighting factor α, we calculate the fitness of the set of features F' using m and p as

$$fitness(F', m, p) = \\ \alpha * |F'|/|F| + (1 - \alpha) * (1 - R^2(est(F', m, p), orig)). \quad (1)$$

As an alternative to the coefficient of determination function we can also use a classification specific function that calculates the ratio of correctly classified samples, either in total or as the average of all classification accuracies of the given classes (as for example described in [32], Section 8.2): For all samples that are to be considered we know the original classifications $origCl$, and using (predefined or dynamically chosen) thresholds we get estimated classifications $estCl(F', m, p)$ for estimated target values $est(F', m, p)$. The total classification accuracy $ca_k(F', m, p)$ is calculated as

$$ca(F', m, p) = \frac{|\{j : estCl(F', m, p)[j] = origCl[j]\}|}{|estCl|} \quad (2)$$

Class-wise classification accuracies $cwca$ are calculated as the average of all classification accuracies for each given class $c \in C$ separately:

$$ca(F', m, p)_c = \\ \frac{|\{j : estCl(F', m, p)[j] = origCl[j] = c\}|}{|\{j : origCl[j] = c\}|} \quad (3)$$

$$cwca(F', m, p) = \frac{\sum_{c \in C} ca(F', m, p)_c}{|C|} \quad (4)$$

We can now define the classification specific fitness of feature selection F' using m and p as

$$fitness_{ca}(F', m, p) = \\ \alpha * |F'|/|F| + (1 - \alpha) * (1 - ca(F', m, p)) \quad (5)$$

or

$$fitness_{cwca}(F', m, p) = \\ \alpha * |F'|/|F| + (1 - \alpha) * (1 - cwca(F', m, p)). \quad (6)$$

In [3], for example, the use of evolutionary algorithms for feature selection optimization is discussed in detail in the context of gene selection in cancer classification; in [34] we have analyzed the sets of features identified as relevant in the modeling of tumor markers AFP and CA15-3.

We have now used evolutionary algorithms for finding optimal feature sets as well as optimal modeling parameters for models for tumor diagnosis; this approach is schematically shown in Figure 1. A solution candidate is here represented as $[s_{1,\ldots,n} p_{1,\ldots,q}]$ where s_i is a bit denoting whether feature F_i is selected or not and p_j is the value for parameter j of the chosen modeling method m. This rather simple definition of solution candidates enables the use of standard concepts for genetic operators for crossover and mutation of bit vectors and real valued vectors: We use uniform, single point, and 2-point crossover operators for binary vectors and bit flip mutation that flips each of the given bits with a given probability. Explanations of these operators can for example be found in [15] and [12].

We have used strict offspring selection [1] which means that individuals are accepted to become members of the next generation if they are evaluated better than both parents.

Standard fitness evaluation as given in Equation 1 has been used during the execution of the evolutionary processes, and classification specific fitness evaluation as given in Equation 6 has been used for selecting the solution candidate eventually returned as the algorithm's result.

3.1.2 Modeling Methods Used in this Research Project

The following techniques for training classifiers have been used in this research project: Linear regression, neural networks, the k-nearest-neighbor method, support vector machines, and genetic programming. All these machine learning methods have been implemented using the HeuristicLab framework[2] [29], a framework for prototyping and analyzing optimization techniques for which both generic concepts of evolutionary algorithms and many functions to evaluate and analyze them are available; we have used these implementations for producing the results summarized in the following section. In this section we give information about these training methods; details about the HeuristicLab implementation of these methods can for example be found in [33].

Linear modeling

Given a data collection including m input features storing the information about N samples, a linear model is defined by the vector of coefficients $\theta_{1...m}$. For calculating the vector of modeled values e using the given input values matrix $u_{1...m}$, these input values are multiplied with the corresponding coefficients and added: $e = u_{1...m} * \theta$. The coefficients vector can be computed by simply applying matrix division. For conducting the test series documented here we have used an implementation of the matrix division function: $\theta = InputValues \backslash TargetValues$. Additionally, a constant additive factor is also included into the model; i.e., a constant offset is added to the coefficients vector. Theoretical background of this approach can be found in [22].

kNN Classification

Unlike other data based modeling methods, k-nearest-neighbor classification [10] works without creating any explicit models. During the training phase, the samples are simply collected; when it comes to classifying a new, unknown sample x_{new}, the sample-wise distance between x_{new} and all other training samples x_{train} is calculated and the classification is done on the basis of those k training samples (x_{NN}) showing the smallest distances from x_{new}.

