Interpretable Classifiers in Precision Medicine: Feature Selection and Multi-class Categorization

Lyn-Rouven Schirra[1,2], Florian Schmid[1], Hans A. Kestler[1(✉)], and Ludwig Lausser[1]

[1] Institute of Medical Systems Biology, Ulm University, 89069 Ulm, Germany
hans.kestler@uni-ulm.de

[2] Institute of Number Theory and Probability Theory, Ulm University, 89069 Ulm, Germany

Abstract. Growing insight into the molecular nature of diseases leads to the definition of finer grained diagnostic classes. Allowing for better adapted drugs and treatments this change also alters the diagnostic task from binary to multi-categorial decisions. Keeping the corresponding multi-class architectures accurate and interpretable is currently one of the key tasks in molecular diagnostics.

In this work, we specifically address the question to which extent biomarkers that characterize pairwise differences among classes, correspond to biomarkers that discriminate one class from all remaining. We compare one-against-one and one-against-all architectures of feature selecting base classifiers. They are validated for their classification performance and their stability of feature selection.

1 Introduction

The analysis of molecular profiles adds a new instrument to the toolbox of medical diagnoses. It allows for a deeper insight in the molecular processes of a cell or a tissue. Due to their high dimensionality, the interpretation of these profiles is often quite challenging. Comprising tens of thousands of molecular measurements, the size of a profile typically exceeds the possibility of a direct visual inspection. Computer-aided classification algorithms are needed for diagnostic purposes [9,18,22]. Training these models often incorporates an internal feature selection process [14,21], which basically yields at a limitation of the measurements in the final prediction [8,19]. The resulting feature signature typically optimizes heuristic criteria [12]. It is often constructed in a purely data-driven or model-driven procedure [2]. Alternatively, feature selection can also be conducted from (prior) domain knowledge about the subject or the measuring process of an experiment [15].

One of the most important findings from the analysis of molecular profiles, is the insight that an observable phenotype or disease that was thought to be

L.-R. Schirra and F. Schmid—Contributed equally.

H.A. Kestler and L. Lausser—Joint senior authors.

F. Schwenker et al. (Eds.): ANNPR 2016, LNAI 9896, pp. 105–116, 2016.
DOI: 10.1007/978-3-319-46182-3_9

a uniform entity can be evoked by varying molecular causes [11]. These refinements of the traditional phenotypes bring up the possibility of more specific treatments and can be seen as a starting point for the field of precision medicine or personalized medicine [4]. From a diagnostic point of view, the challenge of identifying a correct phenotype has changed due to the increased number of diagnostic classes [7]. Primarily designed for binary categorization problems many classification models cannot be directly applied to such a multi-class scenario [6,20]. Fusion architectures for combining an ensemble of binary classifiers are needed [16]. These combining techniques can either follow a predefined scheme [5] or are adapted during the overall training phase of the classifier system [10]. Fusion architectures can also incorporate known relationships among the diagnostic categories [13].

In this work, we focus on the interactions between data-driven feature selection strategies and predefined multi-class architectures. We analyze the influence of the one-against-one and the one-against-all training on the selection strategy of feature selecting base classifiers [16]. We compare the multi-class architectures according to their classification performance and evaluate the selection stability of the individual feature sets.

2 Methods

We denote classification as the task of predicting a category $y \in \mathcal{Y}$ of an object according to a vector of measurements $\mathbf{x} \in \mathcal{X}$. Here, $\mathcal{Y}$ denotes a finite label space with $|\mathcal{Y}| \geq 2$ and $\mathcal{X} \in \mathbb{R}^n$ denotes the feature space. The single categories will be represented by natural numbers $\mathcal{Y} = \{1, \ldots, |\mathcal{Y}|\}$. The ordering of these numbers is not assumed to reflect a known ordering of categories. The elements of a feature vector will be denoted by $\mathbf{x} = (x^{(1)}, \ldots, x^{(n)})^T$. A classification function, a classifier, will be seen as a function mapping

$$c : \mathcal{X} \to \mathcal{Y}. \tag{1}$$

A classifier is typically chosen from a predefined concept class $\mathcal{C}$ and adapted in a data-driven procedure l. Here, the classifier is trained on a set of labeled training sampled $\mathcal{T} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$,

$$l : \mathcal{C} \times \mathcal{T} \to c_{\mathcal{T}}. \tag{2}$$

After this initial training phase, the classifier can be applied for predicting the class label of new unseen samples. Its generalization ability is typically quantified on a separate set of validation samples $\mathcal{V} = \{(\mathbf{x}'_i, y'_i)\}_{i=1}^{m'}$, $\mathcal{T} \cap \mathcal{V} = \emptyset$. We will mainly focus on the overall accuracy of a classifier $c(\mathbf{x})$

