Analysis of Selected Evolutionary Algorithms in Feature Selection and Parameter Optimization for Data Based Tumor Marker Modeling*

Stephan M. Winkler[1], Michael Affenzeller[1], Gabriel Kronberger[1], Michael Kommenda[1], Stefan Wagner[1], Witold Jacak[1], and Herbert Stekel[2]

[1] Heuristic and Evolutionary Algorithms Laboratory
School of Informatics, Communications and Media
Upper Austrian University of Applied Sciences, Campus Hagenberg
Softwarepark 11, 4232 Hagenberg, Austria
{stephan.winkler,michael.affenzeller,gabriel.kronberger, michael.kommenda,stefan.wagner,witold.jacak}@fh-hagenberg.at

[2] Central Laboratory, General Hospital Linz
Krankenhausstraße 9, 4021 Linz, Austria
herbert.stekel@akh.linz.at

Abstract. In this paper we report on the use of evolutionary algorithms for optimizing the identification of classification models for selected tumor markers. Our goal is to identify mathematical models that can be used for classifying tumor marker values as normal or as elevated; evolutionary algorithms are used for optimizing the parameters for learning classification models. The sets of variables used as well as the parameter settings for concrete modeling methods are optimized using evolution strategies and genetic algorithms. The performance of these algorithms is analyzed as well as the population diversity progress. In the empirical part of this paper we document modeling results achieved for tumor markers *CA 125* and *CYFRA* using a medical data base provided by the Central Laboratory of the General Hospital Linz; empirical tests are executed using HeuristicLab.

1 Research Goal: Identification of Models for Tumor Markers

In general, tumor markers are substances (found in blood and / or 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 be used as indicators for the presence of cancer. As a matter of fact, elevated tumor marker values themselves are not diagnostic, but rather only suggestive; tumor markers can be used to monitor

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

R. Moreno-Díaz et al. (Eds.): EUROCAST 2011, Part I, LNCS 6927, pp. 335–342, 2011.

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 [1] (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) and [2].

The general goal of the research work described here is to identify models for estimating selected tumor marker values on the basis of routinely available blood values; in detail, estimators for the tumor markers CA 125 and CYFRA have been identified. The documented tumor marker values are (using limits known from literature) classified as "normal", "slightly elevated", "highly elevated", and "beyond plausible"; our goal is to design classifiers for the 2-class-classification problem classifying samples into "normal" vs. "elevated".

In the research work reported on in this paper we use evolutionary algorithms for optimizing the selection of features that are used by machine learning algorithms for modeling the given target values. This approach is related to the method described in [3] where the authors compared the use of a particle swarm optimization (PSO) and a genetic algorithm (GA), both augmented with support vector machines, for the classification of high dimensional microarray data.

2 Optimization of Feature Selections and Modeling Parameters Using Evolutionary Algorithms

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

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

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

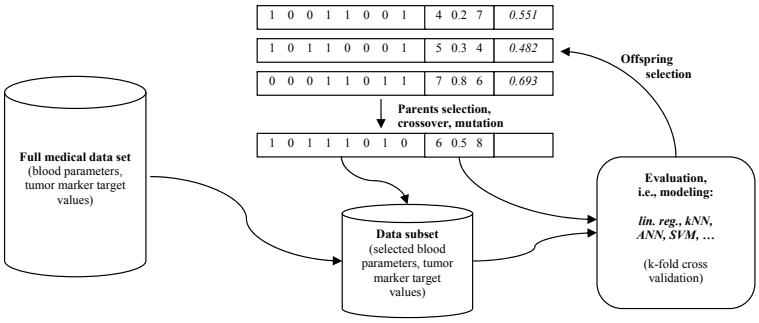

Fig. 1. A hybrid evolutionary algorithm for feature selection and parameter optimization in data based modeling

classification. In [4] we have analyzed the sets of features identified as relevant in the modeling of selected tumor markers; we have now used evolutionary algorithms for finding optimal feature sets as well as optimal modeling parameters for models for tumor markers. This approach is schematically shown in Figure 1: A solution candidate is 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.