In the context of classification, the numbers of instances (of the k nearest neighbors) are counted for each given class and the algorithm automatically predicts that class that is represented by the highest number of instances (included in x_{NN}). In the test series documented in this paper we have applied weighting to kNN classification: The distance between x_{new} and x_{NN} is relevant for the classification statement, the weight of "nearer" samples is higher than that of samples that are "further away" from x_{new}.

In this research work we have varied k between 1 and 10.

Artificial Neural Networks

For training artificial neural network (ANN) models, three-layer feed-forward neural networks with one linear output neuron were created using backpropagation; theoretical background and details can for example be found in [24] (Chapter 11, "Neural Networks"). In the tests documented in this paper the number of hidden (sigmoidal) nodes hn has been varied from 5 to 100; we have applied ANN training algorithms that use internal validation sets, i.e., training algorithms use 30% of the given training data as validation data and eventually return those network structures that perform best on these internal validation samples.

Support Vector Machines

Support vector machines (SVMs) are a widely used approach in machine learning based on statistical learning theory [28]. The most important aspect of SVMs is that it is possible to give bounds on the generalization error of the models produced, and to select the corresponding best model from a set of models following the principle of structural risk minimization [28].

In this work we have used the LIBSVM implementation described in [7], which is used in the respective SVM interface implemented for HeuristicLab; here we have used Gaussian radial basis function kernels with varying values for the cost parameters c ($c \in [0, 512]$) and the γ parameter of the SVM's kernel function ($\gamma \in [0, 1]$).

3.2 Genetic Programming

As an alternative to the approach described in the previous sections we have also applied a classification algorithm based on genetic programming (GP) [19] using a structure identification framework described in [32] and [2], in combination with strict offspring selection; this GP approach has been implemented as a part of HeuristicLab.

We have used the following parameter settings for our GP test series: The mutation rate was set to 20%, gender specific parents selection [30] (combining random and roulette selection) was applied as well as strict offspring selection [1] (OS, with success ratio as well as comparison factor set to 1.0). The functions set described in [32] (including arithmetic as well as logical ones) was used for building composite function expressions.

In addition to splitting the given data into training and test data, the GP based training algorithm implemented in HeuristicLab has been designed in such a way that a part of the given training data is not used for training models and serves as validation set; in the end, when it comes to returning classifiers, the algorithm returns those models that perform best on validation data. This approach has been chosen because it is assumed to help to cope with over-fitting; it is also applied in other GP based machine learning algorithms as for example described in [5].

4. MODELING RESULTS

Five-fold cross-validation [17] training / test series have been executed; this means that the available data are separated in five (approximately) equally sized, complementary subsets, and in each training / test cycle one data subset is chosen is used as test and the rest of the data as training samples.

In this section we document test accuracies (μ, σ) for the investigated cancer types; we here summarize test results for modeling cancer diagnoses using tumor markers (TMs) as well as for modeling without using tumor markers. Linear modeling, kNN modeling, ANNs, and SVMs have been applied for identifying estimation models for the selected tumor

[2] http://dev.heuristiclab.com

types, genetic algorithms with strict OS have been applied for optimizing variable selections and modeling parameters; standard fitness calculation as given in (1) has been used by the evolutionary process, the classification specific one as given in (6) has been used for selecting the eventually returned model. The probability of selecting a variable initially was set to 30%. Additionally, we have also applied simple linear regression using all available variables. Finally, genetic programming with strict offspring selection (OSGP) has also been applied.

In all test series the maximum selection pressure [1] was set to 100, i.e., the algorithms were terminated as soon as the selection pressure reached 100. The population size for genetic algorithms optimizing variable selections and modeling parameters was set to 10, for GP the population size was set to 700. In all modeling cases except kNN modeling regression models have been trained, the threshold for classification decisions was in all cases set to 0.5 (since the absence of the specific tumor is represented by 0.0 in the data and its presence by 1.0).

Details about the size of the optimized variable sets as well as optimized modeling parameters are summarized in the appendix of this paper.

