Detecting Ordinal Subcascades

Ludwig Lausser¹ · Lisa M. Schäfer¹ · Silke D. Kühlwein¹ · Angelika M. R. Kestler³ · Hans A. Kestler^{1,2}

Accepted: 1 October 2020 / Published online: 19 October 2020
© The Author(s) 2020

Abstract

Ordinal classifier cascades are constrained by a hypothesised order of the semantic class labels of a dataset. This order determines the overall structure of the decision regions in feature space. Assuming the correct order on these class labels will allow a high generalisation performance, while an incorrect one will lead to diminished results. In this way ordinal classifier systems can facilitate explorative data analysis allowing to screen for potential candidate orders of the class labels. Previously, we have shown that screening is possible for total orders of all class labels. However, as datasets might comprise samples of ordinal as well as non-ordinal classes, the assumption of a total ordering might be not appropriate. An analysis of subsets of classes is required to detect such hidden ordinal substructures. In this work, we devise a novel screening procedure for exhaustive evaluations of all order permutations of all subsets of classes by bounding the number of enumerations we have to examine. Experiments with multi-class data from diverse applications revealed ordinal substructures that generate new and support known relations.

Keywords Ordinal classification · Classifier cascades · Error bounds · Subsets · Supersets

1 Introduction

Extensive data collections are considered valuable resources for hypothesis generation as well as theory building and confirmation. Providing samples of major concepts or categories, they can be seen as the foundation of data-driven learning and reasoning. Classical machine learning techniques focus on the discrimination of individual concepts. Depending on the

Ludwig Lausser and Lisa M. Schäfer have contributed equally to this work.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11063-020-10362-0>) contains supplementary material, which is available to authorized users.

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

¹ Institute of Medical Systems Biology, Albert-Einstein-Allee 11, 89081 Ulm, Germany

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

³ Internal Medicine I, University Hospital Ulm, 89069 Ulm, Germany

chosen model type, they allow an interpretation of the underlying discrimination rule and the generation of hypotheses on its intrinsic characteristics [10,27,45]. For example, features with a high impact on a decision boundary can be reported when screening for potential causes of class differences [21,34,36]. Dense regions can be extracted when the definition of prototypic cases is of interest [16]. Combined with external domain knowledge classification models can also give hints to higher-level processes involved [35,47].

Hypotheses on the relations of the categories are only rarely provided [9,22,52]. Nevertheless, they have tremendous explanatory potential as they link different concepts collected for a specific subject. This information is complementary to the intrinsic characteristics mentioned above. Together they allow a holistic overview of the matter. One type of classifier that utilises interclass relations are ordinal classifiers [11]. An ordinal classifier relies on an ordinal relationship of the kind $a < b < c$, which is assumed for the semantic concepts (class labels) of a dataset but not guaranteed for its feature representation. The assumed semantic ordering guides the training process of an ordinal classifier. If it is reflected in the feature space, the ordinal classifier can achieve high sensitivities. Otherwise, it will show a diminished performance [33].

A dependency on a predefined class order can be established in different ways. The overall structure, as well as the training algorithm of a classification model, can be modified. The first one is most prominently used for ordinal adaptations of hierarchical multi-class architectures based on sequences of binary classifiers [40]. The order of these sequences can be designed to reflect assumed class orders. Examples can be ordinal classifier cascades [18], which are modifications of general decision lists [45], ordinal versions [23] of nested dichotomies [19] or systems based on directed acyclic graphs [44]. More specific modifications can be applied to different types of base classifiers [11,12]. A training algorithm can be adapted for ordinal classification by utilising specific performance measures [50] or cost-sensitive base classifiers [31,39].

Here, we now do not assume any order on the class labels but instead, screen for potential candidate orders on the semantic level. Nevertheless, the assumption of a total order of classes might not be appropriate in the presence of non-ordinal categories. We propose a novel screening procedure for generating hypotheses on ordinal relations among subsets of these categories. It extensively replaces *de novo* modelling by memoisation techniques bringing exhaustive evaluation into range. We utilise ordinal classifier cascades to screen through all class combinations and highlight those sub-cascades that achieve a predefined minimal class-wise sensitivity which is more difficult for longer sub-cascades. We, therefore, focus on those that cover as many classes as possible. We found that accurate sub-cascades can also result in a compact assignment of the remaining categories and conclude that they hence allow for hypothesis generation on these remaining non-ordinal classes.

The remaining article is organised as follows. Section 2 provides the underlying notation and concepts of ordinal classification. In Sect. 3, we describe our screening method. We outline the design of our experiments and results in Sects. 4 and 5 and discuss them in Sect. 6.

2 Methods

The basic building block of our screening experiments is the training and evaluation of classification models, the classifiers. In its basic version, a classifier is a function designed

for categorising an object into one out of $|\mathcal{Y}|$ predefined and fixed classes $y \in \mathcal{Y}$

$$c: \mathbb{R}^n \longrightarrow \mathcal{Y}. \quad (1)$$

The corresponding classification task is typically named binary ($|\mathcal{Y}| = 2$) or multi-class ($|\mathcal{Y}| > 2$). The object is presented as a vector $\mathbf{x} = (x_1, \dots, x_n)^T$ of measurements which we assume to be embedded in the n -dimensional real-valued space $\mathbb{R}^n$. A classifier is adapted to a classification task via a set of labeled training examples $\mathcal{T} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{|\mathcal{T}|}$, $y_i \in \mathcal{Y}$. The training algorithm and the concept class of the classifier is chosen a priori.

The performance of a classifier is typically evaluated on an independent set of validation samples $\mathcal{V} = \{(\mathbf{x}'_i, y'_i)\}_{i=1}^{|\mathcal{V}|}$. Although a classifier is trained to predict a predefined set of classes $\mathcal{Y}$, it is applied in altered scenarios in which a subset $y'_i \in \mathcal{Y}' \subset \mathcal{Y}$ or a superset $y'_i \in \mathcal{Y}' \supset \mathcal{Y}$ of classes might be used. We will therefore define our performance measures in dependency on the label set of interest.

All performance measures will be based on conditional prediction rates of type

$$p_c(y' | \mathcal{X}_y) = \frac{1}{|\mathcal{X}_y|} \sum_{x \in \mathcal{X}_y} \mathbb{I}_{[c(x)=y']}. \quad (2)$$

The symbol $\mathbb{I}_{[\cdot]}$ denotes the indicator function and $\mathcal{X}_y$ denotes a set of samples of class y . Other (re-)sampling schemes might be applied but will not change the theoretical characteristics discussed in this work. Three types of conditional prediction rates can be distinguished:

$$\text{sensitivities} \quad \text{if} \quad y = y' \quad \text{and} \quad y, y' \in \mathcal{Y} \quad (3)$$

$$\text{confusions} \quad \text{if} \quad y \neq y' \quad \text{and} \quad y, y' \in \mathcal{Y} \quad (4)$$

$$\text{external rates} \quad \text{if} \quad y \notin \mathcal{Y} \quad \text{and} \quad y' \in \mathcal{Y}. \quad (5)$$

While (class-wise) sensitivities and confusions occur for the learned set of classes $\mathcal{Y}$, the external rates describe the classifiers behaviour on samples of foreign classes $y \in \mathcal{Y}' \setminus \mathcal{Y}$.

A standard performance measure for classification tasks is the empirical accuracy. In dependency on a label set $\mathcal{Y}$ it can be formulated as

$$\text{Acc}_c(\mathcal{Y}) = \frac{\sum_{y \in \mathcal{Y}} p_c(y | \mathcal{X}_y) |\mathcal{X}_y|}{\sum_{y \in \mathcal{Y}} |\mathcal{X}_y|}. \quad (6)$$

In its standard version, the empirical accuracy is applied to the full set of classes but it might also be applied to a sub- or superset of classes. In the following, we focus on the minimal class-wise sensitivity as a quality measure for a multi-class classifier system

$$\text{Smin}_c(\mathcal{Y}) = \min_{y \in \mathcal{Y}} p_c(y | \mathcal{X}_y). \quad (7)$$

It can be seen as a lower bound on the overall accuracy

$$\text{Acc}_c(\mathcal{Y}) = \frac{\sum_{y \in \mathcal{Y}} p_c(y | \mathcal{X}_y) |\mathcal{X}_y|}{\sum_{y \in \mathcal{Y}} |\mathcal{X}_y|} \quad (8)$$

$$\geq \frac{\sum_{y \in \mathcal{Y}} \min_{y' \in \mathcal{Y}} p_c(y' | \mathcal{X}_{y'}) |\mathcal{X}_y|}{\sum_{y \in \mathcal{Y}} |\mathcal{X}_y|} \quad (9)$$

$$= \min_{y' \in \mathcal{Y}} p_c(y' | \mathcal{X}_{y'}) \quad (10)$$

$$= \text{Smin}_c(\mathcal{Y}). \quad (11)$$

In the context of analysing subsets of classes, $S_{\min_c}$ shows a monotonicity on the number of classes. For a non empty subset of classes

$$\mathcal{Y}' \subseteq \mathcal{Y}, \mathcal{Y}' \neq \emptyset : S_{\min_c}(\mathcal{Y}') \geq S_{\min_c}(\mathcal{Y}). \quad (12)$$

