Evolutionary Identificatio of Cancer Predictors Using Clustered Data - A Case Study for Breast Cancer, Melanoma, and Cancer in the Respiratory System

Stephan M. Winkler
Upper Austria University of Applied Sciences
Heuristic and Evolutionary Algorithms Laboratory
Softwarepark 11
4232 Hagenberg, Austria
stephan.winkler@fh-hagenberg.at

Michael Affenzeller
Upper Austria University of Applied Sciences
Heuristic and Evolutionary Algorithms Laboratory
Softwarepark 11
4232 Hagenberg, Austria
michael.affenzeller@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 discuss the effects of using pre-clustered data on the 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, the goal is to identify mathematical models for estimating cancer diagnoses. We have applied a hybrid clustering and classification approach that first identifies data clusters (using standard patient data and tumor markers) and then learns prediction models on the basis of these data clusters.

In the empirical section we analyze the clusters of patient data samples formed using k-means clustering: The optimal number of clusters is identified, and we investigate the homogeneity of these clusters. Several evolutionary modeling approaches implemented in HeuristicLab have been applied for subsequently 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. As we show in the results section, the investigated diagnoses of breast cancer, melanoma, and respiratory system cancer can be estimated correctly in up to 84.2%, 80.3%, and 94.1% of the analyzed test cases, respectively; without tumor markers up to 78.2%, 78%, and 93.3% of the test samples are correctly estimated, respectively.

Keywords

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

GECCO'13 Companion, July 6–10, 2013, Amsterdam, The Netherlands.

General Terms

Algorithms, Reliability, Experimentation, Standardization

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

1. INTRODUCTION AND OVERVIEW

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: 28 routinely measured blood values 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). This paper describes research that is a continuation of the results presented at further GECCO Workshops on Medical Applications of Genetic and Evolutionary Computation: In [25] we reported on the data based identification of mathematical models for tumor markers (i.e., virtual tumor markers), and in [26] we discussed the use of several evolutionary machine learning techniques for identifying predictors for cancer diagnoses.

In this paper we discuss the use of an approach presented in [28], namely the integrated use of automated clustering and classification algorithms for identifying even more accurate classifiers for cancer diagnoses.

In the following section (Section 2) we describe the database we have used for our research work and also the tumors 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.

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

In Section 3 we describe our integrated clustering and classification approach. In Section 3.2 we describe the here applied clustering approach and define functions to estimate the homogeneity of the so formed clusters; in Section 3.3 we discuss the classification methods used in this research project as well as the parameter settings applied.

In Section 4 we summarize and analyze the modeling results we have achieved; the conclusion of this paper is given in Section 5.

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 1; 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 and the applied preprocessing methods can be found in [25] and [26].

2.1.1 Standard Parameters

Information about the standard parameters (general patient information and standard blood values) stored in the AKH database (which are listed in the upper part of Table 1) can be found in [13], [23], and [25], e.g.

2.1.2 Tumor Markers

In general, tumor markers (TMs) 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 to detect the presence of cancer. 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 [10] (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), [19], [29], [5], and [30].

Information about the tumor markers stored in the AKH database are listed in the lower part of Table 1.

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).

Table 1: List of patient data variables collected at AKH Linz: Blood parameters, general patient information, and tumor markers

Standard patient information and blood values
ALT, AST, BSG1, BUN, CBAA, CEOA, CH37, CHOL, CLYA, CMOA, CNEA, CRP, FE, FER, GT37, HB, HDL, HKT, HS, KREA, LD37, MCV, PLT, RBC, TBIL, TF, WBC, AGE, SEX
Tumor markers
AFP, CA 125, CA 15-3, CA 19-9, CEA, CYFRA, fPSA, NSE, PSA, S-100, SCC, TPS

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 is divided by the difference between min_i and the maximum plausible value $maxplau_i$; all values greater than $maxplau_i$ 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.
- 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 interval 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 2 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.