Table 2: Modeling results for breast cancer diagnosis

Modeling Method	Using TMs Test accuracies		Not using TMs Test accuracies	
	μ	σ	μ	σ
LR, full features set	79.32%	*1.06*	70.63%	*1.28*
OSGA + LR, $\alpha = 0.0$	81.78%	*0.21*	73.13%	*0.36*
OSGA + LR, $\alpha = 0.1$	81.49%	*1.18*	72.66%	*0.14*
OSGA + LR, $\alpha = 0.2$	81.44%	*0.37*	71.40%	*0.57*
OSGA + kNN, $\alpha = 0.0$	79.21%	*0.78*	74.22%	*2.98*
OSGA + kNN, $\alpha = 0.1$	78.99%	*0.57*	75.55%	*0.87*
OSGA + kNN, $\alpha = 0.2$	78.33%	*1.04*	74.50%	*0.20*
OSGA + ANN, $\alpha = 0.0$	81.41%	*1.14*	75.60%	*2.47*
OSGA + ANN, $\alpha = 0.1$	80.19%	*1.68*	72.38%	*6.08*
OSGA + ANN, $\alpha = 0.2$	79.37%	*1.17*	70.54%	*6.10*
OSGA + SVM, $\alpha = 0.0$	81.23%	*1.10*	73.90%	*2.36*
OSGA + SVM, $\alpha = 0.1$	80.46%	*1.80*	72.19%	*0.94*
OSGA + SVM, $\alpha = 0.2$	77.43%	*3.55*	71.89%	*0.70*
OSGP, $ms = 50$	79.72%	*1.80*	75.32%	*0.45*
OSGP, $ms = 100$	75.50%	*4.95*	71.63%	*2.75*
OSGP, $ms = 150$	79.20%	*6.60*	75.75%	*2.16*

Table 3: Modeling results for melanoma diagnosis

Modeling Method	Using TMs Test accuracies		Not using TMs Test accuracies	
	μ	σ	μ	σ
LR, full features set	73.81%	*3.39*	71.09%	*4.14*
OSGA + LR, $\alpha = 0.0$	72.45%	*4.69*	72.36%	*2.30*
OSGA + LR, $\alpha = 0.1$	74.73%	*2.35*	72.09%	*4.01*
OSGA + LR, $\alpha = 0.2$	73.85%	*2.54*	72.70%	*2.02*
OSGA + kNN, $\alpha = 0.0$	68.77%	*2.38*	71.00%	*1.97*
OSGA + kNN, $\alpha = 0.1$	71.33%	*0.27*	70.21%	*3.41*
OSGA + kNN, $\alpha = 0.2$	67.33%	*0.31*	69.65%	*3.14*
OSGA + ANN, $\alpha = 0.0$	74.78%	*1.63*	69.17%	*2.97*
OSGA + ANN, $\alpha = 0.1$	73.81%	*2.23*	71.82%	*0.61*
OSGA + ANN, $\alpha = 0.2$	74.12%	*1.03*	71.40%	*0.49*
OSGA + SVM, $\alpha = 0.0$	69.72%	*7.57*	68.87%	*4.78*
OSGA + SVM, $\alpha = 0.1$	71.75%	*4.88*	68.22%	*1.88*
OSGA + SVM, $\alpha = 0.2$	61.48%	*3.99*	63.20%	*2.09*
OSGP, $ms = 50$	71.24%	*9.54*	74.89%	*3.66*
OSGP, $ms = 100$	69.91%	*5.20*	65.16%	*13.06*
OSGP, $ms = 150$	71.79%	*4.31*	70.13%	*3.60*

5. CONCLUSION

As documented in the previous section, the investigated diagnoses of breast cancer, melanoma, and respiratory system cancer can be estimated correctly in up to 81%, 74%, and 91% of the analyzed test cases, respectively; without tumor markers up to 75%, 74%, and 88% of the test samples are correctly estimated, respectively. Linear modeling performs well in all modeling tasks, feature selection using genetic algorithms and nonlinear modeling yield even better results for all analyzed modeling tasks. No modeling method performs best for all diagnosis prediction tasks.

Table 4: Modeling results for respiratory system cancer diagnosis

Modeling Method	Using TMs Test accuracies		Not using TMs Test accuracies	
	μ	σ	μ	σ
LR, full features set	91.32%	*0.37*	85.97%	*0.27*
OSGA + LR, $\alpha = 0.0$	91.57%	*0.46*	86.41%	*0.36*
OSGA + LR, $\alpha = 0.1$	91.16%	*1.18*	85.80%	*0.45*
OSGA + LR, $\alpha = 0.2$	89.45%	*0.37*	85.02%	*0.15*
OSGA + kNN, $\alpha = 0.0$	90.98%	*0.84*	87.09%	*0.46*
OSGA + kNN, $\alpha = 0.1$	90.01%	*2.63*	87.01%	*0.83*
OSGA + kNN, $\alpha = 0.2$	90.16%	*0.74*	86.92%	*0.81*
OSGA + ANN, $\alpha = 0.0$	90.28%	*1.63*	85.97%	*4.07*
OSGA + ANN, $\alpha = 0.1$	90.99%	*1.97*	85.82%	*4.52*
OSGA + ANN, $\alpha = 0.2$	88.64%	*1.87*	87.24%	*1.91*
OSGA + SVM, $\alpha = 0.0$	89.03%	*1.38*	83.12%	*3.79*
OSGA + SVM, $\alpha = 0.1$	89.91%	*1.58*	86.25%	*0.79*
OSGA + SVM, $\alpha = 0.2$	88.33%	*1.94*	84.66%	*2.06*
OSGP, $ms = 50$	89.58%	*2.75*	85.98%	*5.74*
OSGP, $ms = 100$	90.44%	*3.02*	86.54%	*6.02*
OSGP, $ms = 150$	89.58%	*3.75*	87.97%	*5.57*

Further research shall focus on the practical application of the here presented research results in the treatment of patients, and the authors also plan to analyze in how far separately estimated tumor markers (as discussed in [33] and [34]) can help improve cancer diagnosis predictions without having to use original tumor marker values.

