Chained correlations for feature selection

Ludwig Lausser¹ · Robin Szekely¹ · Hans A. Kestler^{1,2}

Received: 30 June 2019 / Revised: 28 January 2020 / Accepted: 25 April 2020 / Published online: 9 June 2020
© The Author(s) 2020

Abstract

Data-driven algorithms stand and fall with the availability and quality of existing data sources. Both can be limited in high-dimensional settings ($n \gg m$). For example, supervised learning algorithms designed for molecular pheno- or genotyping are restricted to samples of the corresponding diagnostic classes. Samples of other related entities, such as arise in differential diagnosis, are usually not utilized in this learning scheme. Nevertheless, they might provide domain knowledge on the background or context of the original diagnostic task. In this work, we discuss the possibility of incorporating samples of foreign classes in the training of diagnostic classification models that can be related to the task of differential diagnosis. Especially in heterogeneous data collections comprising multiple diagnostic categories, the foreign ones can change the magnitude of available samples. More precisely, we utilize this information for the internal feature selection process of diagnostic models. We propose the use of chained correlations of original and foreign diagnostic classes. This method allows the detection of intermediate foreign classes by evaluating the correlation between class labels and features for each pair of original and foreign categories. Interestingly, this criterion does not require direct comparisons of the initial diagnostic groups and therefore, might be suitable for settings with restricted data access.

Keywords Classification · Feature selection · High-dimensional data · Differential diagnosis

L. Lausser and R. Szekely have contributed equally to this work.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11634-020-00397-5>) contains supplementary material, which is available to authorized users.

✉ Hans A. Kestler
hans.kestler@uni-ulm.de; hans.kestler@leibniz-fli.de

¹ Institute of Medical Systems Biology, Ulm University, 89069 Ulm, Germany

² Leibniz Institute on Aging – Fritz Lipmann Institute, 07745 Jena, Germany

Mathematics Subject Classification 62H30 Classification and discrimination · 62H20 Measures of association (correlation, canonical correlation, etc.) · 68T10 Pattern recognition · 62P10 Applications to biology and medical sciences

1 Introduction

Data mining and machine learning are the key technologies for molecular pheno- and genotyping as required for personalized medicine (Kraus et al. 2018). As those techniques are designed for high-dimensional data, they extend the human capability of extracting and aggregating (molecular) patterns from these profiles of tens of thousands of measurements. Especially in medical applications, the resulting diagnostic models are required to be both accurate and interpretable. Human experts should be able to comprehend and intervene in automated decisions and their consequences. Furthermore, interpretable decision models also aid in generating hypotheses on the molecular causes or mechanisms of disease.

One of the most prominent techniques for improving the interpretability of high-dimensional data is feature selection (Guyon and Elisseeff 2003). It constructs low-dimensional signatures of primary measurements selected from the original high-dimensional profiles. Often these signatures are the basis for all subsequent processing steps and therefore the only interpretation of the final model. The chosen features describe the discriminative outline of the underlying molecular network.

Although they have great potential, the use of data mining and machine learning techniques can have limitations for molecular data. Due to ethical, economic or technical reasons, sample collections are typically limited in size leading to a high contrast of feature and sample numbers ($n \gg m$). This imbalance causes various effects summarized under the issue of the *curse of dimensionality* (Bellman 1957). For example, the linear separability of m data point dichotomies increases with the dimensionality n (Cover 1965). Simultaneously, the Euclidean distances among data points become less distinguishable, which can affect the reliability of neighbourhood networks (Hinneburg et al. 2000). The imbalance mainly influences the possibility of precise and unique parameter estimation due to the high dimensionality of the search space (Bühlmann and van de Geer 2011). For classification models, the complexities of model classes increase with the dimensionality n (Kearns and Vazirani 1994), leading to overfitting and decreased generalization performance in high-dimensional settings.

Here, learning tasks can be improved by two primary strategies. The first one is to focus on classification models that are designed for operating on a relatively small set of samples. It comprises the selection of fast converging learning algorithms and the regulation of model complexity (Vapnik 1998). The second one is the acquisition of additional information and data sources for guiding a training process. This strategy of integrating domain knowledge comprises a broad spectrum of options that can be used for navigating through the search space. Knowledge of the relationships of diagnostic classes can outline their neighborhood (Lattke et al. 2015). Known interactions of univariate features can highlight multivariate processes (Taudien et al. 2016; Lausser et al. 2016a). Identified sources of noise might be counteracted a priori (Lausser et al. 2016b).