Table 2: Overview of the data sets compiled for selected cancer types

Cancer Type	*Input Variables*	*Total Samples*	*Samples in Class 0*	*Samples in Class 1*	Missing Values
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%

3. METHODS

3.1 An Integrated Clustering and Classification Approach for the Analysis of Medical Data

The here applied analysis approach integrates clustering and classification algorithms:

First, the available patient data are clustered; this clustering is done on the one hand only for standard blood data and on the other hand for standard data plus tumor markers. The so identified clusters of samples are analyzed and compared with each other; we especially analyze the size of the clusters and to which extent samples which are assigned the same clusters regarding standard data are also assigned to the same clusters on the basis of standard and tumor marker data.

In this research project we apply k-means clustering [18], [16]. As simpler models are to be preferred over more complex ones, the quality of clusterings is calculated considering not only their quantization error, but also the number of clusters formed; the Davies-Bouldin index [7] is used in this context.

The so clustered data are subsequently (in combination with tumor diagnosis data) used for learning tumor diagnosis predictors; each cluster is used individually for training these models.

The so identified models are analyzed and compared to each other with respect to their structure and their relevant variables.

Figure 1 graphically summarizes this integrated clustering and classification approach.

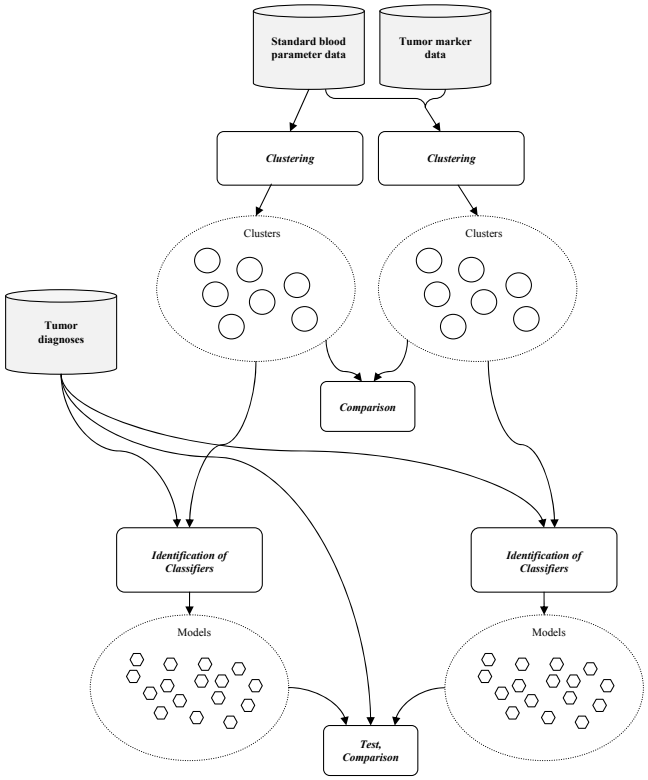

Figure 1: An integrated clustering and classification approach for the analysis of medical data: Data clusters are formed using standard data and optionally also tumor marker data; these clusters are the basis for the identification of classifiers that can be used as predictors for cancer diagnoses.

3.2 Clustering

For clustering the available data we have used the k-means algorithm [18], [16] with varying numbers of clusters k: The cluster centers are initially set at random and then iteratively adapted until the quantization error is minimized; each sample is assigned to the cluster whose center has the minimum distance to the sample (distance is here calculated using the Euclidean distance function). As on the one hand the optimal number clusters is unknown and different values for k have to be tried, and on the other hand simpler models are to be preferred over more complex ones, the quality of clusterings is calculated considering not only their quantization error, but also the number of clusters formed; the Davies-Bouldin index [7] is used in this context. Information about the samples' classification (as diseased or not diseased) is of course not available for the clustering algorithm.

The mean quantization error (MQE) of $cluster_i$ is defined as the average distance of its samples to its center ce_i, and the Davies-Bouldin Index (DBI) for a complete clustering hypothesis takes into account the compactness of the formed clusters (via their MQE) as well as their distance:

$$MQE_i = \frac{\sum_{s_j \in cluster_i} dist(s_j, ce_i)}{|cluster_i|} \quad (1)$$

$$DBI = \frac{1}{k} \cdot \sum_i (max_{j,i \neq j} \frac{MQE_i + MQE_j}{dist(ce_i, ce_j)}) \quad (2)$$

We assume that optimal clustering minimizes the DBI, i.e. we will eventually use that number of clusters k that leads to minimal DBI-values.