6. ACKNOWLEDGMENTS

The work described in this paper was done within the Josef Ressel Centre for Heuristic Optimization (*Heureka!*) sponsored by the Austrian Research Promotion Agency (FFG).

7. REFERENCES

[1] M. Affenzeller and S. Wagner. SASEGASA: A new generic parallel evolutionary algorithm for achieving highest quality results. *Journal of Heuristics - Special Issue on New Advances on Parallel Meta-Heuristics for Complex Problems*, 10:239–263, 2004.

[2] M. Affenzeller, S. Winkler, S. Wagner, and A. Beham. *Genetic Algorithms and Genetic Programming - Modern Concepts and Practical Applications.* Chapman & Hall / CRC, 2009.

[3] E. Alba, J. G.-N. L. Jourdan, and E.-G. Talbi. Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms. *IEEE Congress on Evolutionary Computation 2007*, pages 284 – 290, 2007.

[4] G. L. Andriole, E. D. Crawford, R. L. Grubband, S. S. Buys, D. Chia, T. R. Church, et al. Mortality results from a randomized prostate-cancer screening trial. *New England Journal of Medicine*, 360(13):1310–1319, 2009.

[5] W. Banzhaf and C. Lasarczyk. Genetic programming of an algorithmic chemistry. In U. O'Reilly, T. Yu, R. Riolo, and B. Worzel, editors, *Genetic Programming Theory and Practice II*, pages 175–190. Ann Arbor, 2004.

[6] N. Bitterlich and J. Schneider. Cut-off-independent tumour marker evaluation using ROC approximation. *Anticancer Research*, 27:4305–4310, 2007.

[7] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for*

support vector machines, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[8] N. Clegg, C. Ferguson, L. True, H. Arnold, A. Moorman, J. Quinn, R. Vessella, and P. Nelson. Molecular characterization of prostatic small-cell neuroendocrine carcinoma. *Prostate*, 55(1):55–64, 2003.

[9] G. Crombach, H. Würz, F. Herrmann, R. Kreienberg, V. Möbus, P. Schmidt-Rhode, G. Sturm, H. Caffier, and H. Kaesemann. The importance of the scc antigen in the diagnosis and follow-up of cervix carcinoma. *Deutsche Medizinische Wochenschrift*, 114(18):700–705, 1989.

[10] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley Interscience, 2^{nd} edition, 2000.

[11] M. J. Duffy and J. Crown. A personalized approach to cancer treatment: how biomarkers can help. *Clinical Chemistry*, 54(11):1770–1779, 2008.

[12] A. Eiben and J. Smith. *Introduction to Evolutionary Computation*. Natural Computing Series. Springer-Verlag Berlin Heidelberg, 2003.

[13] B. Frey, R. Morant, H. Senn, and W. Riesen. Clinical assessment of the new tumor marker tps. *International Journal for Cancer Research and Treatment*, 17:270–276, 1994.

[14] P. Gold and S. O. Freedman. Demonstration of tumor-specific antigens in human colonic carcinomata by immunological tolerance and absorption techniques. *The Journal of Experimental Medicine*, 121:439–462, 1965.

[15] J. H. Holland. *Adaption in Natural and Artifical Systems*. University of Michigan Press, 1975.

[16] J. A. Koepke. Molecular marker test standardization. *Cancer*, 69:1578–1581, 1992.

[17] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th international joint conference on Artificial intelligence*, volume 2, pages 1137–1143. Morgan Kaufmann, 1995.

[18] H. Koprowski, M. Herlyn, Z. Steplewski, and H. Sears. Specific antigen in serum of patients with colon carcinoma. *Science*, 212(4490):53–55, 1981.

[19] J. R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. The MIT Press, 1992.

[20] M. LaFleur-Brooks. *Exploring Medical Language: A Student-Directed Approach*. St. Louis, Missouri, USA: Mosby Elsevier, 7^{th} edition, 2008.

[21] R. S. Lai, C. C. Chen, P. C. Lee, and J. Y. Lu. Evaluation of cytokeratin 19 fragment (cyfra 21-1) as a tumor marker in malignant pleural effusion. *Japanese Journal of Clinical Oncology*, 29(9):421–424, 199.