3 Population Diversity Analysis

For analyzing the internal dynamics of optimization algorithms we analyze population diversity progress in analogy to the method proposed in [5] for various problem classes (such as TSP, CVRP, and symbolic regression and classification) and different evolutionary algorithms: All solutions in a given population are compared to each other and the similarities of these solutions are calculated. For the features and parameters optimization approach described in this paper we define the following similarity estimation function:

Let solution candidate sc_i be defined as $[s_{i1,\ldots,in}p_{i1,\ldots,iq}]$. For calculating the similarity between two solutions sc_i and sc_j we calculate the number of feature selection decisions ($s_{i1,\ldots,in}$ and $s_{j1,\ldots,jn}$) that are equal in both solutions as well as the relative differences of the selected modeling parameters ($p_{i1,\ldots,iq}$ and $p_{j1,\ldots,jq}$); a factor β is used for weighting these two aspects. Thus, we define the similarity function $sim(sc_i, sc_j)$ as

$$sim(sc_i, sc_j) = \beta * \frac{1}{n} * |\{k : s_{ik} = s_{jk}\}| + (1-\beta) * \frac{1}{q} * \sum_{k=1}^{q} (1 - \frac{|p_{ik} - p_{jk}|}{r_k}) \quad (2)$$

where r_k is the range of modeling parameter k.

4 Empirical Tests

4.1 The AKH Data Base and Selected Tumor Markers

In this research work our goal is to identify models for the tumor markers **CA 125** and **CYFRA**:

- **CA 125:** Cancer antigen 125 (CA 125) ([6]), 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, especially recurring ovarian cancer [7]. Even though CA 125 is best known as a marker for ovarian cancer, it may also be elevated in the presence of other types of cancers; for example, increased values are seen in the context of cancer in fallopian tubes, lungs, the endometrium, breast and gastrointestinal tract.

- **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. Due to elevated values in the presence of lung cancer CYFRA is often used for detecting and monitoring malign lung tumors. Elevated CYFRA values have already been reported for several different kinds of tumors, especially for example in stomach, colon, breast, and ovaries. The use of CYFRA 21-1 as a tumor marker has for example been discussed in [8].

Data of thousands of patients of the General Hospital (AKH) Linz, Austria, have been analyzed in order to identify mathematical models for tumor markers. We have used a medical data base compiled at the Central Laboratory of the General Hospital Linz, Austria, in the years 2005 - 2008: 28 routinely measured blood values of thousands of patients are available as well as several tumor markers; not all values are measured for all patients, especially tumor marker values are determined and documented if there are indications for the presence of cancer. Details about this data base and the variables available therein as well as the data preprocessing steps that were necessary (such as elimination of features with too many missing values, etc.) can be found in [9].

From the AKH data base we have compiled the following data sets for these two selected tumor markers:

- *CA 125* data set: 1,053 samples, 50.52% belonging to class 0 ("normal"); target variable: CA 125, variables available for modeling: Age, sex, ALT, AST, BUN, CRP, GT37, HB, HKT, HS, KREA, LD37, MCV, PLT, RBC, TBIL, WBC.
- *CYFRA* data set: 419 samples, 70.64% belonging to class 0 ("normal"); target variable: CYFRA, variables available for modeling: Age, sex, ALT, AST, BUN, CH37, CHOL, CRP, CYFS, GT37, HB, HKT, HS, KREA, MCV, PLT, RBC, TBIL, WBC.

4.2 Modeling Algorithms

The following techniques for training classifiers have been used in this research project: Linear regression, neural networks, k-nearest-neighbor classification, and support vector machines. All these machine learning methods have been implemented using the HeuristicLab framework[1] [10], 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 [9].

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 $\theta_{1...m}$; a constant additive factor is also included into the model. Theoretical background of this approach can be found in [11].

kNN Classification. Unlike other data based modeling methods, k-nearest-neighbor classification [12] 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 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 [13]. 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 30% of the given training samples as internal validation data.

Support Vector Machines. Support vector machines (SVMs) are a widely used approach in machine learning based on statistical learning theory [14]. 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 [14]. In this work we have used the LIBSVM implementation described in [15], 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 parameter c ($c \in [0, 512]$) and the γ parameter of the SVM's kernel function ($\gamma \in [0, 1]$).