For external classes $y' \notin \mathcal{Y}$, we define the percentage of the most frequently predicted class label $y \in \mathcal{Y}$ as the purity of y'

$$\text{pur}_c(y') = \max_{y \in \mathcal{Y}} p_c(y \mid \mathcal{X}_{y'}). \quad (13)$$

2.1 Ordinal Classification

Ordinal classification is a multi-class classification task ($|\mathcal{Y}| > 2$), in which we hypothesise that the classes are related via an ordinal relationship

$$y_{(1)} < \dots < y_{(|\mathcal{Y}|)}, \quad (14)$$

where the subscript $y_{(i)}$ indicates the position of the class within the ordering (i th class). The symbol $<$ indicates that this ordering is only known (or assumed) for the verbal concepts. Its reflection in feature space cannot be guaranteed. Nevertheless, ordinal classifiers are constrained to this ordering and try to identify an embedding that reflects this ordinality. If such an embedding can be found, the chosen feature representation might be seen as evidence for the hypothesised ordinal relationship; otherwise the ordinal classifier will suffer from a decreased generalisation performance (Fig. 1). The classifier will typically not be able to predict all classes of $\mathcal{Y}$. Its performance might be measured by the minimal class-wise sensitivity $S_{\min_c}(\mathcal{Y})$. The susceptibility of ordinal classifiers to incorrect class orders is therefore an important characteristic for the screening processes, in which we try to select reflected/relevant class orders among all possible class orders.

2.1.1 Ordinal Classifier Cascades

Our screening procedure is based on ordinal classifier cascades, which can be seen as ensemble multi-class classifiers

$$h : \mathbb{R}^n \longrightarrow \mathcal{Y} \quad (15)$$

based on an ensemble $\mathcal{E} = \{c_{(i)}\}_{i=1}^{|\mathcal{Y}|-1}$ of $|\mathcal{Y}| - 1$ binary base classifiers

$$c_{(i)} : \mathbb{R}^n \longrightarrow \{y_{(i)}, y_{(i+1)}\} \quad (16)$$

designed for predicting a pair of two consecutive classes within the assumed class ordering. We utilise pairwise training in the following meaning that each base classifier $c_{(i)}$ is trained on the samples of two consecutive classes

$$\mathcal{T}_{(i)} = \{(\mathbf{x}, y) \mid (\mathbf{x}, y) \in \mathcal{T}, y \in \{y_{(i)}, y_{(i+1)}\}\}. \quad (17)$$

The cascade itself can be seen as an untrainable fusion architecture, which evaluates the base classifiers sequentially according to the assumed class order

$$h(\mathbf{x}) = \begin{cases} y_{(1)} & \text{if } c_{(1)}(\mathbf{x}) = y_{(1)} \\ y_{(i)} & \text{if } \left(\bigwedge_{j < i} (c_{(j)}(\mathbf{x}) = y_{(j+1)}) \right) \wedge c_{(i)}(\mathbf{x}) = y_{(i)} \\ y_{(|\mathcal{Y}|)} & \text{else.} \end{cases} \quad (18)$$

(a) original order:

$$y_1 \prec y_2 \prec y_3 \prec y_4 \prec y_5 \prec y_6 \prec y_7 \prec y_8$$

(c) wrong order:

$$y_1 \prec y_2 \prec y_3 \prec y_5 \prec y_4 \prec y_6 \prec y_7 \prec y_8$$

(b) inverse order:

$$y_8 \prec y_7 \prec y_6 \prec y_5 \prec y_4 \prec y_3 \prec y_2 \prec y_1$$

(d) wrong order:

$$y_1 \prec y_2 \prec y_3 \prec y_4 \prec y_8 \prec y_7 \prec y_6 \prec y_5$$

$$y_1: \bullet \quad y_2: \bullet \quad y_3: \bullet \quad y_4: \bullet \quad y_5: \circ \quad y_6: \circ \quad y_7: \circ \quad y_8: \circ$$

Fig. 1 Influence of class order on ordinal classifier cascades. The figure shows four ordinal classifier cascades trained under different assumptions of the class order. Panel **a**: A class order that is reflected by the data. The inverse order is shown in Panel **b**. Note, that the decision regions differ in both cases. Panels **c**, **d**: Examples for class orders that are not reflected by the data. Here, decision regions of individual classes can be empty or even do not exist

2.1.2 Pairwise Training of Base Classifiers

The pairwise training of the base classifiers improves the training and the evaluation time of ordinal classifier cascades when multiple class orders are tested. As the pairwise training of a base classifier $c_{(i)}$ does not take into account any foreign classes $\mathcal{Y} \setminus \{y_{(i)}, y_{(i+1)}\}$, a classifier

Table 1 Scheme of predicted class labels

Classifier $c_{i,j}(\mathbf{x})$	Validation samples $(\mathbf{x}, y) \in \mathcal{V}$				
	$\mathbf{x} \in \mathcal{X}_{y_1}$	...	$\mathbf{x} \in \mathcal{X}_{y_i}$	...	$\mathbf{x} \in \mathcal{X}_{y_{ \mathcal{Y} }}$
	$\mathbf{x}_1, \dots, \mathbf{x}_{ \mathcal{X}_{y_1} }$	...	$\mathbf{x}_1, \dots, \mathbf{x}_{ \mathcal{X}_{y_k} }$	...	$\mathbf{x}_1, \dots, \mathbf{x}_{ \mathcal{X}_{y_{ \mathcal{Y} }} }$
...	...	...	...	...	...
y_i, y_j	...	...	$c_{i,j}(\mathbf{x}_1), \dots, c_{i,j}(\mathbf{x}_{ \mathcal{X}_{y_k} })$	...	...
...	...	...	...	...	...

The table provides the predicted class labels of a set of pairwise trained base classifiers $c_{i,j}$ for a validation set $\mathcal{V} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{|\mathcal{V}|}$. As pairwise trained base classifiers do not depend on the chosen class order, all base classifiers and predictions have to be calculated only once. The table is of size $\binom{|\mathcal{Y}|}{2} \times |\mathcal{V}|$ for classes $\mathcal{Y}$. The scheme of the table can be extended for arbitrary resampling strategies

designed for predicting classes y_i and y_j

$$c_{i,j} : \mathbb{R}^n \longrightarrow \{y_i, y_j\} \quad (19)$$

will be identical for all class orders. The classifier $c_{i,j}$ is not based on the class order. The screening process through all class orderings can therefore be based on a common set of $\binom{|\mathcal{Y}|}{2}$ base classifiers and their predictions, which can be memoized beforehand. We propose to construct a prediction table as shown in Table 1, which allows the evaluation of all cascades without training and evaluation of the base classifiers.

2.1.3 Error Bounds

As a consequence of their invariant predictions, also the conditional prediction rates of pairwise trained base classifiers are invariant against the order of classes. They can again be precalculated and looked up if necessary (Table 2). Previously, it has been shown that the class-wise sensitivities of an ordinal classifier cascade are upper bounded by a set of conditional prediction rates of the corresponding base classifiers [37].

Theorem 1 *Let h denote an ordinal classifier cascade*

$$h : \mathbb{R}^n \longrightarrow \mathcal{Y} = \{y_{(1)}, \dots, y_{(|\mathcal{Y}|)}\} \quad (20)$$

with base classifiers $\mathcal{E} = \{c_{(1)}, \dots, c_{(|\mathcal{Y}|-1)}\}$. Let furthermore $\mathcal{X}_{(i)}$ be a non-empty set of samples of class $y_{(i)}$. Then the sensitivity of h for $y_{(i)}$ is limited by

$$p_h(y_{(i)} | \mathcal{X}_{y_{(i)}}) \leq p_{c_{(i)}}(y_{(i)} | \mathcal{X}_{y_{(i)}}) \quad (21)$$

$$p_h(y_{(i)} | \mathcal{X}_{y_{(i)}}) \leq \min_{k < i} p_{c_{(k)}}(y_{(k+1)} | \mathcal{X}_{y_{(i)}}). \quad (22)$$

Note that neither the construction nor the evaluation of an ordinal classifier cascade is required for providing these upper limits. A limit to the minimal class-wise sensitivity of an ordinal classifier cascade can be given by a set of lookups. In this way a large proportion of those cascades that will not pass a predefined threshold $t \leq \text{Smin}_c(\mathcal{Y})$ can be rejected before starting the more expensive calculations on the prediction table (Table 1). We utilise the CASCADES algorithms [37], which realises this strategy as preprocessing step in order to reduce the computational burden. Only those cascades, which pass this filter will be evaluated for their real minimal class-wise sensitivity.