[22] L. Ljung. *System Identification – Theory For the User, 2nd edition*. PTR Prentice Hall, Upper Saddle River, N.J., 1999.

[23] G. J. Mizejewski. Alpha-fetoprotein structure and function: relevance to isoforms, epitopes, and conformational variants. *Experimental biology and medicine*, 226(5):377–408, 2001.

[24] O. Nelles. *Nonlinear System Identification*. Springer Verlag, Berlin Heidelberg New York, 2001.

[25] Y. Niv. Muc1 and colorectal cancer pathophysiology considerations. *World Journal of Gastroenterology*, 14(14):2139–2141, 2008.

[26] D. Nonaka, L. Chiriboga, and B. Rubin. Differential expression of s100 protein subtypes in malignant melanoma, and benign and malignant peripheral nerve sheath tumors. *Journal of Cutaneous Pathology*, 35(11):1014–1019, 2008.

[27] A. J. Rai, Z. Zhang, J. Rosenzweig, I. ming Shih, T. Pham, E. T. Fung, L. J. Sokoll, and D. W. Chan. Proteomic approaches to tumor marker discovery. *Archives of Pathology & Laboratory Medicine*, 126(12):1518–1526, 2002.

[28] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[29] S. Wagner. *Heuristic Optimization Software Systems – Modeling of Heuristic Optimization Algorithms in the HeuristicLab Software Environment*. PhD thesis, Johannes Kepler University Linz, 2009.

[30] S. Wagner and M. Affenzeller. SexualGA: Gender-specific selection for genetic algorithms. In N. Callaos, W. Lesso, and E. Hansen, editors, *Proceedings of the 9^{th} World Multi-Conference on Systemics, Cybernetics and Informatics (WMSCI) 2005*, volume 4, pages 76–81. International Institute of Informatics and Systemics, 2005.

[31] P. W. Williams and H. D. Gray. *Gray's anatomy*. New York: C. Livingstone, 37^{th} edition, 1989.

[32] S. Winkler. *Evolutionary System Identification - Modern Concepts and Practical Applications*. PhD thesis, Institute for Formal Models and Verification, Johannes Kepler University Linz, 2008.

[33] S. Winkler, M. Affenzeller, W. Jacak, and H. Stekel. Classification of tumor marker values using heuristic data mining methods. In *Proceedings of the GECCO 2010 Workshop on Medical Applications of Genetic and Evolutionary Computation (MedGEC 2010)*, 2010.

[34] S. Winkler, M. Affenzeller, G. Kronberger, M. Kommenda, S. Wagner, W. Jacak, and H. Stekel. Feature selection in the analysis of tumor marker data using evolutionary algorithms. In *Proceedings of the 7th International Mediterranean and Latin American Modelling Multiconference*, pages 1 – 6, 2010.

[35] B. W. Yin, A. Dnistrian, and K. O. Lloyd. Ovarian cancer antigen CA125 is encoded by the MUC16 mucin gene. *International Journal of Cancer*, 98(5):737–40, 2002.

[36] K. Yonemori, M. Ando, T. S. Taro, N. Katsumata, K. Matsumoto, Y. Yamanaka, T. Kouno, C. Shimizu, and Y. Fujiwara. Tumor-marker analysis and verification of prognostic models in patients with cancer of unknown primary, receiving platinum-based combination chemotherapy. *Journal of Cancer Research and Clinical Oncology*, 132(10):635–642, 2006.

[37] L. Zhong, X. Zhou, K. Wei, X. Yang, C. Ma, C. Zhang, and Z. Zhang. Application of serum tumor markers and support vector machine in the diagnosis of oral squamous cell carcinoma. *Shanghai Kou Qiang Yi Xue (Shanghai Journal of Stomatology)*, 17(5):457–460, 2008.

Appendix

In this appendix we summarize results of the executing modeling test series:

In Table 5 we summarize the effort of the modeling approaches applied in this research work: For the combination of GAs and machine learning methods we document the number of modeling executions, and for GP we give the number of evaluated solutions (i.e., models).

For the combination of genetic algorithms with linear regression, kNN modeling, ANNs, and SVMs (with varying variable ratio (vr) weighting factors) as well as GP with varying maximum tree sizes ms we give the sizes of selected variable sets, and (where applicable) also k, hn, c, and γ. Obviously there are different variations in the parameters identified as optimal by the evolutionary process: The numbers of variables used as well as the neural networks' hidden nodes vary to a relatively small extent, e.g., whereas especially the SVMs' parameters (especially the c factors) vary very strongly.