Additional data sources can be used to extract domain knowledge within the training process of a classifier. They might comprise additional samples of the original diagnostic classes as well as unlabeled samples or samples from different categories. While samples from the original classes might be seen as a traditional extension of the training data, the other two options lie beyond the scope of supervised learning. Unlabeled samples can be incorporated via partially supervised learning techniques, such as transductive or semi-supervised learning (Vapnik 1998; Chapelle et al. 2010; Lausser et al. 2014). Samples of different classes are utilized, for example, in transfer or multi-task learning approaches (Pan and Yang 2010; Caruana 1997).

In our previous work, we have systematically investigated the potential of transferring specific feature signatures from one molecular classification task to another (Lausser et al. 2018a). We have shown that multi-class classifier systems can utilize this strategy for achieving highly accurate multi-categorical predictions (Lausser et al. 2018b). In this work, we propose a correlation-based feature selection criteria for binary classification tasks that utilize foreign classes for extending their training sets.

2 Methods

In the following, we will view classification as the task of assigning an object to a class selected from a predefined set of distinct classes $y \in \mathcal{Y}$ according to a set of measurements $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^n$

$$c : \mathcal{X} \longrightarrow \mathcal{Y}. \quad (1)$$

We restrict ourselves to binary classification tasks ($|\mathcal{Y}| = 2$). The original classification function c is typically unknown. It has to be reconstructed in a data-driven learning procedure

$$l : \mathcal{C} \times \mathcal{T} \rightarrow c_{\mathcal{T}} \in \mathcal{C}. \quad (2)$$

Here, $\mathcal{C}$ denotes an a priori chosen concept class and $\mathcal{T} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{|\mathcal{T}|}$ a set of labeled training examples. Subscript $\mathcal{T}$ will be omitted for simplicity. In the classical supervised scheme, it is assumed that the training set $\mathcal{T}$ comprises only samples related to the current classification task $\forall i : y_i \in \mathcal{Y}$. This assumption can be restrictive as data collections might also consist of additional samples of other related classes. In the following, we assume our original classification task to be embedded in a larger context comprising additional and distinct classes $\mathcal{Y}' \supset \mathcal{Y}$ as they are collected in multi-class classification tasks. Samples $(\mathbf{x}, y) \in \mathcal{T}$ can be representative for each of these classes $y \in \mathcal{Y}'$. The training algorithm will be allowed to utilize all samples in $\mathcal{T}$. If a subprocess requires only a subset of two classes a and b this will be denoted as

$$\mathcal{T}_{ab} = \{(\mathbf{x}, \mathbb{I}_{[y=b]}) \mid (\mathbf{x}, y) \in \mathcal{T}, y \in \{a, b\}\}. \quad (3)$$

In this case, the original class labels $\{a, b\}$ are replaced by labels $\{0, 1\}$ for simplicity.

In this work, we propose to utilize the full training set $\mathcal{T}$ for the internal feature selection process of the classifier training. The restricted set $\mathcal{T}_{ab}$ will be used for the final adaptation of the classification model. The trained classifier will later on be tested on an analogously restricted validation set $\mathcal{V}_{ab}$.

2.1 Feature selection

Especially in high-dimensional settings, the training procedure of a classification rule can incorporate a feature selection process excluding features that are believed to be noisy, uninformative or even misleading for the adaptation of the classifier. This selection process is typically implemented as a data-driven procedure yielding at the selection of $\hat{n} \leq n$ feature indices

$$\mathbf{i} \in \mathcal{I} = \{\mathbf{i} \in \mathbb{N}^{\hat{n} \leq n} \mid i_{k-1} < i_k, 1 \leq i_k \leq n\}. \quad (4)$$

Subsequent training steps and the final classification model will operate on the reduced feature representation

$$\mathbf{x}^{(i)} = (x^{(i^{(1)})}, \dots, x^{(i^{(\hat{n})})})^T. \quad (5)$$

In the following we concentrate on univariate feature selection strategies. That is each feature is assessed via a quality score $s(i)$ that does not take into account interactions with other candidate features. The selection is based on a vector of quality scores

$$\mathbf{s} = (s(1), \dots, s(n))^T. \quad (6)$$

The top $\hat{n}$ features with the best scores are selected

$$\mathbf{i} = \text{top}_{\hat{n}}(\mathbf{s}) = \{(\dots, i, \dots)^T \mid \text{rk}_{\mathbf{s}}(s(i)) \leq \hat{n}\}, \quad (7)$$

where $\text{rk}_{\mathbf{s}}$ denotes the ranking function of the elements in $\mathbf{s}$.