$$\mathrm{acc}_{\mathcal{V}} = \frac{1}{|\mathcal{V}|} \sum_{(\mathbf{x},y)\in\mathcal{V}} \mathbb{I}_{[c(\mathbf{x})=y]} \tag{3}$$

and the multi-class extensions of the sensitivity $se_{\mathcal{V}}(y)$ and specificity $sp_{\mathcal{V}}(y)$

$$se_{\mathcal{V}}(y) = \frac{1}{|\mathcal{V}_y|} \sum_{(\mathbf{x},y')\in\mathcal{V}_y} \mathbb{I}_{[c(\mathbf{x})=y]} \quad \text{and} \quad sp_{\mathcal{V}}(y) = \frac{1}{|\mathcal{V}\setminus\mathcal{V}_y|} \sum_{(\mathbf{x},y')\in\mathcal{V}\setminus\mathcal{V}_y} \mathbb{I}_{[c(\mathbf{x})\neq y]}. \tag{4}$$

Here, $\mathcal{V}_y = \{(\mathbf{x}, y') \in \mathcal{V} \,|\, y' = y\}$ denotes those samples in $\mathcal{V}$ with class label y and $\mathbb{I}_{[]}$ denotes the indicator function.

2.1 Multi-class Classification

In this work, we are mainly interested in training strategies that construct multi-class classifiers ($|\mathcal{Y}| > 2$) by the combination of binary ones. These strategies are allowed to train an ensemble of base classifiers $\mathcal{E} \subset \mathcal{C}$

$$\mathcal{E} = \{c_i : \mathcal{X} \to \mathcal{Y}_i\}_{i=1}^{|\mathcal{E}|} \tag{5}$$

that are finally combined via a subsequently applied fusion strategy

$$h_{\mathcal{E}} : \mathcal{Y}_1 \times ... \times \mathcal{Y}_{|\mathcal{E}|} \to \mathcal{Y} \text{ with } h_{\mathcal{E}}(c_1(\boldsymbol{x}), ..., c_{|\mathcal{E}|}(\mathbf{x})) = y. \tag{6}$$

We will restrict ourselves to untrainable fusion strategies in the following.

The elements of $\mathcal{E}$ can in principle be trained on different subsets of the overall training set and they can also map to distinct label spaces with distinct interpretation. A base classifier will be denoted by

$$c_{(y,y')} : \mathcal{X} \to \{y, y'\}, \tag{7}$$

if its label space is of interest. In the following, we will restrict ourselves to the well known one-against-one scheme and the one-against-all scheme.

One-against-one Scheme (OaO). The one-against-one (OaO) architecture comprises ensemble members for each pairwise classification between the elements in $\mathcal{Y}$, therefore it consists of $|\mathcal{E}| = \frac{|\mathcal{Y}|(|\mathcal{Y}|-1)}{2}$ base classifiers. Each ensemble member utilizes solely the training samples $\mathcal{T}_y \cup \mathcal{T}_{y'}$ of the selected classes

$$c_{(y,y')} : \mathcal{X} \to \{y, y'\} \quad \text{for all} \quad y, y' \in \mathcal{Y},\, y < y'. \tag{8}$$

For the prediction of a multi-class label $y \in \mathcal{Y}$, all base classifiers of the OaO ensemble are applied simultaneously. Their individual predictions are finally combined via an unweighted majority vote

$$h_{OaO}(\mathbf{x}) = \arg\max_{y\in\mathcal{Y}} \sum_{c\in\mathcal{E}} \mathbb{I}_{[c(\mathbf{x})=y]}. \tag{9}$$

One-against-all Scheme (OaA). The one-against-all (OaA) architecture can be seen as an ensemble of one-class detectors. It consists of $|\mathcal{E}| = |\mathcal{Y}|$ members. Each base classifier $c \in \mathcal{E}$ is trained to separate a single class $y \in \mathcal{Y}$ from the remaining ones

$$c_{(y,\bar{y})} : \mathcal{X} \to \{y, \bar{y}\} \quad \text{for all} \quad y \in \mathcal{Y},\, \bar{y} = \mathcal{Y} \setminus y. \tag{10}$$