Table 5: Effort in terms of executed modeling runs and evaluated model structures

Modeling Method	Modeling executions					
	$vrfw = 0.0$		$vrfw = 0.1$		$vrfw = 0.2$	
	μ	σ	μ	σ	μ	σ
LR	3260.4	717.8	2339.6	222.6	2465.2	459.2
kNN	2955.3	791.8	3046.0	362.4	3791.3	775.9
ANN	3734.0	855.9	3305.0	582.6	3297.0	475.9
SVM	2950.0	794.8	2846.0	391.4	3496.7	859.8

Modeling Method	Evaluated solutions (models)		
	$ms = 50$	$ms = 100$	$ms = 150$
	μ, σ	μ, σ	μ, σ
OSGP	1483865.0, 674026.2	1999913.3, 198289.1	2238496.7, 410123.6

Table 6: Optimized parameters for linear regression

Problem Instance, vr weighting		Variables used	
		μ	σ
BC, TM	$\alpha = 0.0$	16.6	2.10
	$\alpha = 0.1$	11.8	1.50
	$\alpha = 0.2$	6.4	0.60
BC, no TM	$\alpha = 0.0$	9.6	1.15
	$\alpha = 0.1$	8.8	0.58
	$\alpha = 0.2$	6.4	1.20
Mel, TM	$\alpha = 0.0$	16.6	0.55
	$\alpha = 0.1$	12.2	0.84
	$\alpha = 0.2$	9.2	4.09
Mel, no TM	$\alpha = 0.0$	10.8	1.79
	$\alpha = 0.1$	8.8	2.28
	$\alpha = 0.2$	8.2	1.92
RSC, TM	$\alpha = 0.0$	17.2	2.95
	$\alpha = 0.1$	13.4	2.51
	$\alpha = 0.2$	9.0	2.55
RSC, no TM	$\alpha = 0.0$	16.0	4.64
	$\alpha = 0.1$	9.6	0.89
	$\alpha = 0.2$	8.6	3.21

Table 7: Optimized parameters for kNN modeling

Problem Instance, vr weighting		Variables used		k	
		μ	σ	μ	σ
BC, TM	$\alpha = 0.0$	18.2	2.20	9.8	2.10
	$\alpha = 0.1$	14.0	3.60	12.6	4.60
	$\alpha = 0.2$	11.0	1.80	11.2	3.00
BC, no TM	$\alpha = 0.0$	14.4	1.67	11.2	1.64
	$\alpha = 0.1$	14.0	2.45	13.8	3.11
	$\alpha = 0.2$	11.8	0.84	18.8	1.10
Mel, TM	$\alpha = 0.0$	15.6	1.82	17.8	2.86
	$\alpha = 0.1$	16.4	1.34	14.4	5.90
	$\alpha = 0.2$	13.6	1.67	19.4	1.34
Mel, no TM	$\alpha = 0.0$	15.0	1.58	14.2	1.10
	$\alpha = 0.1$	10.4	1.52	18.2	1.64
	$\alpha = 0.2$	9.6	1.14	16.8	2.05
RSC, TM	$\alpha = 0.0$	14.6	1.67	20.0	0.00
	$\alpha = 0.1$	13.6	1.67	16.8	6.06
	$\alpha = 0.2$	10.4	1.52	12.8	3.90
RSC, no TM	$\alpha = 0.0$	15.6	2.79	15.2	1.64
	$\alpha = 0.1$	12.2	1.10	10.6	1.82
	$\alpha = 0.2$	10.2	2.95	13.2	2.95

Table 8: Optimized parameters for ANNs

Problem Instance, vr weighting		Variables used		hn	
		μ	σ	μ	σ
BC, TM	$\alpha = 0.0$	17.0	1.20	75.6	20.80
	$\alpha = 0.1$	14.8	3.40	51.0	5.80
	$\alpha = 0.2$	11.0	0.80	35.8	13.00
BC, no TM	$\alpha = 0.0$	12.6	1.41	82.4	23.46
	$\alpha = 0.1$	12.2	0.89	70.8	26.40
	$\alpha = 0.2$	11.2	1.10	68.2	14.58
Mel, TM	$\alpha = 0.0$	19.6	2.19	56.8	13.31
	$\alpha = 0.1$	12.8	2.28	61.0	2.55
	$\alpha = 0.2$	15.6	5.18	51.2	6.98
Mel, no TM	$\alpha = 0.0$	15.4	2.51	68.6	14.24
	$\alpha = 0.1$	8.2	1.64	59.8	3.83
	$\alpha = 0.2$	8.0	1.00	58.6	5.81
RSC, TM	$\alpha = 0.0$	13.4	3.44	64.6	10.97
	$\alpha = 0.1$	11.2	2.28	68.2	6.69
	$\alpha = 0.2$	8.2	1.64	60.2	13.92
RSC, no TM	$\alpha = 0.0$	13.2	2.28	71.2	12.38
	$\alpha = 0.1$	12.2	2.05	70.6	12.99
	$\alpha = 0.2$	11.6	2.19	64.4	14.24