Table 2 Scheme of prediction rates

Classifier $c_{i,j}(\mathbf{x})$	Classes $y \in \mathcal{Y}$				
	y_1	$\dots$	y_k	$\dots$	$y_{ \mathcal{Y} }$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
y_i, y_j	$\dots$	$\dots$	$p_{c_{i,j}}(y_i \mathcal{X}_{y_k})$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$

The table provides the conditional prediction rates of a set of pairwise trained base classifiers $c_{i,j}$ for a set of validation $\mathcal{V} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{|\mathcal{V}|}$. As pairwise trained base classifiers do not depend on the chosen class order, all conditional prediction rates have to be calculated only once. The table is of size $\binom{|\mathcal{Y}|}{2} \times |\mathcal{Y}|$ for classes $\mathcal{Y}$

Fig. 2 Schematic drawing of ordinal cascades. **a** A scheme of a full ordinal cascade that comprises all available classes $y_{(1)}, \dots, y_{(6)}$. **b** An ordinal subcascade. Here classes $y_{(1)}, \dots, y_{(6)}$ are embedded in a larger collection of classes (black circles), which can not be linked via one single ordinal relationship

3 Screening for Ordinal (Sub-)structures

Here, we focus on the analysis of multi-class datasets that are not originally designed for ordinal classification (Fig. 2). That is, we do not assume a unique ordinal relationship for all classes in $\mathcal{Y}$. The classes in $\mathcal{Y}$ can instead be seen as a loose collection of known categories within a common context of interest. Nevertheless, subsets of classes $\mathcal{Y}' \subseteq \mathcal{Y}$ can fulfil our quality criteria for ordinal classification. The corresponding decision regions can be interpreted as an ordinal structure, and we might hypothesise an ordinal relation between the classes in $\mathcal{Y}'$. Here, memoization tables (Tables 1, 2) proposed for the full set of classes $\mathcal{Y}$ can directly be repurposed for the analysis of any subset of classes $\mathcal{Y}'$.

3.1 Size of the Search Space

The number of candidate cascades is dependent on the chosen screening task. In all cases, it can be decomposed in the number of class orders $|\mathcal{Y}|!$ and the number of subsets $\binom{|\mathcal{Y}|}{k}$ of a predefined size k ¹:

Full cascades: $o(|\mathcal{Y}|) = |\mathcal{Y}|!$

Subcascades of size k : $s_k(|\mathcal{Y}|) = \binom{|\mathcal{Y}|}{k} k!$

¹ For simplification we allow cascades of length $k = 1$ in this section.

Table 3 Number of candidate cascades

	$ \mathcal{Y} = 5$	$ \mathcal{Y} = 10$	$ \mathcal{Y} = 15$	$ \mathcal{Y} = 20$	$ \mathcal{Y} = 25$
$s_{all}(\mathcal{Y})$	325	9.8×10^6	3.5×10^{12}	6.6×10^{18}	4.2×10^{25}
$o(\mathcal{Y})$	120	3.6×10^6	1.3×10^{12}	2.4×10^{18}	1.5×10^{25}
$s_{all}(\mathcal{Y})/o(\mathcal{Y})$	2.708	2.718	2.718	2.718	2.718

The table provides the number of possible full cascades $o(|\mathcal{Y}|)$ and subcascades $s_{all}(|\mathcal{Y}|)$ for $|\mathcal{Y}| = \{5, 10, 15, 20, 25\}$ classes

$$\text{All subcascades: } s_{all}(|\mathcal{Y}|) = \sum_{k=1}^{|\mathcal{Y}|} \binom{|\mathcal{Y}|}{k} k!$$

It might be noteworthy that $s_k(|\mathcal{Y}|) < s_{|\mathcal{Y}|-k}(|\mathcal{Y}|)$ for $k < |\mathcal{Y}|/2$.

The ratio of the number of all subcascades and the number of full cascades is approximately given by the Euler number as

$$s_{all}(|\mathcal{Y}|) = \sum_{k=1}^{|\mathcal{Y}|} \binom{|\mathcal{Y}|}{k} k! \quad (23)$$

$$= |\mathcal{Y}|! \sum_{k=1}^{|\mathcal{Y}|} \frac{1}{k!} \quad (24)$$

and the ratio itself

$$\frac{s_{all}(|\mathcal{Y}|)}{o(|\mathcal{Y}|)} = \sum_{k=1}^{|\mathcal{Y}|} \frac{1}{k!} \xrightarrow{|\mathcal{Y}| \rightarrow \infty} e \sim 2.718. \quad (25)$$

Examples for numbers of candidate cascades can be found in Table 3. Note that for other (traditional) types of ordinal classifiers the evaluation of each candidate cascade requires the de novo training of the corresponding model. As these training phases by far exceed the time complexity of evaluating the memoization tables (Tables 1, 2), the proposed screenings are infeasible for traditional algorithms and implementations.

3.1.1 Properties of Subcascades

The combinatorics show that the overall number of subcascades exceeds the overall number of full cascades by only a constant factor. Nevertheless, we expect a much higher number of candidate subcascades to pass a threshold on the minimal class-wise sensitivity than a number of candidate full cascades.

Theorem 2 Let $y_{(1)} < \dots < y_{(i)} < \dots < y_{(|\mathcal{Y}|)}$ denote an arbitrary but fixed class order of the classes in $\mathcal{Y}$ and $h : \mathbb{R}^n \rightarrow \mathcal{Y}$ an ordinal classifier cascade designed for this class order and based on an ensemble of pairwise trained base classifiers $\mathcal{E} = \{c_{(i)}\}_{i=1}^{|\mathcal{Y}|-1}$. Let furthermore $\mathcal{Y}' = \mathcal{Y} \setminus \{y_{(i)}\}$ and $h' : \mathbb{R}^n \rightarrow \mathcal{Y}'$ denote an ordinal classifier cascade designed for class order $y_{(1)} < \dots < y_{(i-1)} < y_{(i+1)} < \dots < y_{(|\mathcal{Y}|)}$ based on $\mathcal{E}' = \mathcal{E} \setminus \{c_{(1)}\}$, if $i = 1$, $\mathcal{E}' = \mathcal{E} \setminus \{c_{(|\mathcal{Y}|-1)}\}$, if $i = |\mathcal{Y}|$ and

$$\mathcal{E}' = (\mathcal{E} \setminus \{c_{(i-1)}, c_{(i)}\}) \cup \{c'_{(i-1)}\} \quad (26)$$

otherwise, where

$$c'_{(i-1)}(\mathbf{x}) := \begin{cases} y_{(i-1)} & \text{if } c_{(i-1)}(\mathbf{x}) = y_{(i-1)} \\ y_{(i+1)} & \text{if } c_{(i-1)}(\mathbf{x}) = y_{(i)} \end{cases}. \quad (27)$$

In this case

$$Smin_h(\mathcal{Y}) \leq Smin_{h'}(\mathcal{Y}'). \quad (28)$$

Proof From the monotonicity of the minimal class-wise sensitivity (Equation 12) we get

$$Smin_h(\mathcal{Y}) \leq Smin_{h'}(\mathcal{Y}'). \quad (29)$$

Cascade h' is constructed from the same base classifiers as h . Only one classifier is removed, which closes the option of predicting class label $y_{(i)}$. In case $i = 1$, all samples that would have been classified as $y_{(1)}$ by h are directly sent to $c_{(2)}$ by h' . This is of course also true for all remaining samples with $c_{(1)}(\mathbf{x}) = y_{(2)}$ in h . Therefore for each class label $y \in \mathcal{Y}'$, the set of samples that receive this class label y by h' comprises at least all samples that receive class label y by h . The corresponding class-wise sensitivities and their minimum can only be increased.

In case $1 < i < |\mathcal{Y}|$, additionally classifier $c_{(i-1)}$ and $c_{(i)}$ are replaced by $c'_{(i-1)}$, which redirects those samples that would be sent from $c_{(i-1)}$ to $c_{(i)}$. The common preamble of base classifiers of h and h' leads to identical class-wise sensitivities for classes $y_{(1)}, \dots, y_{(i-1)}$. Those classes that received class label $y_{(i)}$ by h are again reclassified into one of the subsequent classes leading to equivalent or increased class-wise sensitivities.

In case $i = |\mathcal{Y}|$, the cascades h and h' share the longest possible preamble of base classifiers. Only the samples receiving class label $y_{(|\mathcal{Y}|)}$ by h will be reclassified as $y_{(|\mathcal{Y}|-1)}$ by h' leading to at least the class-wise sensitivity for class $y_{(|\mathcal{Y}|)}$.

We therefore get for all cases

$$Smin_h(\mathcal{Y}') \leq Smin_{h'}(\mathcal{Y}'). \quad (30)$$

□

Theorem 2 states that a subcascade h' constructed from a full cascade h by removing a single class $y_{(i)}$ and by reconnecting the corresponding base classifiers is guaranteed to achieve at least the same minimal class-wise sensitivity on $\mathcal{Y}'$ than h on $\mathcal{Y}$. It can recursively be applied to h' and its own subcascades leading to a transitive relationship between a full cascade and all its subcascades. The more classes a cascade comprises the more difficult it will be to reach a predefined minimal class-wise sensitivity. We will therefore focus on longer cascades.

4 Experiments

We evaluated our screening procedure empirically in 10 x 10 cross-validation experiments [24]. All ordinal classifier cascades were based on linear SVMs (cost=1) [49]. The TunePareto Software was used for the evaluation of base classifiers [41]. All time measure experiments were performed on an AMD Opteron(tm) Processor 6276 with 2.6 GHz (32 cores with HT) and 512 GB RAM.

