An Integrated Clustering and Classification Approach for the Analysis of Tumor Patient Data

Stephan M. Winkler[1], Michael Affenzeller[1], and Herbert Stekel[2]

[1] Heuristic and Evolutionary Algorithms Laboratory; Bioinformatics Research Group
University of Applied Sciences Upper Austria, Hagenberg Campus
Softwarepark 11, 4232 Hagenberg, Austria
{stephan.winkler,michael.affenzeller}@fh-hagenberg.at

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

Abstract. Standard patient parameters, tumor markers, and tumor diagnosis records are used for identifying prediction models for tumor markers as well as cancer diagnosis predictions. In this paper we present 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. The so formed clusters are analyzed and their homogeneity is calculated; the models learned on the basis of these clusters are tested and compared to each other with respect to classification accuracy and variable impacts.

1 An Integrated Clustering and Classification Approach for the Identification of Predictors for Tumor Diagnoses

The overall goal of the research described here is to identify prediction models for tumor markers (TM) and tumor diagnoses. In previous work ([13], [16]) we have identified classification models that can be used as virtual tumor markers for estimating TM values on the basis of standard blood parameters. Tumor markers are substances (found in blood and/or body tissues) that can be used as indicators for certain types of cancer ([4], [14]). Moreover, in [14] and [15] we have published research results achieved in the identification of prediction models for tumor diagnoses. As described in [15], the use of TM prediction models as virtual tumor markers increases the achievable classification accuracy.

The here proposed analysis approach (schematically shown in Figure 1) 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

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).

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. Within the *Heureka!* research project we have applied several clustering approaches including k-means clustering and soft k-means clustering ([9], [8]) as well as the identification of Gaussian mixture models using expectation maximization techniques [18]. 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 [3] as well as the Akaike information criterion [2] can be used, e.g.

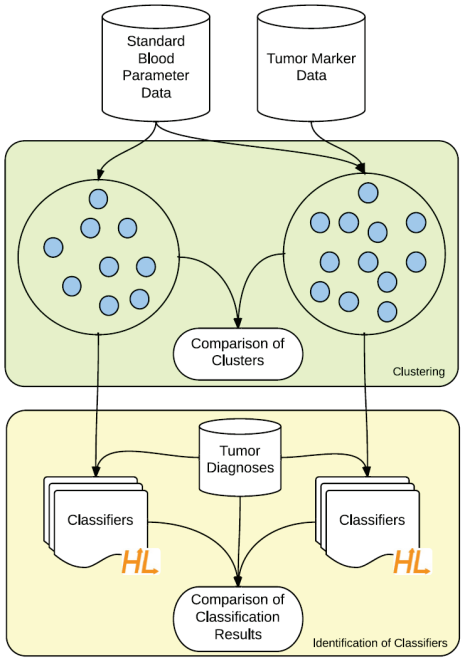

Fig. 1. An integrated clustering and classification approach for the analysis of medical data: Data clusters are formed using standard data and optionally also tumor marker data; these clusters are the basis for the identification of classifiers that can be used as predictors for cancer diagnoses.

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. We use the following two modeling methods for identifying predictors for tumor markers and cancer diagnoses: Hybrid modeling using machine learning algorithms and evolutionary algorithms (that optimize feature selection and the modeling algorithms' parameters) as well as genetic programming.

The so identified models are analyzed and compared to each other with respect to classification accuracy and variable impacts.

2 Empirical Test Study: Clustering and Classification of Breast Cancer Patient Data

2.1 Data Basis

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 standard values of patients are available as well as several tumor markers. 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 [13] and [14].

Information about cancer diagnoses is also available in the AKH database: If a patient is diagnosed with any kind of cancer, then this information 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 breast cancer (BC, cancer class C50 according to the International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10)). Following the data preprocessing approach described in [13] and [14] we have compiled a data set specific for this kind of tumor: 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 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 for the analysis of breast cancer patient data; this data set contains 706 samples (45.89% of not diseased patients forming class 0 and 54.11% of diseased patients forming class 1) containing routinely measured values of patients as well as tumor markers. This data set is the same as the BC data set used in [14].