In this context, the artificial complement class $\bar{y}$ of y is often referred to as the rest class of y. As a class label $y \in \mathcal{Y}$ can only be predicted by one member of an OaA ensemble, an unweighted majority-vote would be prone to ties. It is typically replaced by some kind of weighting scheme of the single decisions. In the basic version of the OaA, it is assumed that the chosen type of base classifier $\mathcal{C}$ is able to provide an additional certainty measure $p_{(y,\bar{y})}(\mathbf{x})$ of a prediction $c_{(y,\bar{y})}(\mathbf{x}) = y$. The final OaA multi-class decision is then given by

$$h_{OaA}(\mathbf{x}) = \arg\max_{y \in \mathcal{Y}} p_{(y,\bar{y})}(\mathbf{x}). \tag{11}$$

2.2 Feature Selection for Multi-class Classification

The training of classification algorithms can incorporate a feature selection phase in which the number of available measurements is reduced to a small number of important markers. Especially in high-dimensional settings, such as the analysis of gene expression profiles, a reduction to a small, interpretable and measurable set of markers is of interest. The selected features can lead to new hypotheses on the causes for a certain class difference.

Formally a feature selection process can be seen as a function

$$f : \mathcal{C} \times \mathcal{T} \rightarrow \mathcal{I} = \{\mathbf{i} \in \mathbb{N}^{\hat{n} \leq n} | i_k < i_{k+1}, 1 \leq i_k \leq n\}, \tag{12}$$

which maps to the space of index vectors $\mathbf{i} \in \mathcal{I}$ of maximal length n. The element of $\mathbf{i} = (i^{(1)}, \ldots, i^{(\hat{n})})^T$ indicate the features $\mathbf{x}^{(\mathbf{i})} = (x^{(i^{(1)})}, \ldots, x^{(i^{(\hat{n})})})$ that will be passed to a subsequent processing step.

In the context of multi-class ensembles, feature selection can be applied in different ways. It can either be seen as an initial step of the overall ensemble training (TYPE I) or as a component of the individual training procedures (TYPE II) of the corresponding base classifiers. If feature selection is incorporated in the overall ensemble training, one common feature signature is extracted for all base classifiers. A suitable (supervised) selection criterion for this type of feature selection must be able to handle multi-class labels.

If feature selection is incorporated in the training of the base classifiers, an individual feature signature $\mathcal{I}_{\mathcal{E}} = \{\mathbf{i}_1, \ldots, \mathbf{i}_{|\mathcal{E}|}\}$ is extracted for each of the ensemble members $\mathcal{E} = \{c_1, \ldots, c_{|\mathcal{E}|}\}$. Here, the feature selection process does not only provide a reduction of the initial feature set. The process also distributes the selected features among the base classifiers. As the ensemble members are designed for binary classification tasks, feature selection criteria for two-class scenarios can be applied in this framework.

As basic two-class feature selection criteria we have chosen the Threshold Number of Misclassification (TNoM) score proposed by Ben-Dor et al. [1], the Pearson's Correlation Coefficient (PCC) and the T-Test (TT).

Threshold Number of Misclassificaitons (TNoM): The TNoM is a univariate filter criterion which applies the (re-)classification error of a single threshold classifier $c_t(\mathbf{x})$ as selection score s_i

$$s_i = \frac{1}{n} \sum_{(\mathbf{x},y)\in\mathcal{T}} \mathbb{I}_{[c_t(\mathbf{x})=y]} \quad \text{with} \quad c_t(\mathbf{x}) = \begin{cases} y' & \text{if} \quad d(x^{(i)} - t) \geq 0 \\ y'' & \text{otherwise.} \end{cases} . \tag{13}$$

The parameters $t \in \mathbb{R}$ and $d \in \{-1, +1\}$ are initially optimized for minimizing the classification error (TNoM) on $\mathcal{T}$ or the mean classwise error (TNoM$_{cw}$).

Pearson's Correlation Coefficient (PCC): The PCC is designed for the detection of linear correlations of two measurements. If applied to a single feature and the corresponding class label, it can be utilized as a score for univariate feature selection. We will utilize the absolute PCC in the following

$$s_i = \left| \frac{\sum_{(\mathbf{x},y)\in\mathcal{T}} (x^{(i)} - \bar{x}^{(i)})(y - \bar{y})}{\sqrt{\sum_{(\mathbf{x},y)\in\mathcal{T}} (x^{(i)} - \bar{x}^{(i)})^2 \sum_{(\mathbf{x},y)\in\mathcal{T}} (y - \bar{y})^2}} \right| . \tag{14}$$

Here $\bar{x}^{(i)}$ and $\bar{y}$ denote the average feature value of the ith feature and the average class label.