Additionally, we also analyze how well this unsupervised clustering approach solves the original classification task by calculating the homogeneity of $cluster_j$ as the ratio r of the samples of the most prominent class in the cluster:

$$r(class_i, cluster_j) = \frac{|s{:}class(s)=class_i \wedge s \in cluster_j|}{|cluster_j|} \quad (3)$$

$$homogeneity(cluster_j) = \max_i (r(class_i, cluster_j)) \quad (4)$$

As we are in the total homogeneity of a whole clustering (i.e., a set of clusters formed for a given data collection), we calculate $homogeneity_{total}$ as the weighted average of all homogeneities:

$$homogeneity_{total}(clusters) = \frac{1}{n} \cdot \sum_{c \in clusters} (homogeneity(c) \cdot |c|) \quad (5)$$

where n is the total number of samples.

3.3 Identificatio of Classifier

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 3.3.1), on the other hand we apply genetic programming (as described in Section 3.3.2).

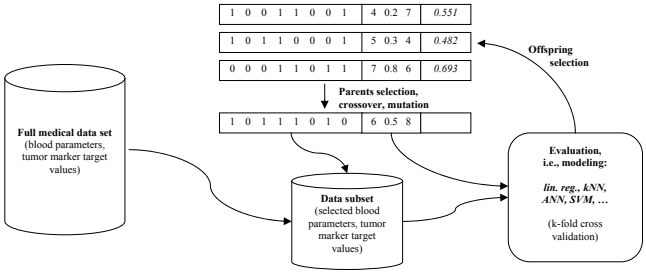

Figure 2: A hybrid evolutionary algorithm for feature selection and parameter optimization in data based modeling.

3.3.1 Hybrid Modeling Using Machine Learning Algorithms and Evolutionary Algorithms for Parameter Optimization and Feature Selection

Given a set of n features $F = \{f_1, f_2, \ldots, f_n\}$, the goal in feature selection 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. We have therefore used an evolutionary algorithm to optimize feature selections as well as parameter settings for various modeling methods:

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

$$fitness(F', m, p) = \alpha * |F'|/|F| + (1-\alpha) * (1 - R^2(est(F', m, p), orig)). \quad (6)$$

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 [24], 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

$$ca(F', m, p) = \frac{|\{j : estCl(F', m, p)[j] = origCl[j]\}|}{|estCl|} \quad (7)$$

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

$$ca(F', m, p)_c = \frac{|\{j : estCl(F', m, p)[j] = origCl[j] = c\}|}{|\{j : origCl[j] = c\}|} \quad (8)$$

$$cwca(F', m, p) = \frac{\sum_{c \in C} ca(F', m, p)_c}{|C|} \quad (9)$$

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

$$fitness_{cwca}(F', m, p) = \alpha * |F'|/|F| + (1-\alpha) * (1 - cwca(F', m, p)). \quad (10)$$

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 [27] we have analyzed the sets of features identified as relevant for modeling 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 2. 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 [9].

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; i.e., the success ratio as well as the comparison factor were set to 1.0.

Standard fitness evaluation as given in Equation 6 has been used during the execution of the evolutionary processes, and classification specific fitness evaluation as given

in Equation 10 has been used for selecting the solution candidate eventually returned as the algorithm's result.

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] [21], 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 [25].

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. Theoretical background of this approach can be found in [15].

k-Nearest-Neighbor Classification

Unlike other data based modeling methods, k-nearest-neighbor classification [8] (kNN) 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 [17]. 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 [20]. 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 [20].

In this work we have used the LIBSVM implementation described in [6], 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 ($\gamma \in [0, 1]$).

3.3.2 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) [12] using a structure identification framework described in [24] and [2], in combination with strict offspring selection; this GP approach has been implemented in HeuristicLab.

We have used the following parameter settings for our GP test series: The mutation rate was set to 20%, gender specific parents selection [22] (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 [24] (including arithmetic as well as logical ones) was used for building composite function expressions.