As mentioned above, alternative systems are unlikely to cope with the computational burden of exhaustive screenings (Sect. 3). Reference classifiers were therefore directly applied to those orders identified by our screening procedure.

We compared our results to those of other multi-class architectures based on independently trained linear SVMs (cost=1). As ordinal reference architectures we used a splitted ordinal classifier cascade (Scc, pairwise training) denoted as “voter 3” by Jiang et al. [26], the ensembles of nested dichotomies (END, $|\mathcal{E}| = 20$) [19] and directed acyclic graphs

Table 4 Overview of the analysed datasets

n	$ \mathcal{Y} $	$ \mathcal{S} $	$ \mathcal{S}_y $					
d_1 : Handwritten digits								
64	10	5620	$y_1 = 554$	$y_2 = 571$	$y_3 = 557$	$y_4 = 572$	$y_5 = 568$	
			$y_6 = 558$	$y_7 = 558$	$y_8 = 566$	$y_7 = 558$	$y_8 = 566$	
			$y_9 = 554$	$y_{10} = 562$				
d_2 : Isolet								
617	26	7797	$y_1, \dots, y_{26} = 300$					
d_3 : Cancer cell lines								
54613	9	174	$y_1 = 18$	$y_2 = 15$	$y_3 = 21$	$y_4 = 26$	$y_5 = 18$	
			$y_6 = 21$	$y_7 = 23$	$y_8 = 26$	$y_9 = 6$		
d_4 : Leukemia								
54613	18	2096	$y_1 = 40$	$y_2 = 36$	$y_3 = 58$	$y_4 = 48$	$y_5 = 28$	
			$y_6 = 351$	$y_7 = 38$	$y_8 = 37$	$y_9 = 40$	$y_{10} = 237$	
			$y_{11} = 122$	$y_{12} = 448$	$y_{13} = 76$	$y_{14} = 13$	$y_{15} = 206$	
			$y_{16} = 74$	$y_{17} = 70$	$y_{18} = 174$			
d_5 : TCGA								
52762	8	2370	$y_1 = 45$	$y_2 = 424$	$y_3 = 182$	$y_4 = 521$	$y_5 = 177$	
			$y_6 = 321$	$y_7 = 611$	$y_8 = 89$			
d_6 : <i>C.elegans</i>								
22548	10	123	$y_1 = 12$	$y_2 = 12$	$y_3 = 13$	$y_4 = 11$	$y_5 = 13$	
			$y_6 = 13$	$y_7 = 10$	$y_8 = 14$	$y_9 = 11$	$y_{10} = 14$	
d_7 : Wound healing								
54613	8	857	$y_1 = 338$	$y_2 = 68$	$y_3 = 107$	$y_4 = 116$	$y_5 = 104$	
			$y_6 = 54$	$y_7 = 33$	$y_8 = 37$			
d_8 : HPA-axis								
54	7	336	$y_1, \dots, y_7 = 48$					

The domain, the dimensionality n , the number of classes $|\mathcal{Y}|$, samples $|S|$ and the number of samples per class $|S_y|$ are given

(DAG) [44]. As non-ordinal references the one-against-one (OaO) and the one-against-all (OaA) architectures were applied [40].

Additionally, our results were compared to a 1D self-organising map (SOM) [30]. This mapping was performed using the standard settings of the R-package *kohonen* [53,54] (learning rate: [0.05,0.01], but no layer normalisation). The nodes and consequently the clusters were labelled based on the cascade under investigation.

4.1 Datasets

The screening procedure is evaluated on eight multi-class datasets from different domains. The main characteristics of the analysed datasets can be found in Table 4. We used data, which comprises different representations of alphanumeric characters (written and spoken) from the machine learning database UCI [13], and gene expression and methylation profiles which were measured either by microarrays or by deep sequencing.

d_1 : *Handwritten digits dataset*. The handwritten digits dataset was made publicly available by Alpaydin and Kaynak [3] as part of the UCI repository [13]. This dataset is a collection of bitmaps of digits written by 43 different people. The classes $y_1, \dots, y_{10}$ correspond to the labels 0, ..., 9.

d_2 : *Isolet dataset*. The Isolet (Isolated Letter Speech Recognition) dataset by Fandy and Cole [17] and downloaded from the UCI repository [13] consists of spoken letters of the alphabet. The classes $y_1, \dots, y_{26}$ correspond to the labels $a, \dots, z$.

d_3 : *Cancer cell lines*. The NCI-60 dataset (GSE32474) was collected by Pfister et al. [43]. It consists of gene expression profiles from cell lines that derived from different cancer tissue types. The different cancer types are: y_1 : Leukemia, y_2 : Breast cancer, y_3 : Ovarian cancer, y_4 : Melanoma, y_5 : Central nervous system (CNS), y_6 : Colon cancer, y_7 : Renal cancer, y_8 : Non-small cell lung cancer, y_9 : Prostate cancer.

d_4 : *Leukemia*. As a prephase to the MILE Study (Microarray Innovations In LEukemia) program expression data from blood and bone marrow samples from acute and chronic leukemia patients was collected (GSE13159). Kohlmann et al. [29] could hereby show that standardised experimental protocols can lead to comparable results across different laboratories. The samples have been categorised in different leukemia subgroups and assigned to class labels as follows: y_1 : ALL (acute lymphocytic leukaemia) with hyperdiploid karyotype, y_2 : ALL with t(1;19), y_3 : ALL with t(12;21), y_4 : AML (acute myeloid leukaemia) complex aberrant karyotype, y_5 : AML with inv(16)/t(16;16), y_6 : AML with normal karyotype and other abnormalities, y_7 : AML with t(11q23)/MLL, y_8 : AML with t(15;17), y_9 : AML with t(8;21), y_{10} : c-ALL/Pre-B-ALL without t(9;22), y_{11} : c-ALL/Pre-B-ALL with t(9;22), y_{12} : CLL (chronic lymphocytic leukaemia), y_{13} : CML (chronic myeloid leukaemia), y_{14} : mature B-ALL with t(8;14), y_{15} : MDS (myelodysplastic syndromes), y_{16} : Non-leukemia and healthy bone marrow, y_{17} : Pro-B-ALL with t(11q23)/MLL, y_{18} : T-ALL.

d_5 : *TCGA*. This dataset is a collection of datasets that was generated by the TCGA Research Network (<http://cancergenome.nih.gov/>). The feature space consists of the intersection of all features of the single datasets. The different cancer types are assigned to class labels as follows: y_1 : Cholangiocarcinoma (CHOL), y_2 : Liver hepatocellular carcinoma (LIHC), y_3 : Pancreatic adenocarcinoma (PAAD), y_4 : Colon adenocarcinoma (COAD), y_5 : Rectum adenocarcinoma (READ), y_6 : Kidney renal papillary cell carcinoma (KIRP), y_7 : Kidney renal clear cell carcinoma (KIRC), y_8 : Kidney chromophobe carcinoma (KICH).

d_6 : *C. elegans*. Baugh et al. [6] analysed the influence of the homeodomain protein PAL-1 of the C-lineage-specific gene regulatory network in the model organism *C. elegans*. They gathered gene expression data of samples of wild-type embryos and mutant embryos at 10 points in time after the 4-cell-stage of the embryo (GSE2180). We normalised (mas5) and labelled these samples based on the points in time: y_1 : 0 minutes, y_2 : 23 minutes, y_3 : 41 minutes, y_4 : 53 minutes, y_5 : 66 minutes, y_6 : 83 minutes, y_7 : 101 minutes, y_8 : 122 minutes, y_9 : 143 minutes, y_{10} : 186 minutes.

d_7 : *Wound healing*. Xiao et al. [56] analysed the total cellular RNA of blood samples taken from severe blunt trauma patients (GSE36809). They performed a time scale analysis and took samples at different points in time after the injury. We normalised (mas5) and summarised the different points in time, measured in hours and labelled them as follows: y_1 : (0,50), y_2 : [50,100), y_3 : [100,150), y_4 : [150,200), y_5 : [200,400), y_6 : [400,600), y_7 : [600,800), y_8 : control.

d_8 : *HPA-axis*. This dataset by Agba et al. [1] comprises measurements of DNA methylation patterns of different promoters of the glucocorticoid receptor gene (NR3C1) and the imprinting control region of IGF2/H19 of 7 tissues of rats. These profiles especially enclose measurements of the hypothalamic-pituitary-adrenal-axis (HPA-axis), a neuroendocrine sys-

tem which is known to be involved in the regulation of stress reactions. In our experiments, the classes denote the different tissue types y_1 : Cortex, y_2 : Hippocampus, y_3 : Hypothalamus, y_4 : Pituitary, y_5 : Adrenal, y_6 : Skin, y_7 : Liver.

5 Experimental Results and Interpretation

In this section we will present the results on the utilised benchmark data sets and give some interpretation on these results. Table 5 provides an overview on the extracted longest cascades. Between one and five cascades were detected per dataset passing the dataset specific highest threshold. They achieved minimal class-wise sensitivity thresholds ranging from 0.75 to 0.9. The theoretical maximum number of cascades is for all datasets more than 800 times larger than the number of returned cascades. A general overview about the numbers of cascades that pass the first threshold in comparison to the number that pass the second threshold (dataset specific) is given in the Supplementary Information.