T-Test (TT): The T-statistic is designed for detecting differences in the mean value of two normally distributed classes y' and y''. In its empirical version it can be used as a univariate feature selection criterion

$$s_i = \frac{\bar{x}^{(i)}_{y'} - \bar{x}^{(i)}_{y''}}{\sqrt{(\sigma^{(i)}_{y'})^2/|\mathcal{T}_{y'}| + (\sigma^{(i)}_{y''})^2/|\mathcal{T}_{y''}|}} . \tag{15}$$

Here $\bar{x}^{(i)}_y$ and $\sigma^{(i)}_y$ denote the classwise and featurewise mean value and standard deviation.

3 Experiments

In the following we are mainly interested in multi-class feature selection algorithms of TYPE II. We apply this kind of feature selection in experiments with both OaO-ensembles and OaA-ensembles and give a direct comparison of their results. As a base classifier the linear support vector machine was chosen [20]. It constructs linear decision rules of type

$$c_{(y,y')}(\mathbf{x}) = \begin{cases} y & \text{if} \quad p_{(y,y')}(\mathbf{x}) \geq 0 \\ y' & else \end{cases} \quad \text{and} \quad p_{(y,y')}(\mathbf{x}) = \mathbf{w}^t\mathbf{x} - t, \tag{16}$$

where $\mathbf{w} \in \mathbb{R}^n$ and $t \in \mathbb{R}$ are optimized to maximize the margin between the samples of y and y'.

For each base classifier, the top-k features with the best scores will be selected. We have chosen $k = 100$ for our experiments. All experiments have been designed as 10×10 cross-validation experiments. They have been conducted in the TunePareto framework [17].

As an example dataset we utilize the collection of pediatric acute leukemia entities provided by Yeoh et al. [23]. The dataset consists of gene expression profiles of 360 patients which are categorized into 6 risk classes (Table 1). The corresponding gene expression profiles comprise $n = 12558$ measurements.

Table 1. Classwise composition of the leukemia dataset [23].

	$y = 1$	$y = 2$	$y = 3$	$y = 4$	$y = 5$	$y = 6$
Risk class	BCR-ABL	E2A-PBX1	Hyperdip.	MLL	T-ALL	TEL-AML
Samples	15	27	64	20	43	79

4 Results

A summary of the classification performance of the tested multi-class classifier systems can be found in Tables 2 (overall) and 3 (classwise). Without feature selection (noFS) the OaA ensemble outperforms the OaO ensemble in terms of the overall accuracy (OaA: 97.0 %, OaO: 91.2 %). It also achieves the highest mean sensitivity in our experiments (OaA: 97.8 %). Coupled to a TNoM, TNoM_{cw}, PCC feature selection both ensemble types achieve mean overall accuracies over 97.0 %. The best overall accuracy of 97.7 % was achieved for the TNoM-OaO ensemble. The application of the TT slightly decreased the overall accuracy (OaO: 95.6 %, OaA: 96.5 %). The best mean specificity was observed for the TNoM-OaO (99.6 %).

On a classwise level, highest differences in sensitivity can be observed for class $y = 1$. Here, the sensitivity of the feature selecting OaO ensembles is up to 15.4 % higher than the sensitivity of their OaA counterparts. A similar tendency can be seen for class $y = 4$. A maximal sensitivity difference of 11.5 % was observed. For classes $y = 3$ and $y = 6$ higher sensitivities can be seen for the feature selecting OaA ensemble. The highest observed differences in these classes is 5.6 %.

Table 2. Overall accuracy, mean sensitivity, mean specificity (mean ± standard deviation in %) achieved in the 10×10 cross-validation experiments on the leukemia dataset. For each performance measure, the highest values are highlighted.

	Overall accuracy		Mean sensitivty		Mean specificity	
	OaO	OaA	OaO	OaA	OaO	OaA
noFS	91.2 ± 0.8	97.0 ± 0.4	93.1 ± 1.2	$\mathbf{97.8 \pm 0.8}$	98.4 ± 0.1	99.5 ± 0.1
TNoM	$\mathbf{97.7 \pm 0.5}$	97.4 ± 0.8	97.6 ± 1.1	95.1 ± 1.4	$\mathbf{99.6 \pm 0.1}$	99.5 ± 0.2
TNoM_{cw}	97.2 ± 0.6	97.1 ± 0.4	97.1 ± 0.9	94.8 ± 1.0	99.5 ± 0.1	99.4 ± 0.1
PCC	97.5 ± 0.7	97.5 ± 0.7	97.6 ± 1.1	95.6 ± 1.1	99.5 ± 0.1	99.5 ± 0.1
TT	95.6 ± 0.8	96.5 ± 0.8	95.5 ± 0.9	93.6 ± 1.2	99.2 ± 0.1	99.3 ± 0.2

Table 3. Classwise sensitivities and specificities (mean in %) achieved in the 10×10 cross-validation experiments on the leukemia dataset. For each performance measure, the highest values are highlighted.