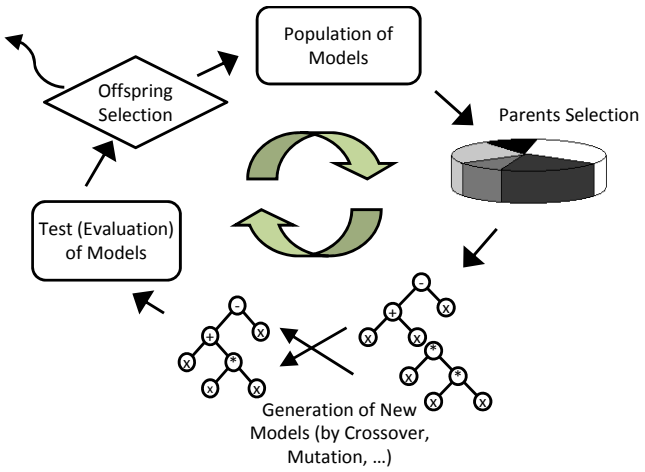

Figure 3: The genetic programming cycle [14] including offspring selection [1].

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 [4].

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

4. EMPIRICAL RESULTS

We have used the data described the data described in Section 2; these data partitions are the same as those used in previous research described in [26].

4.1 Analysis of Identifie Data Clusters

The optimal number of clusters for the given data described in Section 2 was identified in the following way: Different values for the number of clusters k (listed in Table 3) have been applied for clustering each data set; each data partition was clustered five times independently using all variables as well as using all variables except tumor markers. For each clustering we calculated the DB index (as defined in Equation 2) and in Table 3 we list the resulting average values. As we see in the table below and also in Figure 4, setting the number of clusters to 25 seems to be the best decision.

Table 3: Average Davies-Bouldin index values for clusterings of the given data.

k	Davies-Bouldin index ($\mu \pm \sigma$)
3	1.95 ± *0.17*
5	2.10 ± *0.20*
10	1.99 ± *0.18*
15	1.93 ± *0.28*
20	1.76 ± *0.30*
25	1.70 ± *0.35*
30	1.74 ± *0.34*
35	1.75 ± *0.33*
40	1.74 ± *0.34*

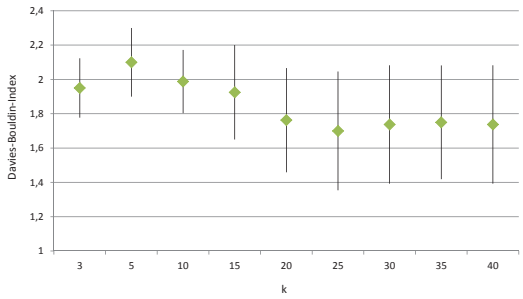

Figure 4: Average Davies-Bouldin index values for clusterings of the given data.

Using this optimal number of clusters ($k = 25$) we clustered the given data partitions an analyzed the quality of the so achieved clustering with respect to their class homogeneity (using Equation 5). The results of this analysis are summarized in Table 4 and shown in Figure 5:

- Breast cancer data can be clustered with approximately 77.9% average homogeneity (with respect to class values) if all variables are used; using no tumor markers or only using tumor markers decreases this value and thus decreases the clustering quality.
- Melanoma data can be clustered with approximately 76.5% average homogeneity using all variables; using only standard values or only tumor markers leads to worse clustering results.
- Respiratory system cancer (RSC) data can be clustered with almost 90% average homogeneity; if tumor markers are omitted, the average class homogeneity in the formed clusters is approximately 85%.

Table 4: Cluster homogeneity with respect to sample classifications for the analyzed data partitions.