Longest cascades. Graph representations of the longest cascades can be found in Figs. 3 and 4. It can be seen that all found cascades overlap in at least one class except for the TCGA data d_5 . The extrema in the amount of overlap are d_5 which shows no overlap and cascades of d_2 and d_6 which overlap in all except one class. Further structures that can be observed are that cascades differ in one class at the beginning (d_2) in the middle (d_6) or at the end (d_1), but especially if more than two cascades pass the minimal sensitivity criterion the cascades also align to more complex graphs (d_3 , d_8).

Classes that are not included into the subcascade. Extended confusion tables for the longest cascades are shown in the Supplementary Information. Figure 5 shows exemplarily the confusion table for the TCGA dataset d_5 . As focusing on the detection of subcascades, classes exist that are not part of the longest cascades, termed **Others** in Figs. 3 and 4. These classes can be split in classes that are not part of any of the longest subcascades (uncovered classes) and classes that are not part of the specific subcascade but part of another of the longest subcascades returned (alternative class). Presenting a sample of one of these classes to a specific subcascade, these samples are necessarily classified as one of the cascade classes as there is no rejection possibility. A general overview of how distinct the allocation of these additional classes to cascade classes is, provides the mean of the purity values (Eq: 13) (Table 5).

De novo calculation vs memoization scheme. In order to demonstrate the impact of the proposed memoization techniques we performed additional experiments with a traditional implementation of an ordinal classifier cascade. That is each training phase is realised as a de novo adaptation of the classification model. Both versions were based on the same implementation of base classifiers (SVM) [41] and did not use any parallelisation. The runtime of both approaches was compared on dataset d_4 ($|\mathcal{Y}| = 18$), which corresponds to $s_{all}(|\mathcal{Y}|) > 10^{18}$ subcascades. Again a 10×10 CV was conducted for each subcascade. The memoization techniques allowed to accomplish this screening in about 1868 minutes, while the de novo calculation of a single 10×10 CV on all $|\mathcal{Y}| = 18$ classes (fixed class order) required 798 minutes. The de novo calculation is therefore not able to perform this screening in a feasible amount of time.

Performance comparison. As the complexity of an ordinal classifier cascade is rather low in comparison to the complexity of the chosen reference classifiers we have chosen to run these algorithms only for the longest subcascades identified by the initial screening. The minimal class-wise sensitivities achieved by the reference classifiers on the longest subcascades are

d_1 : Handwritten digits ($t=0.75$, $l=5$)

Cascades:

d_2 : Isolet ($t=0.9$, $l=10$)

Cascades:

d_3 : Cancer cell lines ($t=0.85$, $l=5$)

Cascades:

Others:

d_4 : Leukemia ($t=0.75$, $l=7$)

Cascades:

Others:

Fig. 3 Longest cascades detected for datasets $d_1 - d_4$. Descriptions of the datasets are given in the text. For each dataset the threshold (t) and the length of the longest cascades (l) is stated. The individual cascades are color encoded differently. For datasets d_1 and d_2 the left out classes (others) are not shown

d_5 : TCGA ($t=0.9$, $l=4$)

Cascade:

Others:

d_6 : C. elegans ($t=0.75$, $l=9$)

Cascades:

d_7 : Wound healing ($t=0.8$, $l=4$)

Cascades:

Others:

d_8 : HPA-axis ($t=0.9$, $l=5$)

Cascades:

Others:

Fig. 4 Longest cascades detected for datasets $d_5 - d_8$. Descriptions of the datasets are given in the text. For each dataset the threshold (t) and the length of the longest cascades (l) is stated. The individual cascades are color encoded differently. Additionally, the left out classes, which are classes that are not included in neither of the cascades are depicted as others

Table 5 The characteristics of the longest cascades

Cascade	S _{min}	pur _{alt}	pur _{unc}
d_1: Handwritten digits ($t=0.75$, $l=5$): 2(30240)		$alt=2$	$unc=3$
$y_7 < y_2 < y_6 < y_{10} < y_4$	79.8	53.6	72.9
$y_7 < y_3 < y_6 < y_{10} < y_8$	76.2	56.4	74.6
d_2: Isolet ($t=0.9$, $l=10$): 2 ($> 1 \cdot 10^{13}$)		$alt=1$	$unc=15$
$y_{19} < y_6 < y_{13} < y_{14} < y_1 < y_2 < y_4 < y_7 < y_{26} < y_3$	90.8	41.8	66.0
$y_{25} < y_6 < y_{13} < y_{14} < y_1 < y_2 < y_4 < y_7 < y_{26} < y_3$	90.4	100	68.6
d_3: Cancer cell lines ($t=0.85$, $l=5$): 4(15120)		$alt=3$	$unc=1$
$y_1 < y_6 < y_3 < y_8 < y_4$	88.5	62.4	55.3
$y_1 < y_6 < y_3 < y_8 < y_5$	88.5	62.3	39.3
$y_1 < y_9 < y_3 < y_7 < y_5$	86.7	51.8	38.7
$y_1 < y_6 < y_3 < y_7 < y_5$	86.1	67.8	42.0
d_4: Leukemia ($t=0.75$, $l=7$): 2($> 1 \cdot 10^{18}$)		$alt=6$	$unc=5$
$y_{12} < y_2 < y_{10} < y_{18} < y_5 < y_9 < y_8$	75.8	69.3	61.1
$y_1 < y_{17} < y_5 < y_7 < y_{13} < y_{15} < y_{16}$	75	77.0	69.8
d_5: TCGA ($t=0.9$, $l=4$): 1(1680)		$alt=0$	$unc=4$
$y_2 < y_7 < y_3 < y_5$	91.2	NA	94.8
d_6: C. elegans ($t=0.75$, $l=9$): 2($> 3 \cdot 10^{16}$)		$alt=1$	$unc=0$
$y_1 < y_2 < y_3 < y_4 < y_6 < y_7 < y_8 < y_9 < y_{10}$	79.3	52.3	NA
$y_1 < y_2 < y_3 < y_4 < y_5 < y_7 < y_8 < y_9 < y_{10}$	79.3	52.3	NA
d_7: Wound healing ($t=0.8$, $l=4$): 2(1680)		$alt=1$	$unc=3$
$y_8 < y_6 < y_3 < y_1$	82.3	89.1	67.0
$y_8 < y_7 < y_3 < y_1$	80.3	82.2	65.3
d_8: HPA-axis ($t=0.9$, $l=5$): 5(2520)		$alt=1$	$unc=1$
$y_7 < y_6 < y_2 < y_4 < y_5$	93.1	85.6	68.3
$y_7 < y_6 < y_2 < y_5 < y_4$	93.1	95.2	68.3
$y_5 < y_4 < y_2 < y_6 < y_7$	92.7	87.7	89.8
$y_5 < y_4 < y_3 < y_6 < y_7$	92.3	90.8	43.3
$y_7 < y_6 < y_3 < y_4 < y_5$	91.4	90.4	36.5

For each dataset the threshold (t), the length of the longest cascade (l) and the number of the returned cascades, as well as the maximal number of theoretical possible cascades (in brackets) is given. Classes that are not part of the cascade are splitted in alternative classes (classes that are part of another returned cascade) and uncovered cascades (classes that are not within any of the longest cascades) and the size of each of these subsets is given as alt and unc . For each cascade the minimal class-wise sensitivity S_{min} is given, as well as the averaged misclassification purity values for the alternative subset pur_{alt} and the uncovered subset pur_{unc} .

shown in Table 6. An overview on all sensitivities is given in the Supplementary Information. The comparison to the non-ordinal architectures (OaO, OaA) demonstrate the impact of (correct) ordinal information by a higher performance of the ordinal classifier cascade (Occ). In five out of eight datasets ($d_4 - d_8$) cascades exist for which the OaO achieved lower minimal sensitivities and did not even pass the minimal threshold (t). For the OaA this was observed only twice (d_6, d_7). The ordinal architectures, the ensembles of nested dichotomies (END) and directed acyclic graphs (DAG), perform well on these datasets and constrained to the selected orders. Both approaches did not pass the minimal threshold t on d_7 . The END did

Table 6 The table shows the minimal class-wise sensitivity performance as % of various methods on the selected subcascades