Table 9: Optimized parameters for SVMs

Problem Instance, vr weighting		Variables used	C	γ
		μ , σ	μ, σ	μ, σ
BC, TM	$\alpha = 0.0$	21.6, 3.50	101.50, 92.30	0.05, 0.06
	$\alpha = 0.1$	18.8, 3.50	12.44, 13.85	0.09, 0.01
	$\alpha = 0.2$	16.0, 2.00	64.79, 67.59	0.04, 0.01
BC, no TM	$\alpha = 0.0$	15.6, 1.83	47.16, 12.63	0.05, 0.05
	$\alpha = 0.1$	15.4, 1.10	22.50, 25.88	0.07, 0.04
	$\alpha = 0.2$	13.0, 2.65	8.14, 10.09	0.07, 0.04
Mel, TM	$\alpha = 0.0$	13.0, 4.53	166.23, 236.61	0.27, 0.25
	$\alpha = 0.1$	10.8, 3.42	204.74, 210.43	0.18, 0.19
	$\alpha = 0.2$	4.2, 2.95	123.08, 44.14	0.26, 0.20
Mel, no TM	$\alpha = 0.0$	21.4, 6.95	116.21, 196.73	0.41, 0.30
	$\alpha = 0.1$	19.8, 1.64	492.73, 8.10	0.48, 0.41
	$\alpha = 0.2$	14.4, 3.29	310.17, 208.60	0.36, 0.35
RSC, TM	$\alpha = 0.0$	21.2, 8.50	183.54, 95.38	0.27, 0.26
	$\alpha = 0.1$	14.6, 1.14	74.56, 67.98	0.09, 0.10
	$\alpha = 0.2$	11.2, 3.83	37.55, 68.31	0.45, 0.35
RSC, no TM	$\alpha = 0.0$	13.4, 4.10	23.14, 31.91	0.35, 0.25
	$\alpha = 0.1$	12.4, 3.21	144.73, 96.79	0.19, 0.08
	$\alpha = 0.2$	12.4, 3.21	376.66, 206.84	0.09, 0.10

Table 10: Number of variables used by models returned by OSGP

Problem Instance, maximum tree size ms		Variables used by returned model	
		μ	σ
BC, TM	$ms = 50$	9.0	2.74
	$ms = 100$	9.6	1.34
	$ms = 150$	17.8	0.45
BC, no TM	$ms = 50$	10.5	0.71
	$ms = 100$	10.0	1.41
	$ms = 150$	11.5	0.71
Mel, TM	$ms = 50$	10.2	2.05
	$ms = 100$	10.0	2.55
	$ms = 150$	12.0	2.00
Mel, no TM	$ms = 50$	8.0	1.58
	$ms = 100$	8.8	0.84
	$ms = 150$	11.4	3.36
RSC, TM	$ms = 50$	7.8	2.05
	$ms = 100$	12.0	2.35
	$ms = 150$	12.0	1.22
RSC, no TM	$ms = 50$	9.4	3.91
	$ms = 100$	12.2	2.17
	$ms = 150$	13.6	3.13

Table 11: List of patient data variables collected at AKH Linz in the years 2005 – 2008: Blood parameters, general patient information, and tumor markers