	$y = 1$		$y = 2$		$y = 3$		$y = 4$		$y = 5$		$y = 6$	
	OaO	OaA	OaO	OaA	OaO	OaA	OaO	OaA	OaO	OaA	OaO	OaA
Sensitivity												
noFS	98.0	**98.7**	96.7	**100.0**	81.2	92.2	95.0	99.5	94.4	97.7	93.5	98.5
TNoM	92.7	78.0	**100.0**	**100.0**	93.6	**96.2**	**100.0**	96.5	**100.0**	**100.0**	99.4	**100.0**
TNoM_{cw}	90.7	75.3	**100.0**	**100.0**	93.3	94.8	**100.0**	98.5	**100.0**	**100.0**	98.5	**100.0**
PCC	93.3	81.3	**100.0**	**100.0**	92.7	96.1	**100.0**	96.0	**100.0**	**100.0**	99.4	**100.0**
TT	86.0	78.0	**100.0**	**100.0**	89.7	95.3	**100.0**	88.5	**100.0**	**100.0**	97.1	99.9
Specificity												
noFS	91.5	98.1	99.9	99.5	99.8	**99.9**	99.6	99.2	**100.0**	**100.0**	99.8	**100.0**
TNoM	98.2	99.4	**100.0**	99.6	99.8	98.3	99.5	99.7	**100.0**	99.8	**100.0**	**100.0**
TNoM_{cw}	97.9	99.0	100.0	100.0	99.6	97.9	99.5	**99.9**	**100.0**	99.8	**100.0**	99.8
PCC	98.2	99.2	**100.0**	99.6	99.8	98.6	99.3	99.6	**100.0**	**100.0**	**100.0**	**100.0**
TT	96.9	**99.8**	**100.0**	99.9	99.0	97.6	99.2	99.4	100.0	99.0	**100.0**	**100.0**

The selection stability of the feature selecting base classifier is summarized in Fig. 1. As a basic measure we apply our stability score presented in Lausser et al. [14]. It is used to characterize the stability of a feature selection strategy in cross-validation experiments. A score near 1 indicates a perfectly stable (fixed) feature selection; a score near zero indicate perfectly instable feature selection (distinct features for each experiment). The scores achieved by the single base classifiers are organized classwise.

In general, all achieved stability scores lie in the range of 0.70 to 0.95. Similar to the classification performance, classwise differences in the selection stability can be observed. For all feature selection strategies, the lowest selection stability of an OaA base classifier is reported for class $y = 1$, which is the class with the lowest number of samples. For this class ($y = 1$) all selection strategies also show the lowest classwise mean stability over all OaO base classifiers. Among the remaining classes, similar observation can be found for class $y = 4$. For this second smallest class, the stability of the OaA base classifier is smaller than for the remaining four classes for TNoM, TNoM_{cw} and PCC. The mean classwise stability of the OaO base classifiers is smaller for all four feature selection strategies. Higher selection stabilities are observed for the larger classes. The highest stabilities of OaA classifiers are observed for the largest class $y = 6$. The highest classwise mean stabilities of OaO base classifiers are observed for class $y = 5$ in three of four cases.

Figure 2 shows the feature selections of the TNoM score. For each class $y \in \mathcal{Y}$, a panel reports on the features of all base classifiers that where trained for recognizing y. The corresponding features are sorted according to the number of

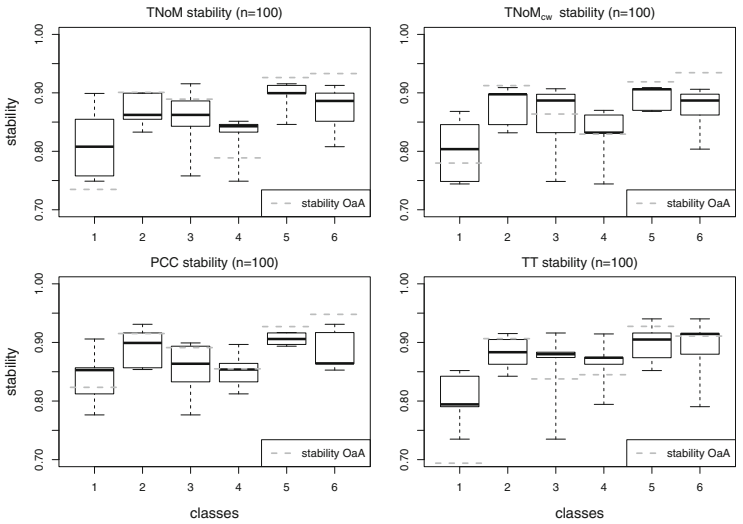

Fig. 1. Selection stabilities of the 10×10 cross validation experiments: The figure gives a comparison of the stability scores of the OaA ensemble and the OaO ensemble. For each of the class y, the dashed lines give the stability score of the corresponding OaA base classifiers. The boxplots summarize the stability scores of the OaO base classifiers that are trained to predict y.