2.2 Clustering Results

The so compiled data set of patients was clustered using k-means algorithm ([9], [8]) 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 of 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 [3] 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:

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

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

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:

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

$$homogeneity(cluster_j) = \max_i (r(class_i, cluster_j)) \tag{4}$$

As we are interested 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:

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

where n is the total number of samples.

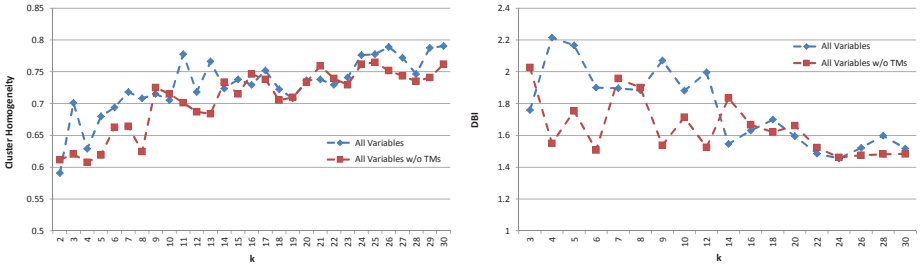

Fig. 2. Progress of cluster homogeneity (left) and DBI (right) for clusterings with varying numbers of clusters

In Figure 2 we show the progress of the clusters' homogeneity over k as well as the progress of the DBI over k (averages of 5 independent clusterings for each k and data partition are shown). We see that setting $k = 25$ seems to yield optimal clustering results as the DBI is minimized and the cluster homogeneity

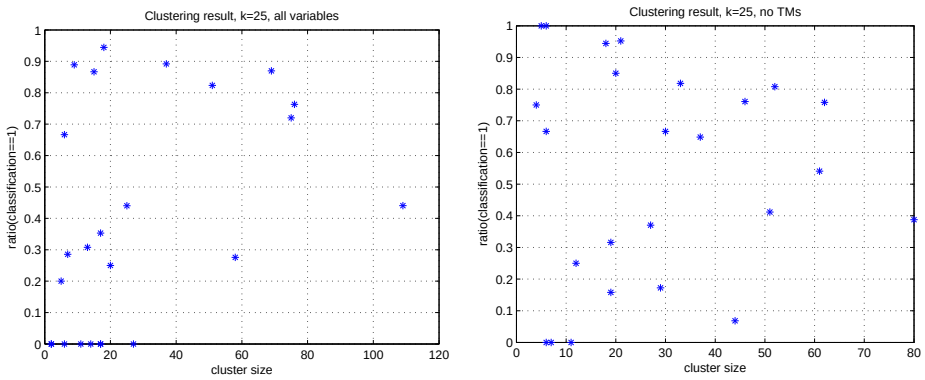

Fig. 3. Clustering result for $k = 25$: Size and homogeneity of clusters formed by k-means clustering using all variables (left) and all variables except tumor markers (right)

reaches relatively high values using all variables as well as using all variables except tumor makers; tumor diagnoses were not used for clustering. These results are consistent with result presented in [17]. In Figure 3 we show an overview of exemplary clustering results achieved for $k = 25$; the sizes as well as the homogeneity of the formed clusters are shown (each spot represents one cluster).

2.3 Identification of Classifiers Using Clustered Data

Finally, using the previously identified clusters 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 [5] training / test series have been executed; in order to avoid overfitting, all clusters with less than 45 samples were (for each clustering separately) combined into "rest" clusters.

The following approaches have been applied for learning BC classifiers for the previously identified clusters:

- Hybrid modeling using support vector machines (SVMs, [10]) and a genetic algorithm (GA) with strict offspring selection (OS, [1]) for parameter optimization and feature selection as described in [14], e.g., and shown in Figure 4; this approach is referred to as "OSGA+SVM".
- Genetic programming (GP) ([6], [7], [12]) in combination with strict offspring selection as shown in Figure 5, referred to as "OSGP".