	Occ	ScC	END	DAG	SOM	OaO	OaA
d_1: Handwritten digits ($t=0.75$, $l=5$)							
$y_7 < y_2 < y_6 < y_{10} < y_4$	79.8	30.6	97.8	97.5	0.0	96.6	96.7
$y_7 < y_3 < y_6 < y_{10} < y_8$	76.2	22.1	98.6	98.7	2.6	96.6	98.0
d_2: Isolet ($t=0.9$, $l=10$)							
$y_{19} < y_6 < y_{13} < y_{14} < y_1 < y_2 < y_4 < y_7 < y_{26} < y_3$	90.8	24.2	93.0	92.7	11.3	92.6	91.1
$y_{25} < y_6 < y_{13} < y_{14} < y_1 < y_2 < y_4 < y_7 < y_{26} < y_3$	90.4	11.0	92.5	92.7	0.1	92.6	91.2
d_3: Cancer cell lines ($t=0.85$, $l=5$)							
$y_1 < y_6 < y_3 < y_8 < y_4$	88.5	37.7	96.2	96.5	5.7	96.2	98.5
$y_1 < y_6 < y_3 < y_8 < y_5$	88.5	63.9	100.0	100.0	11.7	100.0	100.0
$y_1 < y_9 < y_3 < y_7 < y_5$	86.7	6.7	100.0	100.0	11.4	100.0	100.0
$y_1 < y_6 < y_3 < y_7 < y_5$	86.1	66.1	100.0	100.0	12.4	100.0	100.0
d_4: Leukemia ($t=0.75$, $l=7$)							
$y_{12} < y_2 < y_{10} < y_{18} < y_5 < y_9 < y_8$	75.8	0.0	85.0	84.4	0.1	81.6	87.5
$y_1 < y_{17} < y_5 < y_7 < y_{13} < y_{15} < y_{16}$	75.0	17.1	74.9	77.7	1.3	66.5	77.8
d_5: TCGA ($t=0.9$, $l=4$)							
$y_2 < y_7 < y_3 < y_5$	91.2	93.8	95.2	95.8	10.1	50.5	96.5
d_6: <i>C. elegans</i> ($t=0.75$, $l=9$)							
$y_1 < y_2 < y_3 < y_4 < y_6 < y_7 < y_8 < y_9 < y_{10}$	79.3	0.0	81.4	79.3	6	79.3	79.0
$y_1 < y_2 < y_3 < y_4 < y_5 < y_7 < y_8 < y_9 < y_{10}$	79.3	50.8	81.4	79.3	5	66.2	68.2
d_7: Wound healing ($t=0.8$, $l=4$)							
$y_8 < y_6 < y_3 < y_1$	82.3	76.1	80.6	82.8	9.7	33.0	83.5
$y_8 < y_7 < y_3 < y_1$	80.3	56.4	65.1	72.4	10.3	22.7	69.4
d_8: HPA-axis ($t=0.9$, $l=5$)							
$y_7 < y_6 < y_2 < y_4 < y_5$	93.1	92.7	95.8	93.1	91.7	76.2	97.5
$y_7 < y_6 < y_2 < y_5 < y_4$	93.1	92.7	95.8	93.1	1.5	76.2	97.5
$y_5 < y_4 < y_2 < y_6 < y_7$	92.7	92.7	95.8	93.1	0.0	76.2	97.5
$y_5 < y_4 < y_3 < y_6 < y_7$	92.3	92.3	95.2	92.3	0.0	62.3	95.4
$y_7 < y_6 < y_3 < y_4 < y_5$	91.5	92.3	95.4	92.3	58.5	62.3	95.4

Various ordinal classifier methods, as the investigated ordinal classifier cascade (Occ), Nested Dichotomies (END), directed acyclic graph SVMs (DAG) or splitted (ordinal) classifier cascade (ScC) are applied to the order returned by the cascades algorithm within the screening. Additionally, a linear self-organising map (SOM) was used to cluster the data and the cluster performance was evaluated based on the order. Furthermore, two non-ordinal multi-class approaches, one-against-one (OaO) and one-against-all (OaA) are used

also slightly pass the threshold on d_4 . The splitter ordinal classifier cascade (ScC) achieved high minimal class-wise sensitivities only on d_8 dataset and d_5 . It would have deselected individual classes in our experiments. With the exception of the first cascade of d_8 SOM was also not able to separate all classes of the cascades.

Handwritten digits dataset. The two longest cascades of length five which are found in d_1 the digit dataset overlap at three classes. A meaningful interpretation of the direction is not obvious, but the returned cascades show possible structural properties. There are three digits (0, 4, 8) that are uncovered classes and in both cascades the 8 is mainly allocated to the second cascade position, corresponding once to the digit 2 and once to 1. If 1 is placed at the

second position, 0 is mainly classified as 6, whereas in the other cascade (2 at position two), 0 is classified as 5 (the third position of the cascade), which means that in the first case the first decision region is placed in a way that the samples of class 6 and 0 are on the same side, whereas in the second case they lay on opposite sides and hence the directions of the first relation in the feature space for both cascades differ. The two cascades split at the second and at the last position of the cascade and the purity values for the alternative classes show that on average half of the samples of the alternative classes are assigned to one intra-cascade class.

Isolet dataset. Dataset d_2 corresponds to a spoken representation of letters. Two cascades of length 10 pass a minimal class-wise sensitivity threshold of 0.9. The uncovered classes are classified to around 70% on average to the same class and the heatmaps (Supplementary Information) show that they are not well-distributed as many entries are zero. The two cascades differ only in their first position (s vs. y). If not part of the cascade the class s is completely assigned to the second cascade class f, whereas in the case in which y is not part of the cascade its assignment fades out till the fifth intra-cascade class. The different classes might be grouped based on the phonetic alphabet. In this case s, f, m and n share the sound [ɛ], but y does not, which might be one possible interpretation of what is observed. The last five classes (b, d, g, z, c) of both cascades share the sound [i:]. The further letters (e, p, t, v) that are pronounced using this sound are also mainly classified as one of these classes. Why they are not included in the main cascade is not clear from the semantic level. The cascade, however, reveals that the samples and hence the classes are located in a way that including one of those classes seem to change the direction of the decision region sequence in a way that would lead to shorter sequences.

Cancer cell lines. The cancer cell line dataset d_3 shows four cascades of length five. There is only one class, y_2 : breast cancer, that is not part of any cascade and reveals a low purity. All four cascades share their first class, y_1 : leukemia and their center class, y_3 : ovarian cancer. If not part of the returned cascade no sample of y_6 : colon cancer or y_8 : lung cancer is classified as y_9 : prostate cancer, however the other way round y_9 : prostate cancer splits half-half between those classes, which reveals that the decision regions differ depending on the assumed order and outlines the importance of the fact that the assumed order has to be appropriately reflected in the feature representation. The correlation within this dataset can be explained by messenger molecules. Both prostate as well as ovarian are affected by sex hormones [28]. The fact that leukemia is twice as prevalent in men as in women [2] does not rule out the influence of sex hormones and may explain why leukemia is closer to prostate cancer. A similar relationship can also be found between ovarian and kidney cancer. Here, it is assumed that estrogen has a protective effect on the kidney, while testosterone damages the kidney [7,48]. Likewise, an effect of sex hormones on the CNS is clearly confirmed [57]. Here, a decrease in hormone concentration during aging can be attributed to loss of neuronal functions. Various studies support a neuroprotective role of estrogens on this neuronal decline [58]. The other cascade connects neurotransmitter-controlled tissues. The central nerve system is controlled by a diversity of neurotransmitters [42]. Non-adrenergic, non-cholinergic nerves are known to control the gut as well as the urogenital tract [5]. Since the lung develops embryonically seen from the foregut, it is not surprising that non-adrenergic, non-cholinergic nerves are also present [5]. In addition, secreted neurotransmitter such as glutamate are known to be involved in T-cell leukemia [20].

Leukemia. Both cascades found in d_6 share exactly one class y_5 : AML with inv(16)/t(16,16). In both cascades classes that correspond to lymphoid leukemia (LL) are placed before this class and afterwards classes that correspond to myeloid leukemia (ML) or syndromes that might develop into a myeloid leukemia, with the exception of the healthy bone

Fig. 5 Longest cascade of the TCGA dataset d_5 (extended confusion matrix). The sensitivity values and the conditional prediction rates for the classes belonging to the selected cascades as well as the external rates of classes that do not belong to the cascade are shown. The dataset specific threshold (t) and the minimal class-wise sensitivity is given in the heading. The columns give the order of classes predicted by the cascade (from left to right). The rows give all classes of a dataset. They are split in two blocks (separated by a red line): classes covered by the own cascade (top), other classes (bottom)

marrow. It can be observed that the AML classes that are not included into one of the cascades are not classified as an own entity according to the WHO classification system [4]. The samples of those classes might be considered as not distinct enough to be included. Furthermore, it was shown that certain subtypes of LL show a higher incidence for the expression of myeloid antigens and these are pro B-ALL and early T-ALL [55] which are the LL classes at the found ML-LL transition. This might lead to the hypothesis that the found subcascades represent a sequence of similar but still distinct subtypes. The cascade showing a higher average purity for both, the alternative and the uncovered classes (others), reflects the grouping of ML and LL also in these classes. The observed characteristic that classes of similar concepts form subcascades within the context of a larger research topic, reveals that the proposed screening procedure is able to find orders within such a group but is also at the same time able to relate these groups using the same consecutive and hence overarching pattern.

TCGA data. For the TCGA dataset d_5 consisting of eight classes, a cascade of length four achieved a minimal class-wise sensitivity higher than 90%. For d_5 all uncovered classes are assigned with a purity of at least 80%. An order for most tissue derived tumors of endodermal or mesodermal origin can be found here. Only the data for kidney renal clear cell carcinoma (KIRC) cannot be assigned to the mesodermal germ layer. Similar findings were made by Berman [8] in his classification analysis. Here, he found that epithelial tumors with mesodermal origin in particular exhibit class independent behavior [8]. In line with this, there are various differentiation protocols to derive functional pancreatic or liver cells from human pluripotent stem cells (hPSCs) while few protocols are able to induce effective kidney differentiation [32]. Thus, differentiation of renal cells and associated tumors is somehow more complex in comparison to other tissues. One reason for this might be the complex architecture of the kidney with all its functional units [32] that require a variety of differentiation factors not only specific for kidney.