2.2 Foreign classes in feature selection

As we want to analyze the possibility of utilizing foreign classes for feature selection, the chosen score will not only be evaluated for the original pair of classes a and b but also for other classes $o \in \mathcal{Y} \setminus \{a, b\}$. The corresponding set of scores (obtained for the i th feature) will be denoted by $\mathcal{S}(i)$. The cardinality of $\mathcal{S}(i)$ depends on the chosen strategy for selecting pairs of classes. For aggregating the scores in $\mathcal{S}(i)$, we will utilize following three strategies:

$$s_{\max}(i) = \max(\mathcal{S}(i)), \quad (8)$$

$$s_{\text{mean}}(i) = \text{mean}(\mathcal{S}(i)), \quad (9)$$

$$s_{\min}(i) = \min(\mathcal{S}(i)). \quad (10)$$

We utilize $s_{\text{for}} \in \{s_{\min}, s_{\text{mean}}, s_{\max}\}$ to denote a general foreign feature selection strategy. A classical feature selection strategy is denoted as s_{orig} .

Here, we construct $\mathcal{S}(i)$ from various scores based on the empirical Pearson correlation between an individual feature and class label. For the i th feature and a fixed pair of classes $a, b \in \mathcal{Y}$, it is given by

$$\text{cor}_{\mathcal{T}_{ab}}(i) = \frac{\frac{1}{|\mathcal{T}_{ab}|-1} \sum_{(\mathbf{x}, y) \in \mathcal{T}_{ab}} (x^{(i)} - \bar{x}^{(i)})(y - \bar{y})}{\sqrt{\frac{1}{|\mathcal{T}_{ab}|-1} \sum_{(\mathbf{x}, y) \in \mathcal{T}_{ab}} (x^{(i)} - \bar{x}^{(i)})^2} \sqrt{\frac{1}{|\mathcal{T}_{ab}|-1} \sum_{(\mathbf{x}, y) \in \mathcal{T}_{ab}} (y - \bar{y})^2}}, \quad (11)$$

where $\bar{x}^{(i)}$ denotes the observed mean value of the i th feature and $\bar{y}$ denotes the average class label.

More precisely, we investigate a score based on the Pearson correlations of foreign classes $o \in \mathcal{Y} \setminus \{a, b\}$ to both original classes a and b ,

$$\text{ccor}_{\mathcal{T}_{ao}, \mathcal{T}_{ob}}(i) = \frac{\text{cor}_{\mathcal{T}_{ao}}(i) + \text{cor}_{\mathcal{T}_{ob}}(i)}{2} \in [-1; +1]. \quad (12)$$

In the following we call this score *chained correlation*. The corresponding sets of scores are given by

$$\mathcal{S}_{\text{ccor}}(i) = \left\{ |\text{ccor}_{\mathcal{T}_{ao}, \mathcal{T}_{ob}}(i)| \mid o \in \mathcal{Y} \setminus \{a, b\} \right\}. \quad (13)$$

The score $\text{ccor}_{\mathcal{T}_{ao}, \mathcal{T}_{ob}}(i)$ is high, if foreign class o fulfils two conditions simultaneously and high correlations are obtained when related to classes a and b . Note that o is the second class in $\text{cor}_{\mathcal{T}_{ao}}(i)$ and the first one in $\text{cor}_{\mathcal{T}_{ob}}(i)$. As

$$\forall c, d \in \mathcal{Y}, c \neq d : \text{cor}_{\mathcal{T}_{cd}}(i) = -\text{cor}_{\mathcal{T}_{dc}}(i), \quad (14)$$

class o will lead to high correlations under opposite conditions.

For $\text{cor}_{\mathcal{T}_{ao}}(i)$, high positive correlations are achieved if the values of class a are lower than those of class o . For $\text{cor}_{\mathcal{T}_{ob}}(i)$, the values of class b are required to be higher.