Parameter	*Description*	*Unit*	*Plausible Range*
ALT	Alanine transaminase, a transaminase enzyme; also called glutamic pyruvic transaminase (GPT).	U/l	[1; 225]
AST	Aspartate transaminase, an enzyme also called glutamic oxaloacetic transaminase (GOT).	U/l	[1; 175]
BSG1	Erythrocyte sedimentation rate; the rate at which red blood cells settle / precipitate within one hour.	mm	[0; 50]
BUN	Blood urea nitrogen; measures the amount of nitrogen in the blood (caused by urea).	mg/dl	[1; 150]
CBAA	Basophil granulocytes; type of leukocytes.	G/l	[0.0; 0.2]
CEOA	Eosinophil granulocytes; type of leukocytes.	G/l	[0.0; 0.4]
CH37	Cholinesterase, an enzyme.	kU/l	[2; 23]
CHOL	Cholesterol, a structural component of cell membranes.	mg/dl	[40; 550]
CLYA	Lymphocytes; type of leukocytes.	G/l	[1; 4]
CMOA	Monocytes; type of leukocytes.	G/l	[0.2; 0.8]
CNEA	Neutrophils; most abundant type of leukocytes.	G/l	[1.8; 7.7]
CRP	C-reactive protein, a protein; inflammations cause the rise of CRP.	mg/dl	[0; 20]
FE	Iron.	ug/dl	[30; 210]
FER	Ferritin, a protein that stores and transports iron in a safe form.	ng/ml	[10; 550]
GT37	γ-glutamyltransferase, an enzyme.	U/l	[1; 290]
HB	Hemoglobin, a protein that contains iron and transports oxygen.	g/dl	[6; 18]
HDL	High-density lipoprotein; this protein enables the transport of lipids with blood.	mg/dl	[25; 120]
HKT	Hematocrit; the packed cell volume, i.e., the proportion of red blood cells within the blood.	%	[25; 65]
HS	Uric acid, also called urate.	mg/dl	[1; 12]
KREA	Creatinine, a chemical by-product produced in muscles.	mg/dl	[0.2; 5.0]
LD37	Lactate dehydrogenase (LDH), an enzyme that can be used as a marker of injuries to cells.	U/l	[5; 744]
MCV	Mean corpuscular / cell volume; the average size (i.e., volume) of red blood cells.	fl	[69; 115]
PLT	Thrombocytes, also called platelets, are irregularly-shaped cells that do not have a nucleus.	G/l	[25; 1,000]
RBC	Erythrocytes, red blood cells that transport and deliver oxygen.	T/l	[2.2; 8.0]
TBIL	Bilirubin, the yellow product of the heme catabolism.	mg/dl	[0; 5]
TF	Transferrin, a protein, delivers iron.	mg/dl	[100; 500]
WBC	Leukocytes, also called white blood cells (WBCs); cells that help the body fight infections or foreign materials.	G/l	[1.5; 50]
AGE	The patient's age.	years	[0; 120]
SEX	The patient's sex.	f/m	{f, m}
AFP	Alpha-fetoprotein ([23]) is a protein found in the blood plasma; during fetal life it is produced by the yolk sac and the liver. AFP is also often measured and used as a marker for a set of tumors, especially endodermal sinus tumors (yolk sac carcinoma), neuroblastoma, hepatocellular, carcinoma and germ cell tumors [11].	IU/ml	[0.0; 90.0]
CA 125	Cancer antigen 125 (CA 125) ([35]), also called carbohydrate antigen 125 or mucin 16 (MUC16), is a protein that is often used as a tumor marker that may be elevated in the presence of specific types of cancers.	U/ml	[0.0; 150]
CA 15-3	Mucin 1 (MUC1), also known as cancer antigen 15-3 (CA 15-3), is a protein used as a tumor marker in the context of monitoring certain cancers [25], especially breast cancer.	U/ml	[0.0; 100.0]
CA 19-9	CA 19-9 is a tumor marker often used to monitor monitor a person's response to cancer treatment and/or cancer progression, for example colon cancer and pancreatic cancer [18].	U/m	[0.0; 120.0]
CEA	Carcinoembryonic antigen (CEA; [14]) is a protein that is in humans normally produced during fetal development. When used as a tumor marker, CEA is mainly used to identify recurrences of cancer after surgical resections.	ng/ml	[0.0; 50.0]
CYFRA	Fragments of cytokeratin 19, a protein found in the cytoskeleton, are found in many places of the human body; especially in the lung and in malign lung tumors high concentrations of these fragments, which are also called CYFRA 21-1, are found [21].	ng/ml	[0.0; 10.0]
fPSA	The free-to-total ratio of the prostate-specific antigen (PSA) is calculated and stored in fPSA.	ratio	[0.0; 1.0]
NSE	The neuron-specific enolase (NSE) is an enzyme frequently used as tumor marker for lung cancer because it can be help to identify neuronal cells and cells with neuroendocrine differentiation. [8]	ng/ml	[0.0; 100.0]
PSA	Prostate-specific antigen (PSA; [4]) is a protein produced in the prostate gland; PSA blood tests are widely considered the most effective test currently available for the early detection of prostate cancer since PSA is often elevated in the presence of prostate disorders.	ng/ml	[0.0; 20.0]
S-100	S-100 is a family of proteins found in vertebrates; members of the S-100 protein family are useful as markers for certain tumors. S-100 values can be found in melanoma and are used as as cell markers for anatomic pathology and also markers for inflammatory diseases. [26]	ug/l	[0.0; 1.2]
SCC	The squamous cell carcinoma antigen (SCC) is used as tumor marker for the diagnosis and follow-up control of epithelial carcinoma [9].	ng/ml	[0.0; 20.0]
TPS	TPS (tissue polypeptide specific antigen) is used as tumor marker indicating cellular proliferation; details about this tumor marker can for example be found in [13].	U/l	[0.0; 300.0]