base classifiers $c_{(y,y')}(\mathbf{x})$, $y' \in \mathcal{Y} \setminus \{y\}$ for which they were selected. For each base classifier $c_{(y,y')}(\mathbf{x})$, those features are reported that are selected in at least 70 % of all training runs. The result of the TNoM score can be seen as exemplarily for the other feature selections.

In this classwise view, the classes $y \in \mathcal{Y}$ are represented by $n = 232$ ($y = 5$) up to $n = 368$ ($y = 3$) features. Most of these features are selected in a single two class comparison in the OaO ensembles ($n = 112$ ($y = 5$) up to $n = 249$ ($y = 3$)). The number of selected features decreases with the number of selections per class. At most $n = 21$ features were selected in all five base classifiers $c_{(y,y')}(\mathbf{x})$ of one class y ($y = 5$) in the OaO ensemble. For two classes ($y = 3$ and $y = 4$), no feature was selected in all corresponding OaO base classifiers. The color of a feature indicates its selection frequency in the base classifiers $c_{(y',y'')}(\mathbf{x})$, $y'' \in \mathcal{Y} \setminus \{y'\}$ of the reference class y' of $c_{y,y'}(\mathbf{x})$. Most of these features are selected only once in the OaO base classifiers of y, which indicates a preference for y'.

The selections of the OaA ensemble members $c_{(y,\bar{y})}(\mathbf{x})$ are shown in the rightmost column of each panel. It can be seen that they have a large overlap to the selections of the OaO base classifiers. Only 20.2 % ($y = 5$) up to 32.6 % ($y = 1$) of the OaA selections are not included in the frequently selected features of the OaO base classifiers. For classes $y = 1$, $y = 3$, $y = 5$ no OaA feature was selected four or five times by the corresponding OaO base classifiers. For class $y = 6$, only one OaA feature was selected more than twice by the OaO base classifiers. These observations can also be seen for the features that were