Top scores of $\text{ccor}_{\mathcal{T}_{ao}, \mathcal{T}_{ob}}(i)$ are achieved, if the samples of class o (projected on the i th feature) lie in between the samples of classes a and b . A high value of $\text{ccor}_{\mathcal{T}_{ao}, \mathcal{T}_{ob}}(i)$ implies that the values of class a are lower than those of class b and therefore indicate high values of $\text{cor}_{\mathcal{T}_{ab}}(i)$. The score $\text{ccor}_{\mathcal{T}_{ao}, \mathcal{T}_{ob}}(i)$ can therefore be seen as a surrogate for $\text{cor}_{\mathcal{T}_{ab}}(i)$. As an analogous argumentation can be given for high negative correlations, we have chosen to consider the absolute value $|\text{ccor}_{\mathcal{T}_{ao}, \mathcal{T}_{ob}}(i)|$ in our experiments.

3 Experiments

We evaluate $\mathcal{S}_{cor}$ using the aggregation schemes $s_{for} \in \{s_{max}, s_{mean}, s_{min}\}$ in experiments with 9 multi-class datasets comprising multiple instances ($m \geq 59$, $|\mathcal{Y}| \geq 4$). Each dataset was collected for a specific research question and is therefore analysed independently. This research question can be regarded as the common semantical (and biological) context of the classes $\mathcal{Y}$. A summary of these datasets can be found in Table 1. All datasets consist of gene expression profiles ($n \geq 8740$). Each feature corresponds to the expression level of a mRNA molecule of a biological sample. Within each dataset the biological samples were prepared according to identical laboratory protocols.

We perform experiments utilizing all pairs of classes $a, b \in \mathcal{Y}$ as original classes. For an individual dataset, the experimental setup therefore consists of $\frac{|\mathcal{Y}|(|\mathcal{Y}|-1)}{2}$ settings. As a reference score, the absolute Pearson correlation is chosen $s_{orig} = |cor_{\mathcal{T}_{ab}}|$.

3.1 Empirical characterization of chained correlations

In order to characterize the relations of the foreign aggregation schemes s_{for} and the original score s_{orig} we provide their empirical joint distributions over all n features gained on datasets $d_1 - d_9$. A detailed example for all individual foreign scores $\mathcal{S}_{cor}$ and the foreign aggregation schemes s_{for} is shown for dataset d_8 . Additionally we show examples for high scoring features of s_{for} gained on datasets $d_1 - d_9$.

Table 1 Table of datasets

id:	Description	$ \mathcal{Y} $	n	m	$m_1, \dots, m_{ \mathcal{Y} }$
d_1 :	Leukemia (Haferlach et al. 2010)	18	54613	2096	40, 36, 58, 48, 28, 351, 38, 37, 40, 237, 122, 448, 76, 13, 206, 74, 70, 174
d_2 :	Cancer cell lines (Pfister et al. 2009)	9	54613	174	18, 15, 21, 26, 18, 21, 23, 26, 6
d_3 :	Cancer cell lines (Bittner 2005)	9	54613	1777	366, 85, 343, 264, 130, 225, 280, 41, 43
d_4 :	Cancer cell lines (TCGA 2008)	8	52762	2370	45, 424, 182, 521, 177, 89, 611, 321
d_5 :	Colorectal cancer (Sheffer et al. 2009)	7	22215	381	49, 47, 20, 54, 13, 186, 12
d_6 :	Renal cell cancer (Jones et al. 2005)	6	22215	92	23, 32, 11, 8, 6, 12
d_7 :	Liposarcoma (Gobble et al. 2011)	5	22215	140	40, 11, 17, 20, 52
d_8 :	Alcohol (Kimpel et al. 2007)	5	8740	59	11, 12, 12, 12, 12
d_9 :	Brain tissue (Berchtold et al. 2008)	4	54613	173	43, 48, 43, 49

The number of classes $|\mathcal{Y}|$, features n , samples m and samples per class m_i are reported

3.2 Classification experiments

We also compare the classification accuracies a_{for} gained by the use of foreign aggregation strategies s_{for} to the classification accuracy a_{orig} gained by the original feature selection score s_{orig} . As classification algorithms linear support vector machines (Vapnik 1998) (SVM, $\text{cost} = 1$), random forests (Breiman 2001) (RF, $\text{n tree} = 500$) and k -nearest neighbor classifiers (Fix and Hodges 1951) (k -NN, $k = 3$) were chosen. The accuracies are estimated in stratified 10×10 cross-validation (10×10 CV) experiments (Japkowicz and Shah 2011). All experiments are performed in the TunePareto-Framework (Müssel et al. 2012).