Clustering task	Cluster homogeneity (weighted average, $\mu \pm \sigma$)
BC data, all variables	77.878 ± *0.882*
BC data, no tumor markers	74.704 ± *1.896*
BC data, only tumor markers	77.552 ± *1.114*
Mel data, all variables	76.510 ± *1.009*
Mel data, no tumor markers	74.720 ± *0.521*
Mel data, only tumor markers	73.876 ± *1.063*
RSC data, all variables	88.598 ± *0.770*
RSC data, no tumor markers	84.988 ± *0.882*
RSC data, only tumor markers	89.646 ± *1.113*

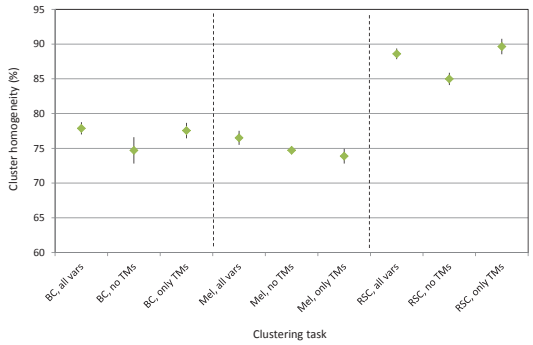

Figure 5: Cluster homogeneity with respect to sample classifications for the analyzed data partitions.

4.2 Identificatio of Cancer Predictors Using Clustered Data

Finally, using the clusters identified as described previously we have performed machine learning in order to learn classifiers for the given samples. All clusters were used separately, i.e., each cluster was used for training classification models. Five-fold cross-validation [11] 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 order to avoid overfitting, all clusters with less than 50 samples were (for each clustering separately) combined into "rest" clusters.

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. The test accuracy is calculated as the ratio of test samples that were classified correctly; as the clusters are inequally sized, we have calculated the test accuracy for each cluster separately, weighted each so resulting classification rate with the size of the cluster and so retrieve the total classification accuracy.

Linear modeling, kNN modeling, ANNs, and SVMs have been applied for identifying estimation models for the selected tumor types, genetic algorithms with strict OS have

been applied for optimizing variable selections and modeling parameters; standard fitness calculation as given in Equation 6 has been used by the evolutionary process, the classification specific one as given in Equation 10 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 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 and the maximum tree size (ms) to 100. 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).

The so achieved results are summarized in the following Tables 5 – 7 and in Figure 6.

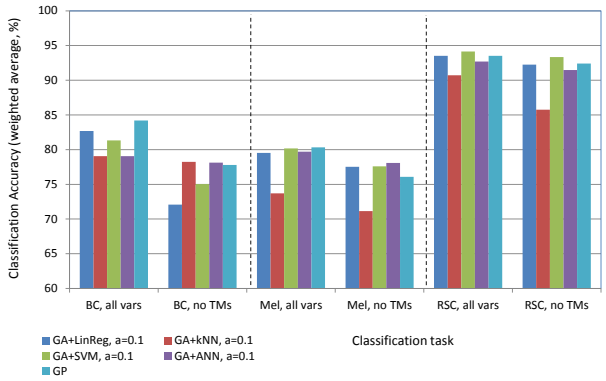

Figure 6: Classification results overview.

5. DISCUSSION AND CONCLUSION

As we clearly see in the classification results section, the here applied approach of using pre-clustered data and evolutionary modeling techniques leads to better results than those reported in previous test series [26]:

Table 5: Classification results for breast cancer diagnosis

Results using all variables	
Modeling method	Test accuracies ($\mu \pm \sigma$)
OSGA + LR, $\alpha = 0.1$	82.690% ± *3.64*
OSGA + kNN, $\alpha = 0.1$	79.069% ± *3.11*
OSGA + ANN, $\alpha = 0.1$	81.319% ± *4.04*
OSGA + SVM, $\alpha = 0.1$	79.058% ± *2.93*
OSGP, $ms = 100$	84.204% ± *2.82*
Results using no tumor markers	
Modeling method	Test accuracies ($\mu \pm \sigma$)
OSGA + LR, $\alpha = 0.1$	72.082% ± *4.30*
OSGA + kNN, $\alpha = 0.1$	78.247% ± *3.71*
OSGA + ANN, $\alpha = 0.1$	75.025% ± *3.40*
OSGA + SVM, $\alpha = 0.1$	78.136% ± *3.08*
OSGP, $ms = 100$	77.787% ± *4.81*