C.elegans and wound healing data. In contrast to the other datasets, the class labels of d_6 and d_7 are given by consecutive time points and therefore might be seen as hypothesised ordinal class labels. These timelines were not completely reconstructed in our experiments. Nevertheless the identified subcascades followed the assumed order.

For *C.elegans* d_6 two cascades passed a minimal class-wise sensitivity threshold of 75% that included all classes despite one. Both orderings correspond to a progression in time and either the class corresponding to samples taken at minute 66 (y_5) or the ones taken at minute 83 (y_6) are skipped. In both cases the uncovered classes were assigned to their assumed ordinal neighbours with a purity over 90%. The high similarity of classes y_5 and y_6 are also indicated by a decreased classification performance of the corresponding base classifier (74% minimal class-wise sensitivity, data not shown) suggesting that in this period the expression does not change a lot. The detected cascades therefore rather constructs an ordinal sequence of discriminable events than a reconstruction of the timeline.

Similar observations can be gained for the wound healing dataset d_7 . Here two cascades of length four achieved a minimal class-wise sensitivity higher than 80%. They differ in their second class. In the first cascade y_6 is chosen, which corresponds to the samples taken between 400 and 600 minutes after injury. In the second one, y_7 is covered, which comprises to the samples taken between 600 and 800 minutes after injury. If uncovered, y_6 receives the class label of y_7 (and vice versa) with a purity of over 80%. Both cascades were not able to cover three classes that reflect earlier time points (y_2 , y_4 , y_5). These external classes are assigned to their assumed ordinal neighbours (within the cascade) with a purity of at least 90%. As the class definitions are again based on prior assumptions they most likely do not correspond to large changes in the process. This can be also seen in the minimal class-wise sensitivity of the binary classifiers, which are below 0.8 for time-based neighbouring classes, except for the first class y_1 and the control class y_8 . Based on the classification tendency of the uncovered classes one might conclude that the expression profile of the first 50 minutes after injury and the control one is quite distinct from the process in between. Furthermore the assignment of the uncovered and alternative classes are consistent with the findings that wound healing consists of overlapping phases [51].

HPA-axis methylation data. Not considering both directions and the classes at the two last cascade positions the subcascades of d_8 reduce to two orderings:

(*liver* < *skin* < *hippocampus* / *hypothalamus* < *pituitary* < *adrenal*). One part of one cascade *hypothalamus* < *pituitary* < *adrenal* is reported as HPA-axis and known to control stress reaction. The role of the hippocampus in the control of stress reaction and its interconnection towards the hypothalamus might lead to the effect that hippocampus and hypothalamus are not discriminable anymore if a relation reflecting the interplay of stress response is considered. The neighbourhood between *skin* and the HPA-axis within the cascade reflects the findings by Agba et al. [1] that methylation patterns in skin are closely related to the patterns within the HPA-axis.

6 Discussion and Conclusion

Although the task of classification is mainly based on the assumption of pairwise distinct concepts, it is highly unlikely that no other relationship among the classes exists. Nevertheless, these relations might be implicit or even unknown. They might be revealed by analysing suitable data representations.

In this work, we utilise ordinal classifier cascades for detecting ordinal relations in each subset of classes. This multi-class architecture is highly sensitive to the order of classes as decision regions of early classes overlay decision regions of later ones. Interestingly, this property causes more rejections than the performance of the corresponding base classifiers. In our study, almost all classes would have been separable in pairwise classification experiments.

From a combinatorial point of view the number of candidate cascades increases exponentially in the number of classes. The proposed exhaustive screening is therefore out of range for approaches that require a de novo generation of all classification models. It would require the training of an ordinal classifier for each subset of classes and each permutation thereof. In our approach we circumvent this complexity by the extensive use of memoization techniques and theoretical error bounds. This allowed us to perform exhaustive screens for datasets with 26 classes, which corresponds to the evaluation of more than 10^{27} candidate orders. The runtime of our approach is mainly determined by the training of the required number of base classifiers which is quadratic in the overall number classes. It therefore brings into range ordinal classifier cascades based on more sophisticated but also more complex base classifiers [14,15,25,38,46,59]. To our knowledge, our screening is the first one that applies memoization techniques to ordinal classification. It might be a blue print for other ordinal classifiers or fusion architectures.

Due to the transitive relationship between the minimal class-wise sensitivities of a cascade and its subcascades, the probability of larger ordinal structures decreases. Long ordinal subcascades must be seen as rare events. As such, we consider the longest cascades that pass the highest threshold on the minimal class-wise sensitivity as being the most informative ones. Of course, both length and threshold are clearly dependent on the classification task and the chosen data representation. There is no guarantee that cascades of suitable length will be found.

The identified subcascades can be seen as ordered subsets of well separable classes. They might be considered as a roadmap of axes organising the complete collection of classes. However, not all classes are connected in this network. The labels of these external classes cannot be predicted by the ordinal classifier cascades. As most base classifiers do not possess a rejection option, the samples of the external classes will in general be assigned to cascade classes. Although the classification of these samples is incorrect, the association to one of the ordinal classes can reveal properties of the external classes. In our experiments, we observed a large set of external classes that are assigned to a small set of consecutive classes of an ordinal classifier cascade. They might be a hint to a too fine granular classification (e.g. caused by inappropriate sampling), which cannot be separated in feature space. The whole process of identifying subcascades can also be seen as a selection of a maximal number of classes that can collectively be discriminated. This in turn can give rise to new hypothesis about the processes involved like “Is this entity really a distinct new group?”, or “There could be some stagnation in the development after a certain time”.

Overall, our experiments show that ordinal substructures of classes can be detected in feature space and that they corroborate existing findings. These structures might be seen as hypotheses for ordinal relations and also for discriminative entities among the corresponding class concepts, which can be investigated in a more detailed analysis. From a theoretical point of view, the aggregation of ordinal subcascades remains an open research question. It might be addressed in form of multi-class architectures that directly address partial orders of classes.

Acknowledgements We thank Konstanze Döhner and Thomas Barth for helpful comments on some of the found orderings. The research leading to these results has received funding from the German Research Founda-

tion (DFG, GRK 2254 HEIST and SFB 1074 project Z1), and the Federal Ministry of Education and Research (BMBF, e:Med, conFirm, id 01ZX1708C) all to HAK.

Funding Open Access funding enabled and organized by Projekt DEAL.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

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

1. Agba O, Lausser L, Huse K, Bergmeier C, Jahn N, Groth M, Bens M, Sahm A, Gall M, Witte O, Kestler HA, Schwab M, Platzer M (2017) Tissue-, sex-, and age-specific DNA methylation of rat glucocorticoid receptor gene promoter and insulin-like growth factor 2 imprinting control region. *Physiol Genomics* 49(11):690–702
2. Allain E, Venzl K, Caron P, Turcotte V, Simonyan D, Gruber M, Le T, Lévesque E, Guillemette C, Vanura K (2018) Sex-dependent association of circulating sex steroids and pituitary hormones with treatment-free survival in chronic lymphocytic leukemia patients. *Ann Hematol* 97(9):1649–1661
3. Alpaydin E, Kaynak C (1998) Cascaded classifiers. *Kybernetika* 34:369–374
4. Arber D, Orazi A, Hasserjian R, Thiele J, Borowitz M, Le Beau M, Bloomfield C, Cazzola M, Vardiman JW (2016) The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* 127(20):2391–2405
5. Barnes P (1984) The third nervous system in the lung: physiology and clinical perspectives. *Thorax* 39(8):561–567
6. Baugh LR, Hill AA, Claggett JM, Hill-Harfe K, Wen JC, Slonim DK, Brown EL, Hunter CP (2005) The homeodomain protein PAL-1 specifies a lineage-specific regulatory network in the *C. elegans* embryo. *Development* 132(8):1843–1854
7. Baylis C (2009) Sexual dimorphism in the aging kidney: differences in the nitric oxide system. *Nat Rev Nephrol* 5(7):384–396
8. Berman J (2004) Tumor classification: molecular analysis meets Aristotle. *BMC Cancer* 4(1):10
9. Bishop C (2006) Pattern recognition and machine learning. Springer, New York
10. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. The Wadsworth statistics/probability series. Chapman and Hall/CRC, Boca Raton
11. Cardoso J, Pinto da Costa J (2007) Learning to classify ordinal data: the data replication method. *J Mach Learn Res* 8:1393–1429
12. Crammer K, Singer Y (2001) Pranking with ranking. In: Dietterich T, Becker S, Ghahramani Z (eds) Proceedings of the 14th international conference on neural information processing systems: natural and synthetic. Advances in neural information processing systems, vol 14. MIT Press, Cambridge, pp 641–647
13. Dheeru D, Karra TE (2017) UCI machine learning repository
14. Ding S, Zhang N, Zhang X, Wu F (2017) Twin support vector machine: theory, algorithm and applications. *Neural Comput Appl* 28(11):3119–3130
15. Ding S, Zhao X, Zhang J, Zhang X, Xue Y (2019) A review on multi-class TWSVM. *Artif Intell Rev* 52(2):775–801
16. Edla D, Jana P (2012) A prototype-based modified DBSCAN for gene clustering. *Procedia Technol* 6:485–492
17. Fanty M, Cole R (1991) Spoken letter recognition. In: Lippmann RP, Moody JE, Touretzky DS (eds) Advances in neural information processing systems 3. Morgan-Kaufmann, New York, pp 220–226