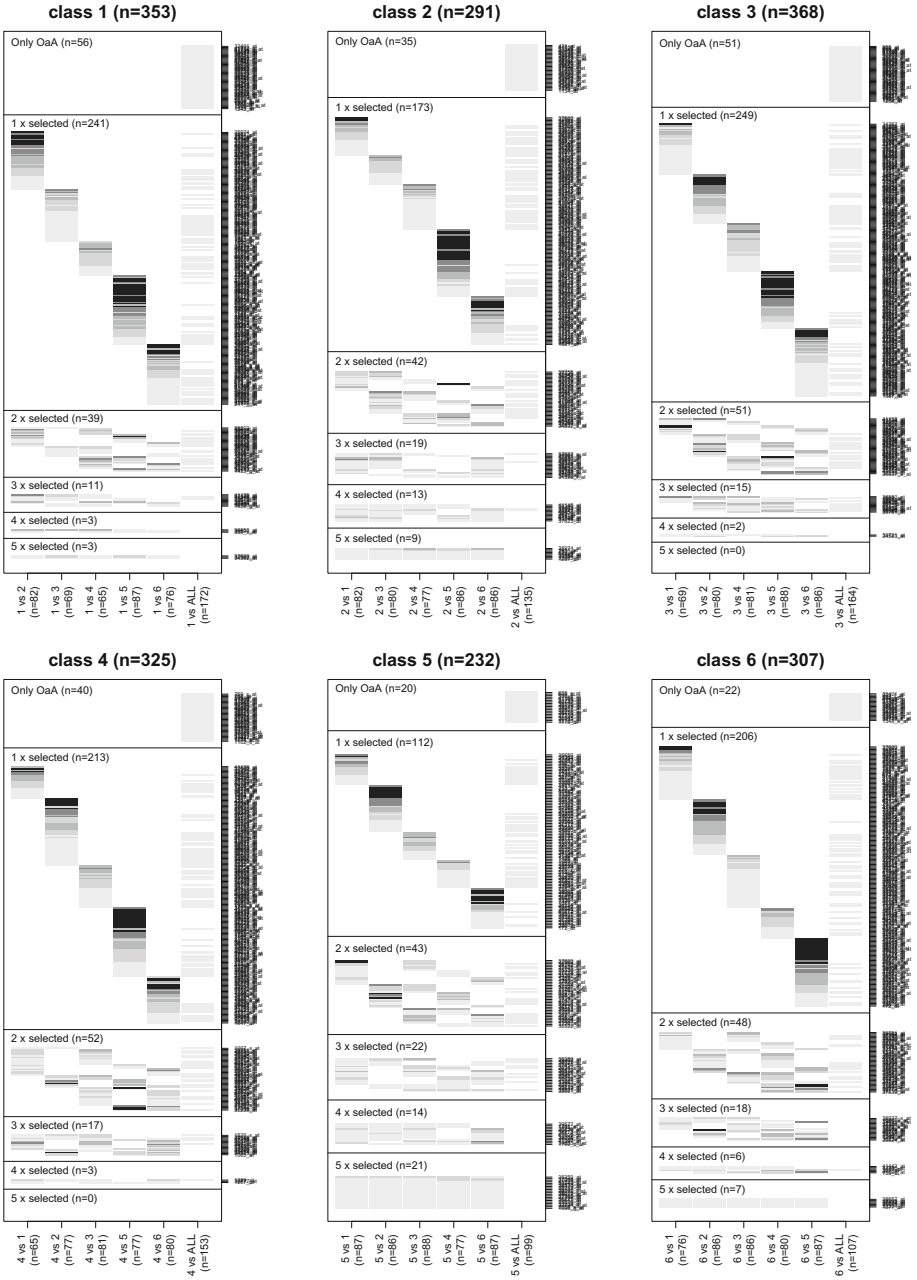

Fig. 2. Overlap of feature signatures (TNoM score): The figure shows the frequently selected features ($\geq$70 %) of the 10×10 cross-validation experiments with the YEOH dataset. The signatures are grouped according to the six classes $y \in \mathcal{Y}$ and sorted according to their selection frequency in the base classifiers $c_{y,y'}(\mathbf{x})$, $y' \in \mathcal{Y} \setminus \{y\}$. The color of the features indicates its selection frequency in the OaO base classifiers $c_{y',y''}(\mathbf{x})$, $y' \in \mathcal{Y} \setminus \{y\}$. A light gray color indicates a selection in one base classifier of y' a black color indicates a selection in all base classifiers of y'.

selected only once by a OaO ensemble. Here, in general the largest number of OaA features can be found. It is interesting to see that the OaA features concentrate on those OaO features, which were selected rarely in y and the corresponding reference classes $y' \in \mathcal{Y} \setminus \{y\}$.

5 Discussion and Conclusion

The interpretability of a decision rule is often as important as the characterization of its generalization performance. It may lead to new hypotheses on the characteristics of the underlying classes. For the analysis of high-dimensional profiles, feature selection is a key concept. A similar benefit can be expected for multi-class decisions [3]. Nevertheless, these architectures allow for large variety of different interactions between feature selection and classification algorithms.

In this work, we analyzed the utility of multi-class architectures of feature selecting base classifiers. They were tested in the task of classifying high-dimensional gene expression profiles. This usually involves classifiers of a low complexity as otherwise over-adaptation would be imminent in these low sample size settings. In our experiments feature selecting base classifiers were able to improve the classification performance of both one against one and one against all ensembles. The improvements were higher for the one against one case, which might be caused by their lower initial performance. However, the overall benefit is not uniformly reflected in all classes. It might be gained on the costs of lower sensitivities for single classes.

The stability analysis shows that classwise differences can also be found on the level of feature signatures. Here, it can be observed that the stability of the signatures depends on the analyzed class (or class combination). This observation can only to some extend be explained by the different class sizes. As the one-against-all training scheme leads uniformly to more imbalanced class ratios than the one-against-one scheme, a common shift (either up or down) would have been expected in all experiments.

Most interestingly, the classwise inspection of the feature signatures allowed an association of the selected features to one of the involved classes. Basically designed for the differentiation of two classes, the signature of a feature selection strategy rather reflects the "neutral" decision boundary than one of the classes. In the context of the analyzed multi-class architectures, we can utilize additional information. As the selection process of the base classifier is not guided by a central partitioning scheme, single features can be included in more than one signature. Our inspection of the classwise selected features of the one against one architecture revealed that a certain proportion of the features cumulates for the analyzed class or one of its counter classes (but not for both). They rather characterize one of the classes than a (pairwise) class difference and might be considered as class attributes. In this line of argumentation, we would have expected an high overlap between these attributes and the features selected by an one against all ensemble. Counterintuitively, this seems not to be the case. The features that were selected in both strategies rather correspond to those

features which are involved in one particular two class decision. They might be seen as a holistic view on the pairwise differences. Nevertheless the base classifiers of the one against all strategy select a certain amount of features, which is not considered by the one-against-one scheme. These features might also contain class attributes but can not be evaluated as mentioned above.