Table 6: Classification results for melanoma diagnosis

Results using all variables	
Modeling method	Test accuracies ($\mu \pm \sigma$)
OSGA + LR, $\alpha = 0.1$	79.526% ± *3.04*
OSGA + kNN, $\alpha = 0.1$	73.710% ± *3.99*
OSGA + ANN, $\alpha = 0.1$	80.171% ± *2.82*
OSGA + SVM, $\alpha = 0.1$	79.700% ± *2.48*
OSGP, $ms = 100$	80.319% ± *4.64*
Results using no tumor markers	
Modeling method	Test accuracies ($\mu \pm \sigma$)
OSGA + LR, $\alpha = 0.1$	77.519% ± *4.52*
OSGA + kNN, $\alpha = 0.1$	71.133% ± *4.04*
OSGA + ANN, $\alpha = 0.1$	77.596% ± *4.27*
OSGA + SVM, $\alpha = 0.1$	78.086% ± *4.29*
OSGP, $ms = 100$	76.072% ± *5.89*

Table 7: Classification results for respiratory system cancer diagnosis

Results using all variables	
Modeling method	Test accuracies ($\mu \pm \sigma$)
OSGA + LR, $\alpha = 0.1$	93.518% ± *3.37*
OSGA + kNN, $\alpha = 0.1$	90.710% ± *3.42*
OSGA + ANN, $\alpha = 0.1$	94.148% ± *2.84*
OSGA + SVM, $\alpha = 0.1$	92.695% ± *4.19*
OSGP, $ms = 100$	93.518% ± *3.05*
Results using no tumor markers	
Modeling method	Test accuracies ($\mu \pm \sigma$)
OSGA + LR, $\alpha = 0.1$	92.242% ± *3.75*
OSGA + kNN, $\alpha = 0.1$	85.760% ± *2.56*
OSGA + ANN, $\alpha = 0.1$	93.346% ± *3.08*
OSGA + SVM, $\alpha = 0.1$	91.462% ± *2.91*
OSGP, $ms = 100$	92.411% ± *2.36*

- The average classification rate for breast cancer using all variables could be raised from maximum ∼82% to ∼84%, omitting tumor markers from maximum ∼75.5% to ∼78%.
- The average classification rate for melanoma using all variables could be raised from maximum ∼75% to ∼80%, omitting tumor markers from maximum ∼75% to ∼78%.
- The average classification rate for respiratory system cancer using all variables could be raised from maximum ∼91.5% to ∼94%, omitting tumor markers from maximum ∼87% to ∼92%.

These results are very encouraging and in the future we will research the capability of this approach to lead to better results on other data sets (real world as well as benchmark data collections).

Furthermore, we plan to use an evolutionary algorithm for optimizing the sets of features for clustering the data in order to even further improve the resulting cluster homogeneities and classification rates.

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] 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.

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

[6] 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`.

[7] D. L. Davies and D. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2:224–227, 1979.

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

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

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

[11] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. pages 1137–1143. Morgan Kaufmann, 1995.

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

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

[14] W. B. Langdon and R. Poli. *Foundations of Genetic Programming*. Springer Verlag, Berlin Heidelberg New York, 2002.

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

[16] D. MacKay. *Information Theory, Inference and Learning Algorithms*, chapter An Example Inference Task: Clustering, pages 284–292. Cambridge University Press, 2003.

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

[18] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Gaussian Mixture Models and k-Means Clustering*. Cambridge University Press, New York, 2007.

[19] 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.

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

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

[22] 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.

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

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

[25] S. M. 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.

[26] S. M. Winkler, M. Affenzeller, W. Jacak, and H. Stekel. Identification of cancer diagnosis estimation models using evolutionary algorithms – a case study for breast cancer, melanoma, and cancer in the respiratory system. In *Proceedings of the GECCO 2011 Workshop on Medical Applications of Genetic and Evolutionary Computation (MedGEC 2011)*, 2011.

[27] S. M. 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 22nd European Modeling & Simulation Symposium*, pages 1 – 6, 2010.

[28] S. M. Winkler, M. Affenzeller, and H. Stekel. An integrated clustering and classification approach for the analysis of tumor patient data. In *Computer Aided Systems Theory - EUROCAST 2013*, pages 230–233, 2013.

[29] 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.

[30] 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 Journal of Stomatology*, 17(5):457–460, 2008.