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

4.3 Optimization Algorithms

The following evolutionary algorithms have been used for optimizing feature sets and modeling parameters:

- *Evolution strategy (ES)* [16]: Population size: 10, random parents selection, number of children per generation: 20, plus selection (i.e., 10+20 ES), 100 iterations.
- *Genetic algorithm (GA)* [17]: Population size: 10, tournament selection (k=2), 30% mutation rate, 200 iterations.
- *Genetic algorithm with strict offspring selection (OSGA)* [5]: Population size 10, random & roulette parents selection, 30% mutation rate, strict offspring selection (success ratio and comparison factor: 1.0), maximum selection pressure: 100, maximum number of evaluated solutions: 2,000.

For all algorithms the (maximum) number of evaluated solutions was set to 2,000 and the initial selection probability of each variable was (for each individual) set to 30%; for mutating solution candidates the bit flip probability for variable selections was set to 30%, σ for the Gaussian mutation of real valued parameters to 0.3. The fitness function described in Equation 2 was used, α was set to 0.1 as this value has been identified as suitable in previous research work (see [4], e.g.).

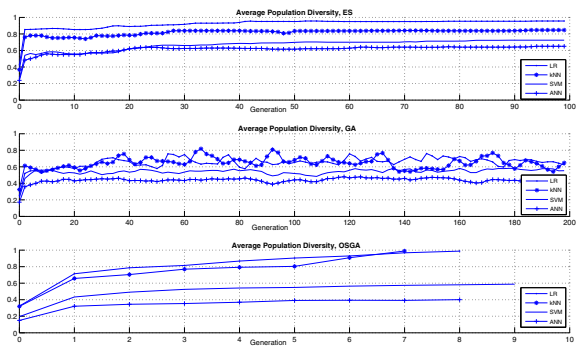

Fig. 2. Average population similarity in test runs using GA, ES, and OSGA for optimizing feature selection and modeling parameters for LinReg, kNN, ANNs, and SVMs

4.4 Results

We here summarize modeling results obtained using the algorithms listed in the previous sections; average classification accuracies are documented here as well as population diversity aspects. Five-fold cross-validation [18] training / test series have been executed for evaluating feature selections and modeling parameter configurations, each evolutionary algorithm was executed five times. The evaluation of the generated models on validation data (selected from the training samples) is used for calculating the fitness of solution candidates of the

Table 1. Modeling results for CA 125 and CYFRA

		CA 125	
		Classification accuracy ($\mu \pm \sigma$)	Variables ($\mu \pm \sigma$)
ES	LinReg	0.6174 (± 0.0167)	2.2 (± 0.84)
	kNN	0.6636 (± 0.0178)	3.8 (± 1.31)
	ANN	0.6281 (± 0.0369)	4.4 (± 0.55)
	SVM	0.6417 (± 0.0205)	3.8 (± 1.48)
GA	LinReg	0.6556 (± 0.0121)	6.8 (± 1.30)
	kNN	0.6631 (± 0.0065)	10.4 (± 1.52)
	ANN	0.6397 (± 0.0152)	5.6 (± 1.14)
	SVM	0.6463 (± 0.0073)	7.2 (± 0.84)
OSGA	LinReg	0.6527 (± 0.0043)	4.8 (± 0.84)
	kNN	0.6745 (± 0.0157)	8.2 (± 1.10)
	ANN	0.6516 (± 0.0217)	4.2 (± 0.84)
	SVM	0.6412 (± 0.0210)	6.8 (± 0.84)
		CYFRA	
		Classification accuracy ($\mu \pm \sigma$)	Variables ($\mu \pm \sigma$)
ES	LinReg	0.7262 (± 0.0107)	3.6 (± 1.52)
	kNN	0.7012 (± 0.0163)	5.0 (± 1.87)
	ANN	0.7232 (± 0.0230)	2.4 (± 0.55)
	SVM	0.7225 (± 0.0328)	3.6 (± 2.07)
GA	LinReg	0.7367 (± 0.0049)	6.4 (± 2.51)
	kNN	0.7057 (± 0.0027)	9.2 (± 0.84)
	ANN	0.7124 (± 0.0185)	8.0 (± 4.18)
	SVM	0.7162 (± 0.0249)	12.6 (± 2.07)
OSGA	LinReg	0.7315 (± 0.0130)	3.4 (± 1.52)
	kNN	0.7279 (± 0.0135)	6.8 (± 1.92)
	ANN	0.7339 (± 0.0072)	5.0 (± 1.87)
	SVM	0.7275 (± 0.0072)	5.4 (± 2.19)

optimization algorithms; these fitness values are the basis for the selection of the best models eventually presented by the optimization processes. Test results presented in Table 1 are "real" test figures, these accuracies have been calculated on independent test samples. In Figure 2 the population diversity progress of the evolutionary optimization runs used here is depicted.