For each multi-class dataset with $|\mathcal{Y}|$ classes we analyze $\frac{|\mathcal{Y}|(|\mathcal{Y}|-1)}{2}$ two-class classification tasks. For each aggregation strategy s_{for} , feature signatures of $\hat{n} \in \{25, 50\}$ features are generated based on the samples of the remaining $|\mathcal{Y}| - 2$ classes.

4 Results

4.1 Empirical joint distributions of chained correlations

Figure 1 shows an example of the individual and aggregated chained correlations for d_8 . An overview for all datasets is given in Fig. 2. The differences $s_{\text{orig}} - s_{\text{for}}$ are shown. The highest scores for s_{for} occur with s_{max} . Over all datasets, the average ranges are $s_{\text{max}} \in [0.00, 0.84]$, $s_{\text{mean}} \in [0.00, 0.74]$, $s_{\text{min}} \in [0.00, 0.67]$.

Fig. 1 Comparison of the original score s_{orig} and the foreign scores s_{for} for dataset d_8 . Each scatterplot shows the profile of all $n = 8740$ features of d_8 . $a = y_1$ and $b = y_5$ were chosen as original classes. The remaining classes were used as foreign classes $o_1 = y_2, o_2 = y_3, o_3 = y_4$. The first row shows the individual foreign scores s_o for $o_1 - o_3$. The second row gives the aggregated scores $s_{\text{min}}, s_{\text{mean}}, s_{\text{max}}$

(a) score differences: $s_{\text{orig}} - s_{\text{for}}$

(b) fraction of features

Fig. 2 Differences of the original score and the foreign scores $s_{\text{orig}} - s_{\text{for}}$. The figure gives the empirical differences between the original score s_{orig} and the aggregated foreign scores s_{min} , s_{mean} , s_{max} observed for each dataset $d_1 - d_9$. The results for the whole feature profiles and for all pairs of original classes $a, b \in \mathcal{Y}$ are aggregated for each dataset. The foreign scores s_{for} were discretized in intervals of width 0.1. Panel (a) gives the ranges of differences in dependency of s_{for} . Panel (b) can be seen as a histogram showing the percentage of features in a specific interval of s_{for} . Each bar is split in the percentage of features that over- or underestimated the original score s_{orig}

Fig. 3 Visualization of expression values of high scoring features for $d_1, \dots, d_9$ using the aggregation strategy s_{max} . For every dataset, the original classes $y_a, y_b \in \mathcal{Y}$ are shown at the top, the value of s_{orig} is shown to the right. For every foreign class $y_c \in \mathcal{Y} \setminus \{y_a, y_b\}$, the expression values of this single class are shown on an axis and the value of s_{for} to the right. The values of s_{for} are sorted in descending order. For each dataset, the limits (minimum and maximum) of the expression values are shown at the bottom

Panel (a) shows boxplots of these differences in fixed intervals of s_{for} and panel (b) the corresponding fraction of features. It can be observed that for the majority of features s_{orig} is underestimated by the use of s_{for} . We quantify this observation by counting all features that over- and underestimate s_{orig} . Over all datasets, s_{orig} is underestimated by s_{min} in 98.52% of all cases and overestimated in 1.47%. For s_{mean} , an underestimation was observed for 92.24% of all experiments. s_{orig} was overestimated in 7.76%. The aggregation scheme s_{max} leads to an underestimation of s_{orig} in 81.46% of all cases and to an overestimation in 18.54%.

Figure 3 shows the expression values of high scoring features for s_{max} for each dataset $d_1 - d_9$. It can be seen that for all these features an underestimation of s_{orig} holds true. Differences from s_{max} to s_{orig} up to 0.21 (d_7) can be observed. In mean

over all datasets, these differences are 0.11. Comparable figures for $s_{\min}$ and s_{mean} can be found in the supplement.

4.2 Evaluation of 10×10 cross-validation experiments

A comparison of the accuracies achieved in the 10×10 CV experiments is given in Fig. 4. All results are shown in triplets indicating the number of wins, ties and losses (w/t/l) for a specific dataset. Panel A shows the result for $\hat{n} = 25$ features. Over all datasets, best results were achieved for $s_{\max}$. For the SVM, $s_{\max}$ achieved better results for 37.71% of all experiments ($t=24.74\%/l=37.55\%$). For the 3NN it gained better results for 47.37% ($t=14.53\%/l=38.10\%$). For RF, it was better in 40.07% of all cases ($t=26.14\%/l=33.79\%$).