18. Frank E, Hall M (2001) A simple approach to ordinal classification. In: Raedt LD, Flach P (eds) Proceedings of the machine learning: ECML 2001—12th European conference on machine learning, Freiburg, Germany, September 5–7, 2001, lecture notes in artificial intelligence, vol 2167. Springer, Berlin, pp 145–156
19. Frank E, Kramer S (2004) Ensembles of nested dichotomies for multi-class problems. In: Proceedings of the 21st international conference of machine learning (ICML-2004). ACM Press, London, pp 305–312
20. Ganor Y, Levite M (2014) The neurotransmitter glutamate and human T cells: glutamate receptors and glutamate-induced direct and potent effects on normal human T cells, cancerous human leukemia and lymphoma T cells, and autoimmune human T cells. *J Neural Transm* 121(8):983–1006
21. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. *J Mach Learn Res* 3:1157–1182
22. Hastie T, Tibshirani R, Friedman JH (2001) The elements of statistical learning. Springer, New York
23. Hühn J, Hüllermeier E (2009) Is an ordinal class structure useful in classifier learning? *J Data Min Model Manag* 1(1):45–67
24. Japkowicz N, Shah M (2011) Evaluating learning algorithms: a classification perspective. Cambridge University Press, New York
25. Jayadeva A, Khemchandani R, Chandra S (2007) Twin support vector machines for pattern classification. *IEEE Trans Pattern Anal Mach Intell* 29(5):905–910
26. Jiang Z, Sun G, Gu Q, Chen D (2014) An ordinal multi-class classification method for readability assessment of Chinese documents. In: Buchmann R, Kifer CV, Yu J (eds) Knowledge science, engineering and management. Springer, Cham, pp 61–72
27. Kestler HA, Lausser L, Lindner W, Palm G (2011) On the fusion of threshold classifiers for categorization and dimensionality reduction. *Comput Stat* 26(2):321–340
28. Key T (1995) Hormones and cancer in humans. *Mutat Res Fundam Mol Mech Mutagen* 333(1):59–67
29. Kohlmann A, Kipps T, Rassenti L, Downing J, Shurtleff S, Mills K, Gilkes A, Hofmann WK, Basso G, Dell'Orto M, Foà R, Chiaretti S, De Vos J, Rauhut S, Papenhausen P, Hernández J, Lumberas E, Yeoh A, Koay E, Li R, Wm Liu, Williams P, Wiecezorek L, Haferlach T (2008) An international standardization programme towards the application of gene expression profiling in routine leukaemia diagnostics: the microarray innovations in LEukemia study prephase. *Br J Haematol* 142(5):802–807
30. Kohonen T (1995) Self-organizing maps, vol I. Springer, Berlin
31. Kotsiantis S, Pintelas P (2004) A cost sensitive technique for ordinal classification problems. In: Vouros G, Panayiotopoulos T (eds) Proceedings of the methods and applications of artificial intelligence: third hellenic conference on AI (SETN 2004), Samos, Greece, May 5–8, 2004. Springer, Berlin, pp 220–229
32. Lam A, Freedman B, Morizane R, Lerou P, Valerius M, Bonventre J (2014) Rapid and efficient differentiation of human pluripotent stem cells into intermediate mesoderm that forms tubules expressing kidney proximal tubular markers. *J Am Soc Nephrol* 25(6):1211–1225
33. Latke R, Lausser L, Müssel C, Kestler HA (2015) Detecting ordinal class structures. In: Schwenker F, Roli F, Kittler J (eds) Proceedings of the multiple classifier systems—12th international workshop (MCS 2015), Günzburg, Germany, June 29–July 1, 2015. Image processing, computer vision, pattern recognition, and graphics, vol 9132. Springer, Cham, pp 100–111
34. Lausser L, Müssel C, Kestler HA (2013) Measuring and visualizing the stability of biomarker selection techniques. *Comput Stat* 28(1):51–65
35. Lausser L, Schmid F, Platzer M, Sillanpää MJ, Kestler HA (2016) Semantic multi-classifier systems for the analysis of gene expression profiles. *Arch Data Sci Ser A* 1(1):157–176
36. Lausser L, Szekely R, Schirra LR, Kestler HA (2017) The influence of multi-class feature selection on the prediction of diagnostic phenotypes. *Neural Process Lett* 48(2):863–880
37. Lausser L, Schäfer LM, Schirra LR, Szekely R, Schmid F, Kestler HA (2019) Assessing phenotype order in molecular data. *Sci Rep* 9(1):11746
38. Lausser L, Szekely R, Klimmek A, Schmid F, Kestler HA (2020) Constraining classifiers in molecular analysis: invariance and robustness. *J R Soc Interface* 17(163):20190612
39. Lin HT, Li L (2012) Reduction from cost-sensitive ordinal ranking to weighted binary classification. *Neural Comput* 24(5):1329–1367
40. Lorena AC, de Carvalho ACPLF, Gama JMP (2009) A review on the combination of binary classifiers in multiclass problems. *Artif Intell Rev* 30:19–37
41. Müssel C, Lausser L, Maucher M, Kestler HA (2012) Multi-objective parameter selection for classifiers. *J Stat Soft* 46(5):1–27
42. Nicoll R, Malenka R, Kauer J (1990) Functional comparison of neurotransmitter receptor subtypes in mammalian central nervous system. *Physiol Rev* 70(2):513–565
43. Pfister T, Reinhold W, Agama K, Gupta S, Khin S, Kinders R, Parchment R, Tomaszewski J, Doroshow J, 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 Therap* 8(7):1878–1884
44. Platt JC, Shawe-Taylor J, Cristianini N (1999) Large margin DAG's for multiclass classification. In: Solla SA, Leen TK, Müller K (eds) *Proceedings of the 12th international conference on neural information processing systems: mini-symposium on causality in time series, advances in neural information processing systems*, vol 12. MIT Press, Cambridge, pp 547–553
 45. Rivest RL (1987) Learning decision lists. *Mach Learn* 2(3):229–246
 46. Schwenker F, Kestler HA, Palm G (2001) Three learning phases for radial-basis-function networks. *Neural Netw* 14(4–5):439–458
 47. Taudien S, Lausser L, Giamarellos-Bourboulis EJ, Sponholz C, FS, Felder M, Schirra LR, Schmid F, Gogos C, SG, 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
 48. Valdivielso J, Jacobs-Cachá C, Soler MJ (2019) Sex hormones and their influence on chronic kidney disease. *Curr Opin Nephrol Hypertens* 28(1):1–9
 49. Vapnik VN (1998) *Statistical learning theory*. Wiley, New York
 50. Waegeman W, Baets BD, Boullart L (2008) Roc analysis in ordinal regression learning. *Pattern Recognit Lett* 29(1):1–9
 51. Wang PH, Huang BS, Horng HC, Yeh CC, Chen YJ (2018) Wound healing. *Chin Med Assoc* 81(2):94–101
 52. Webb AR (2002) *Statistical pattern recognition*, 2nd edn. Wiley, Chichester
 53. Wehrens R, Buydens L (2007) Self- and super-organizing maps in R: the Kohonen package. *J Stat Softw* 21(5):1–19
 54. Wehrens R, Kruisselbrink J (2018) Flexible self-organizing maps in Kohonen 3.0. *J Stat Softw* 87(7):1–18
 55. Wiernik P, Dutcher J, Gertz M (2018) *Neoplastic diseases of the blood*. Springer, Berlin
 56. Xiao W, Mindrinos M, Seok J, Cuschieri J, Cuenca A, Gao H, Hayden D, Hennessy L, Moore E, Minei JP, Bankey P, Johnson J, Sperry J, Nathens A, Billiar T, West M, Brownstein B, Mason P, Baker H, Finnerty C, Jeschke M, López MC, Klein M, Gamelli R, Gibran N, Arnoldo B, Xu W, Zhang Y, Calvano S, McDonald-Smith G, Schoenfeld D, Storey J, Cobb J, Warren H, Moldawer L, Herndon D, Lowry S, Maier R, Davis R, Tompkins R (2011) A genomic storm in critically injured humans. *J Exp Med* 208(13):2581–2590
 57. Young W, Goy R, Phoenix C (1964) Hormones and sexual behavior. *Science* 143(3603):212–218
 58. Zárate S, Stevnsner T, Gredilla R (2017) Role of estrogen and other sex hormones in brain aging: neuro-protection and DNA repair. *Front Aging Neurosci* 9:430
 59. Zhang N, Ding S, Zhang J, Xue Y (2018) An overview on restricted Boltzmann machines. *Neurocomputing* 275:1186–1199

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