The implementations of these approaches in HeuristicLab[1] [11] have been applied; for the evolutionary process the population sizes were set to 10 and 100, respectively, and for evolutionary feature selection the parsimony pressure α was set to 0.1.

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

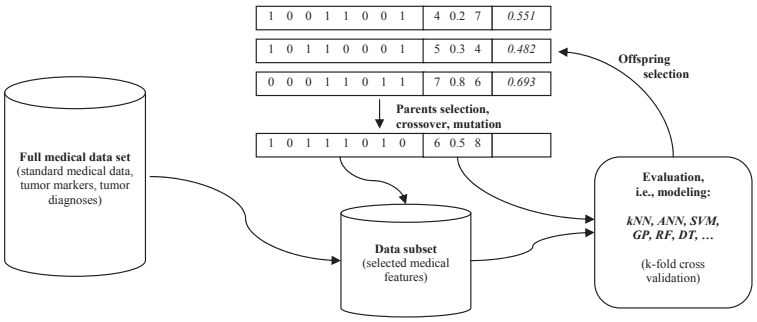

Fig. 4. A hybrid evolutionary algorithm for feature selection and parameter optimization in data based modeling.

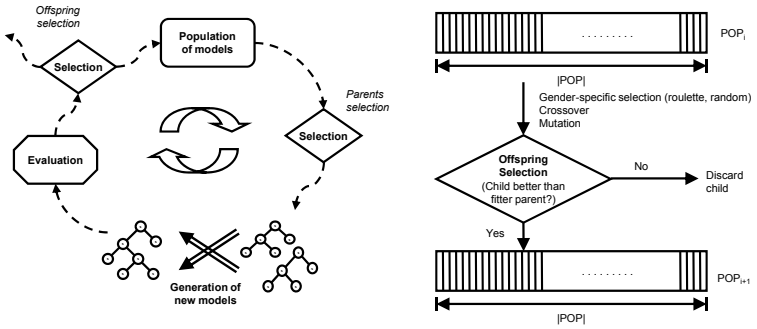

Fig. 5. The genetic programming cycle [7] (left) including strict offspring selection [1] (right).

In [14] we have documented that using all variables except tumor markers up to $\sim 75\%$ of the given samples in the BC data set can be classified correctly; using the here discussed clustering and classification approach we are able to reach the following classification rates:

- Hybrid modeling (OSGA+SVM, $\alpha = 0.1$): 78.136% (± 3.08)
- Genetic programming with offspring selection (OSGP): 77.787% (± 4.81)

Please note that each method was applied 5 times using 5-fold cross validation and that the here stated numbers are averages of weighted averages (calculated as the average of the classification accuracies on the given clusters multiplied with their relative size); no tumor markers were used for clustering or learning classifiers. More result details shall be presented in [17].

One of the major advantages of the here discussed approach is that it first clusters the data in groups of rather similar samples and is then able to separate the classes within these groups; for different clusters the algorithms are able to

use different variables for forming classifiers. In order to analyze this behavior we have analyzed the importance of the given variables: We have exemplarily used the clusters formed using $k = 25$ (and no tumor markers; as described previously, all clusters with size <45 were merged into a cluster here called the "REST" cluster) and documented the frequency of the available variables in the final solutions of the applied evolutionary modeling approaches (OSGA+SVM and OSGP). The results are shown in Table 1: An "X" indicates that a variable was used in at least 80% of the executed test runs while a "x" means that it was used in at least 40% of the executed classification runs.