Comparable results were observed for s_{mean} . Here the SVM based on s_{mean} achieved higher accuracies in 38.43% of all experiments ($t=23.55\%/l=38.02\%$). The 3NN

Fig. 4 Panels showing how classifiers using s_{for} compare to the same classifiers using s_{orig} in 10×10 cross-validation experiments. For $\hat{n} = 25$ (panel A) and $\hat{n} = 50$ (panel B), stacked barplots are depicted according to the aggregation strategies $s_{\text{for}} \in \{s_{\min}, s_{\text{mean}}, s_{\max}\}$. Rows indicate the datasets $d_1, \dots, d_9$ and columns the classification algorithms SVM, 3NN and RF. Each individual barplot shows how often a classifier utilizing s_{for} gains a better, equal or lower accuracy a_{for} compared to its counterpart a_{orig} utilizing s_{orig} . Under each classification algorithm, the mean values over all comparisons are presented

gained better results in 36.93% of all cases ($t = 11.35\%/l = 51.72\%$). The RF was better in 32.01% of all settings ($t = 26.05\%/l = 41.94\%$).

The lowest performance was achieved by $s_{\min}$. For the SVM, $s_{\min}$ won 31.12% comparisons ($t = 23.82\%/l = 45.05\%$). For 3NN, it achieved better results for 35.72% of all cases ($t = 8.90\%/l = 55.39\%$). For RF, 31.07% wins ($t = 22.78\%/l = 46.14\%$) were observed.

In Panel B the results for $\hat{n} = 50$ features are given. Here, the ranking of the foreign scores is similar to the ranking observed for $\hat{n} = 25$ features. Best results were gained for $s_{\max}$. The corresponding SVM was better for 40.92% of all experiments ($t = 26.10\%/l = 32.98\%$), 3NN gained better results in 48.42% ($t = 16.53\%/l = 35.05\%$) and the RF outperformed its counterpart in 40.48% ($t = 25.69\%/l = 33.83\%$).

Coupled to s_{mean} , the SVM achieved better performance in 38.38% of all experiments ($t = 26.39\%/l = 35.22\%$). The 3NN shows higher accuracies in 38.62% of all cases ($t = 13.61\%/l = 47.77\%$). The RF had higher accuracies in 34.49% of all settings ($t = 24.35\%/l = 41.17\%$).

For $s_{\min}$ again the lowest performances were achieved. For SVM, it achieved better results for 35.50% of all cases ($t = 26.16\%/l = 38.34\%$). For 3NN, 28.22% wins ($t = 15.15\%/l = 56.63\%$) were observed. For the RF, the $s_{\min}$ won 32.06% comparisons ($t = 23.80\%/l = 44.14\%$).

5 Discussion and conclusion

In this work, we analyzed the use of chained correlations for incorporating foreign classes in the feature selection processes of binary diagnostic tasks. Here, samples of the original diagnostic classes are related to those of a foreign class for each feature. The chained correlations link the first original class to the foreign class and the foreign class to the second original class. A high score indicates a central position of the foreign class and almost disjoint positions of the original ones. A chained correlation typically underestimates the correlation gained for the original classes. The first one, therefore, might be used as a surrogate marker for the second one.

However, it might be unclear which foreign classes should be used. In our experiments, the original classes, as well as the foreign classes, were selected from the same multi-class data collection. The underlying biological samples were collected for a specific scientific purpose and preprocessed according to identical laboratory protocols. They are therefore comparable and share common semantical (or biological) context. These constraints might be too strict as we mainly screen for (feature-wise) central classes that imply well separable categories. Only the original diagnostic classes require an identical preprocessing to guarantee the correctness of implications from high chained correlations. However, we assume that both constraints increase the probability of detecting foreign feature-wise central classes.

We proposed three scores for combining chained correlations of multiple foreign classes. The minimum, mean or maximum score are discussed which consider a different number of foreign classes. While a high minimum score requires all foreign classes to lie in between the original classes, only one central foreign class is required for a high maximum score. The minimum criterion, therefore, models the original classes as

outlying classes implying an extreme margin on the selected feature dimensions. The maximum criterion only enforces a large margin between the original classes without implications on the other foreign classes.