Acknowledgements. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/20072013) under grant agreement n°602783, the German Research Foundation (DFG, SFB 1074 project Z1), and the Federal Ministry of Education and Research (BMBF, Gerontosys II, Forschungskern SyStaR, ID 0315894A and e:Med, SYMBOL-HF, ID 01ZX1407A) all to HAK.

References

1. Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., Yakhini, Z.: Tissue classification with gene expression profiles. J. Comput. Biol. **7**(3–4), 559–583 (2000)
2. Blum, A., Langley, P.: Selection of relevant features and examples in machine learning. Artif. Intell. **97**(1–2), 245–271 (1997)
3. Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., Benítez, J., Herrera, F.: A review of microarray datasets and applied feature selection methods. Inf. Sci. **282**, 111–135 (2014)
4. Chow, S.C., Song, F.: Some thoughts on precision medicine. J. Biom. Biostat. **6**(5), 1–2 (2015)
5. Dietterich, T.G., Bariki, G.: Solving multiclass problems via error-correcting output codes. J. Artif. Intell. Res. **2**, 263–286 (1995)
6. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. In: Vitányi, P.M.B. (ed.) EuroCOLT 1995. LNCS, vol. 904, pp. 23–37. Springer, Heidelberg (1995)
7. Gress, T.M., Kestler, H.A., Lausser, L., Fiedler, L., Sipos, B., Michalski, C.W., Werner, J., Giese, N., Scarpa, A., Buchholz, M.: Differentiation of multiple types of pancreatico-biliary tumors by molecular analysis of clinical specimens. J. Mol. Med. **90**(4), 457–464 (2011)
8. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
9. Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning. Springer, New York (2001)
10. Huang, Y., Suen, C.: A method of combining multiple experts for the recognition of unconstrained handwritten numerals. IEEE Trans. Pattern Anal. Mach. Intell. **17**, 90–94 (1995)
11. Khan, J., Wei, J., Ringner, M., Saal, L., Westermann, F., Berthold, F., Schwab, M., Antonesco, C., Peterson, C., Meltzer, P.: Classification and diagnostic prediction of cancer using gene expression profiling and artificial neural networks. Nat. Med. **6**, 673–679 (2001)
12. Kraus, J., Lausser, L., Kestler, H.A.: Exhaustive k-nearest-neighbour subspace clustering. J. Stat. Comput. Simul. **85**(1), 30–46 (2015)

13. Lattke, R., Lausser, L., Müssel, C., Kestler, H.A.: Detecting ordinal class structures. In: Schwenker, F., Roli, F., Kittler, J. (eds.) MCS 2015. LNCS, vol. 9132, pp. 100–111. Springer, Heidelberg (2015)
14. Lausser, L., Müssel, C., Maucher, M., Kestler, H.A.: Measuring and visualizing the stability of biomarker selection techniques. Comput. Stat. **28**(1), 51–65 (2013)
15. Lausser, L., Schmid, F., Platzer, M., Sillanpää, M.J., Kestler, H.A.: Semantic multi-classifier systems for the analysis of gene expression profiles. Arch. Data Sci. Ser. A (Online First) **1**(1), 1–19 (2016)
16. Lorena, A., de Carvalho, A., Gama, J.: A review on the combination of binary classifiers in multiclass problems. Artif. Intell. Rev. **30**, 19–37 (2008)
17. Müssel, C., Lausser, L., Maucher, M., Kestler, H.A.: Multi-objective parameter selection for classifiers. J. Stat. Softw. **46**(5), 1–27 (2012)
18. Ripley, B.D.: Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge (1996)
19. Saeys, Y., Iñza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics **23**(19), 2507–2517 (2007)
20. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
21. Völkel, G., Lausser, L., Schmid, F., Kraus, J.M., Kestler, H.A.: Sputnik: ad hoc distributed computation. Bioinformatics **31**(8), 1298–1301 (2015)
22. Webb, A.R.: Statistical Pattern Recognition, 2nd edn. Wiley, New York (2002)
23. Yeoh, E.J., Ross, M.E., Shurtleff, S.A., Williams, W.K., Patel, D., Mahfouz, R., Behm, F.G., Raimondi, S.C., Relling, M.V., Patel, A., Cheng, C., Campana, D., Wilkins, D., Zhou, X., Li, J., Liu, H., Pui, C.H., Evans, W.E., Naeve, C., Wong, L., Downing, J.R.: Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell **1**(2), 133–143 (2002)