5 Conclusion

Comparing the results summarized in Table 1 with those published in [9] we see that for each modeling method the achieved classification results could be improved: The classification accuracies could be increased, and the sizes of the sets of used variables could be decreased significantly. All three evolutionary algorithms were successful in finding improved feature sets and modeling parameters; comparing ES, GA, and OSGA we see that especially ES and OSGA tend to produce significantly smaller feature sets. Regarding the progress of population diversity in the evolutionary algorithms tested here we see that in ES and OSGA the diversity in the population tends to decrease during the algorithm's execution more than when using a GA.

In future research work we will investigate the use of tumor marker estimation models in the prediction of tumor diagnoses: As we have now identified classification models for tumor markers that can be used for estimating tumor marker values on the basis of standard blood parameters, these virtual tumor markers shall be used in combination with standard blood parameters for learning classifiers that can be used for predicting tumor diagnoses.

References

1. Koepke, J.A.: Molecular marker test standardization. Cancer 69, 1578–1581 (1992)
2. Bitterlich, N., Schneider, J.: Cut-off-independent tumour marker evaluation using ROC approximation. Anticancer Research 27, 4305–4310 (2007)
3. Alba, E., Jourdan, J.G.N.L., Talbi, E.G.: Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms. IEEE Congress on Evolutionary Computation, 284–290 (2007)
4. Winkler, S., Affenzeller, M., Kronberger, G., Kommenda, M., Wagner, S., Jacak, W., Stekel, H.: Feature selection in the analysis of tumor marker data using evolutionary algorithms. In: Proceedings of the 7th International Mediterranean and Latin American Modelling Multiconference, pp. 1–6 (2010)
5. Affenzeller, M., Winkler, S., Wagner, S., Beham, A.: Genetic Algorithms and Genetic Programming - Modern Concepts and Practical Applications. Chapman & Hall / CRC (2009)
6. Yin, B.W., Dnistrian, A., Lloyd, K.O.: Ovarian cancer antigen CA125 is encoded by the MUC16 mucin gene. International Journal of Cancer 98, 737–740 (2002)
7. Osman, N., O'Leary, N., Mulcahy, E., Barrett, N., Wallis, F., Hickey, K., Gupta, R.: Correlation of serum ca125 with stage, grade and survival of patients with epithelial ovarian cancer at a single centre. Irish Medical Journal 101, 245–247 (2008)
8. Lai, R.S., Chen, C.C., Lee, P.C., Lu, J.Y.: Evaluation of cytokeratin 19 fragment (cyfra 21-1) as a tumor marker in malignant pleural effusion. Japanese Journal of Clinical Oncology 29(199), 421–424
9. Winkler, S., Affenzeller, M., Jacak, W., Stekel, H.: 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)
10. Wagner, S.: Heuristic Optimization Software Systems – Modeling of Heuristic Optimization Algorithms in the HeuristicLab Software Environment. PhD thesis, Johannes Kepler University Linz (2009)
11. Ljung, L.: System Identification – Theory For the User, 2nd edn. PTR Prentice Hall, Upper Saddle River (1999)
12. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley Interscience, Hoboken (2000)
13. Nelles, O.: Nonlinear System Identification. Springer, Heidelberg (2001)
14. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
15. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001), Software, `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
16. Schwefel, H.P.: Numerische Optimierung von Computer-Modellen mittels der Evolutionsstrategie. Birkhäuser Verlag, Basel (1994)
17. Holland, J.H.: Adaption in Natural and Artifical Systems. University of Michigan Press, Ann Arbor (1975)
18. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, vol. 2, pp. 1137–1143. Morgan Kaufmann, San Francisco (1995)