The design of the proposed combination schemes implies a natural order of the scores achievable for the individual features leading to more pronounced right-skewness of the minimum criterion than for the maximum criterion. In general, all combination schemes underestimate the real correlations between the feature values and the class labels of the original classes. Overestimation only occurs in rare cases. Higher biases and variances can be observed for the minimum strategy than for the maximum approach. Most top scores resulted in high original correlations for all combination schemes. Nevertheless, this definition might be dataset dependent.

Although individual top features for the minimum strategy look more promising than those of the maximum policy, better classification results were obtained for the later one. A reason might be the rarity of features that achieve high scores for the minimum strategy. Such features were mainly not observed for all class combinations. As we have chosen to construct signatures of the top k candidates also inferior features entered the selection. A threshold on the scores or other cut-off strategies might be more suitable (Yu and Príncipe 2019; François et al. 2007). High scores for the maximum approach were available for almost all class combinations leading to an overall better result. This reasoning could be a seeding point for the design of more sophisticated combination strategies implementing concepts from multi-objective optimization (Deb 2001), social choice theory (Chevalyre et al. 2007) or rank aggregation (Burkovski et al. 2014).

Acknowledgements Open Access funding provided by Projekt DEAL. The research leading to these results has received funding from the German Research Foundation (DFG, SFB 1074 Project Z1, and GRK 2254 HEIST), and the Federal Ministry of Education and Research (BMBF, e:Med, conFirm, id 01ZX1708C) all to HAK.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Bellman R (1957) Dynamic programming. Princeton University Press, Princeton
- Berchtold NC, Cribbs DH, Coleman PD, Rogers J, Head E, Kim R, Beach T, Miller C, Troncoso J, Trojanowski JQ, Zielke HR, Cotman CW (2008) Gene expression changes in the course of normal brain aging are sexually dimorphic. *Proc Natl Acad Sci USA* 105(40):15605–15610
- Bittner M (2005) Expression project for oncology (expO). National Center for Biotechnology Information
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
- Bühlmann P, van de Geer S (2011) Statistics for high-dimensional data. Springer Series in Statistics, Springer, Heidelberg