Cluster index	Modeling method	ALTER	ALT & AST	BSG1	BUN	CBAA	CEOA	CHOL	CLYA	CMOA	CNEA & WBC	CRP	FER	GT37	HB & HKT & RBC	HDL	HS	KREA	LD37	MCV	PLT	TBIL	TPS
1	OSGA+SVM	x				X							x		x						x		
	OSGP	x	X			x		X					x	x	X			x		x			
2	OSGA+SVM										X			x	x							x	
	OSGP				x						X		x	x								X	
3	OSGA+SVM	x	X	x	X	x														X			
	OSGP	x	x					x			X			x	x					x			
4	OSGA+SVM	x	x										X					X					
	OSGP						X	x		X					x								
9	OSGA+SVM									x			x			x		x	x				
	OSGP								x	x	x				x			X	x		X		
16	OSGA+SVM										x	x							x				x
	OSGP											x					x			x	x		
REST	OSGA+SVM		X		X		X		x				x		x	x							
	OSGP		X				X	X	x				x		x		x						

Table 1. Relevant variables for classifying pre-clustered samples

3 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: The classification accuracy of potential breast cancer patients (without considering tumor markers) can be increased to ∼78% using evolutionary modeling (hybrid modeling or genetic programming). Furthermore, we see that there are significant differences regarding the importance of variables for classifying pre-clustered data: Some variables are essentially important for classifying samples of certain clusters while they might be irrelevant for forming classifiers for other clusters.

Further research shall focus on 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.

References

1. Affenzeller, M., Winkler, S., Wagner, S., Beham, A.: Genetic Algorithms and Genetic Programming - Modern Concepts and Practical Applications. Chapman & Hall / CRC (2009)
2. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on automatic control 19, 716–723 (1974)
3. Davies, D.L., Bouldin, D.: A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence 2, 224–227 (1979)
4. Koepke, J.A.: Molecular marker test standardization. Cancer 69, 1578–1581 (1992)
5. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. pp. 1137–1143. Morgan Kaufmann (1995)
6. Koza, J.R.: Genetic Programming: On the Programming of Computers by Means of Natural Selection. The MIT Press (1992)
7. Langdon, W.B., Poli, R.: Foundations of Genetic Programming. Springer Verlag, Berlin Heidelberg New York (2002)
8. MacKay, D.: Information Theory, Inference and Learning Algorithms, pp. 284–292. Cambridge University Press (2003)
9. Press, W., Teukolsky, S., Vetterling, W., Flannery, B.: Gaussian Mixture Models and k-Means Clustering. Cambridge University Press, New York (2007)
10. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
11. Wagner, S.: Heuristic Optimization Software Systems – Modeling of Heuristic Optimization Algorithms in the HeuristicLab Software Environment. Ph.D. thesis, Johannes Kepler University Linz (2009)
12. Winkler, S.: Evolutionary System Identification - Modern Concepts and Practical Applications. Ph.D. thesis, Institute for Formal Models and Verification, Johannes Kepler University Linz (2008)
13. Winkler, S., Affenzeller, M., Jacak, W., Stekel, H.: Classification of tumor marker values using heuristic data mining methods. In: Proceedings of the Genetic and Evolutionary Computation Conference GECCO 2010 (2010)
14. Winkler, S., Affenzeller, M., Jacak, W., Stekel, H.: 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 Genetic and Evolutionary Computation Conference GECCO 2011 (2011)
15. Winkler, S., Affenzeller, M., Kronberger, G., Kommenda, M., Wagner, S., Dorfer, V., Jacak, W., Stekel, H.: On the use of estimated tumor marker classifications in tumor diagnosis prediction - a case study for breast cancer. Accepted to be published in: International Journal of Simulation and Process Modelling (2013)
16. Winkler, S., Affenzeller, M., Kronberger, G., Kommenda, M., Wagner, S., Jacak, W., Stekel, H.: On the use of estimated tumor marker classifications in tumor diagnosis prediction - a case study for breast cancer. In: Proceedings of the 23rd European Modeling & Simulation Symposium (2011)
17. Winkler, S., Affenzeller, M., Stekel, H.: Evolutionary identification of cancer predictors using clustered data - a case study for breast cancer, melanoma, and cancer in the respiratory system. In: Proceedings of the Genetic and Evolutionary Computation Conference GECCO 2013 (2013)
18. Xu, L., Jordan, M.I.: On convergence properties of the EM algorithm for Gaussian mixtures. Neural Computation 8, 129–151 (1995)