- Burkovski A, Lausser L, Kraus J, Kestler H (2014) Rank aggregation for candidate gene identification, machine learning and knowledge discovery. In: Spiliopoulou M, Schmidt-Thieme L, Janning R (eds) *Data analysis*. Springer International Publishing, Cham, pp 285–293
- Caruana R (1997) Multitask learning. *Mach Learn* 28(1):41–75
- Chapelle O, Schölkopf B, Zien A (2010) *Semi-supervised learning*, 1st edn. The MIT Press, Cambridge
- Chevalere Y, Endriss U, Lang J, Maudet N (2007) A short introduction to computational social choice. In: van Leeuwen J, Italiano G, van der Hoek W, Meinel C, Sack H, Plášil F (eds) *SOFSEM 2007: theory and practice of computer science*. Springer, Berlin, Heidelberg, pp 51–69
- Cover TM (1965) Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans Electron Comput* 14(3):326–334
- Deb K (2001) *Multi-objective optimization using evolutionary algorithms*. Wiley, Hoboken
- Fix E, Hodges JL (1951) *Discriminatory analysis: nonparametric discrimination: consistency properties*. In: Technical reports project 21-49-004, report number 4. USAF School of Aviation Medicine, Randolph Field, Texas
- François D, Rossi F, Wertz V, Verleysen M (2007) Resampling methods for parameter-free and robust feature selection with mutual information. *Neurocomputing* 70(7–9):1276–1288
- Gobble RM, Qin LX, Brill ER, Angeles CV, Ugras S, O'Connor RB, Moraco NH, DeCarolis PL, Antonescu C, Singer S (2011) Expression profiling of liposarcoma yields a multigene predictor of patient outcome and identifies genes that contribute to liposarcomagenesis. *Cancer Res* 71(7):2697–2705
- Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. *J Mach Learn Res* 3(Mar):1157–1182
- Haeflrich T, Kohlmann A, Wiczorek L, Basso G, Kronnie GT, Béné MC, Vos JD, Hernández JM, Hofmann WK, Mills KI, Gilkes A, Chiaretti S, Shurtleff SA, Kipps TJ, Rassenti LZ, Yeoh AE, Papenhausen PR, Liu WM, Williams PM, Foà R (2010) Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: report from the international microarray innovations in leukemia study group. *J Clin Oncol* 28(15):2529–2537
- Hinneburg A, Aggarwal C, Keim D (2000) What is the nearest neighbor in high dimensional spaces? In: *Proceedings of the 26th international conference on very large data bases*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp 506–515
- Japkowicz N, Shah M (2011) *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, New York
- Jones J, Otu H, Spentzos D, Kolia S, Inan M, Beecken WD, Fellbaum C, Gu X, Joseph M, Pantuck AJ, Jonas D, Libermann TA (2005) Gene signatures of progression and metastasis in renal cell cancer. *Clin Cancer Res* 11(16):5730–5739
- Kearns M, Vazirani U (1994) *An introduction to computational learning theory*. MIT Press, Cambridge
- Kimpel MW, Strother WN, McClintick JN, Carr LG, Liang T, Edenberg HJ, McBride WJ (2007) Functional gene expression differences between inbred alcohol-preferring and non-preferring rats in five brain regions. *Alcohol* 41(2):95–132
- Kraus J, Lausser L, Kuhn P, Jobst F, Bock M, Halanke C, Hummel M, Heuschmann P, Kestler HA (2018) Big data and precision medicine: challenges and strategies with healthcare data. *Int J Data Sci Anal* 6(3):241–249
- Lattke R, Lausser L, Müssel C, Kestler HA (2015) Detecting ordinal class structures. In: Schwenker F, Roli F, Kittler J (eds) *Multiple classifier systems, MCS 2015. Lecture notes in computer science*, vol 9132, pp 100–111. Springer, Cham
- Lausser L, Schmid F, Schmid M, Kestler HA (2014) Unlabeling data can improve classification accuracy. *Pattern Recogn Lett* 37:15–23
- Lausser L, Schmid F, Platzer M, Sillanpää MJ, Kestler HA (2016a) Semantic multi-classifier systems for the analysis of gene expression profiles. *Arch Data Sci Ser A* 1(1):1–19 (Online First)
- Lausser L, Schmid F, Schirra LR, Wilhelm A, Kestler H (2016b) Rank-based classifiers for extremely high-dimensional gene expression data. *Adv Data Anal Classif* 12:1–20
- Lausser L, Szekely R, Kessler V, Schwenker F, Kestler HA (2018a) Selecting features from foreign classes. In: Pancioni L, Schwenker F, Trentin E (eds) *Artificial neural networks in pattern recognition*. Springer International Publishing, Cham, pp 66–77
- Lausser L, Szekely R, Schirra LR, Kestler HA (2018b) The influence of multi-class feature selection on the prediction of diagnostic phenotypes. *Neural Process Lett* 48(2):863–880
- Müssel C, Lausser L, Maucher M, Kestler HA (2012) Multi-objective parameter selection for classifiers. *J Stat Softw* 46(5):1–27

- Pan SJ, Yang Q (2010) A survey on transfer learning. *IEEE Trans Knowl Data Eng* 22(10):1345–1359
- Pfister TD, Reinhold WC, Agama K, Gupta S, Khin SA, Kinders RJ, Parchment RE, Tomaszewski JE, Doroshow JH, Pommier Y (2009) Topoisomerase I levels in the NCI-60 cancer cell line panel determined by validated ELISA and microarray analysis and correlation with indenoisoquinoline sensitivity. *Mol Cancer Ther* 8(7):1878–1884
- Sheffer M, Bacolod MD, Zuk O, Giardina SF, Pincas H, Barany F, Paty PB, Gerald WL, Notterman DA, Domany E (2009) Association of survival and disease progression with chromosomal instability: a genomic exploration of colorectal cancer. *Proc Nat Acad Sci* 106(17):7131–7136
- Taudien S, Lausser L, Giamarellos-Bourboulis EJ, Sponholz C, F S, Felder M, Schirra LR, Schmid F, Gogos C, S G, Petersen BS, Franke A, Lieb W, Huse K, Zipfel PF, Kurzai O, Moepps B, Gierschik P, Bauer M, Scherag A, Kestler HA, Platzer M (2016) Genetic factors of the disease course after sepsis: rare deleterious variants are predictive. *EBioMedicine* 12:227–238
- Vapnik VN (1998) *Statistical learning theory*. Wiley, New York
- Yu S, Príncipe J (2019) Simple stopping criteria for information theoretic feature selection. *Entropy* 21(1):99

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.