Statistical Comparisons of Classifiers by Generalized Stochastic Dominance

Christoph Jansen CHRISTOPH.JANSEN@STAT.UNI-MUENCHEN.DE
Department of Statistics
Ludwig-Maximilians-Universität
Ludwigstr. 33, 80539 Munich, Germany

Malte Nalenz MALTE.NALENZ@STAT.UNI-MUENCHEN.DE
Department of Statistics
Ludwig-Maximilians-Universität
Ludwigstr. 33, 80539 Munich, Germany

Georg Schollmeyer GEORG.SCHOLLMEYER@STAT.UNI-MUENCHEN.DE
Department of Statistics
Ludwig-Maximilians-Universität
Ludwigstr. 33, 80539 Munich, Germany

Thomas Augustin THOMAS.AUGUSTIN@STAT.UNI-MUENCHEN.DE
Department of Statistics
Ludwig-Maximilians-Universität
Ludwigstr. 33, 80539 Munich, Germany

Abstract

Although being a question in the very methodological core of machine learning, there is still no unanimous consensus on how to compare classifiers. Every comparison framework is confronted with (at least) three fundamental challenges: the multiplicity of quality criteria, the multiplicity of data sets and the randomness / arbitrariness of the selection of data sets. In this paper, we add a fresh view to the vivid debate by adopting recent developments in decision theory. Our resulting framework, based on so-called preference systems, ranks classifiers by a generalized concept of stochastic dominance, which powerfully circumvents the cumbersome, and often even self-contradictory, reliance on aggregates. Moreover, we show that generalized stochastic dominance can be operationalized by solving easy-to-handle linear programs and statistically tested by means of an adapted two-sample observation-randomization test. This indeed yields a powerful framework for the statistical comparison of classifiers with respect to multiple quality criteria simultaneously. We illustrate and investigate our framework in a simulation study and with standard benchmark data sets.

Keywords: algorithm comparison, statistical test, generalized stochastic dominance, preference system, decision theory

1. Introduction

1.1 Background

With a surge of new classification algorithms, a statistically sound way to decide if a method improves on its competitors is of great importance. This task has eo ipso a multi-dimensional structure: one typically compares several classifiers over several data sets relying on several criteria, with accuracy, AUC, and Brier score being popular choices. Depending on

the specific field, also other criteria, such as model size, interpretability, and computational demand may be of interest (see, e.g., the systematic taxonomy given in Lavesson and Davidsson (2007) and the minimum requirements discussed in Yu and Kumbier (2020)). In addition, in the last years, following Demšar (2006), attention has also been paid to the question of statistical significance of observed differences, understanding the investigated data sets as a sample from a virtual universe of potential data sets.

If the quality criterion is commensurable over data sets, then tests on the differences of means may be used, e.g., pairwise t-tests. An immediate way to reach conclusions from a multidimensional structure is to summarize the different components by a real-valued quantity, for instance by considering rank aggregates of some quality criteria. Then also tests in the vain of Demšar (2006) can be used to judge statistical significance. Typical studies proceeding that way include, for instance, Fernández-Delgado et al. (2014), where algorithms are ranked based on their mean accuracy on each data set, Ismail Fawaz et al. (2019), where time series classifiers are compared, or Graczyk et al. (2010), where different neural networks for regression are investigated.

Although being quite intuitive at first glance, such approaches also have been shown to have severe shortcomings. Of course, the used aggregation procedure suffers from substantial arbitrariness. Moreover, even rather intuitively plausible aggregation procedures may show quite paradoxical behavior; see, e.g., Benavoli et al. (2016), who substantially relativize any comparison based on rank aggregation by demonstrating that adding classifiers to the evaluation can change the comparison between the originally compared classifiers.

A systematic discussion and a clarified understanding of the problem of classifier comparison could benefit from embedding this problem into the framework of choice theory (see, e.g., Brams and Fishburn (2002) for a classical review), where the proper aggregation of preferences is a well-established topic; see, for instance, Eugster et al. (2012) or Mersmann et al. (2015) for the connection between benchmarking algorithms and social choice theory. However, relying on the classical setting of choice theory typically brings negative results, but helps to understand paradoxical results better and to elaborate on the principled fundamental difficulties of our comparison task. Indeed, following Arrow's impossibility theorem (Arrow, 1950), in general, there is no aggregation rule that satisfies a set of seemingly intuitive minimum requirements. More precisely, any function aggregating tuples of preference relations to a single one while satisfying a Pareto and an independence condition can be shown to be dictatorial: implicitly, only one of the quality criteria is considered. This very general statement includes the Borda rule (originally proposed in de Borda (1781)), in which classifiers are compared on the basis of their average ranks with respect to the various quality criteria.

Unfortunately, even leaving the Arrovian framework and searching for seemingly convincing aggregation rules outside it may not be successful. In particular, also the Condorcet rule (originally proposed in de Condorcet (1785)), where a ranking between any pair of classifiers is derived by counting which one performs better with respect to more quality criteria, has undesirable properties. It is known to produce intransitive and – even worse – cyclic rankings for certain constellations.

Before we look deeper into recent developments of modern choice and decision theory founding our proposal, it seems helpful to precisely formalize our problem.

1.2 Specification and Formalization of the Problem

To formally describe the different levels at which problems can arise when comparing classifiers, we use the following notation, which will be revisited in Section 3: Let $\mathcal{C}$ denote the set of classifiers available in the problem under consideration and $\mathcal{D}$ denote a set of data sets with respect to which the classifiers are to be compared. Let further be $\phi_1, \ldots, \phi_n : \mathcal{C} \times \mathcal{D} \to \mathbb{R}$ different criteria to measure the goodness of classification of the different classifiers on the different data sets (note that the assumption of all criteria mapping to $\mathbb{R}$ is only for the simplicity of presentation and will be dropped later in Section 3). The structure of the problem is summarized in Table 1 for the situation that $\mathcal{C} = \{C_1, \ldots C_q\}$ and $\mathcal{D} = \{D_1, \ldots, D_s\}$. A closer look at Table 1 now directly shows three different levels of

classifier \ data sets	D_1	$\ldots$	D_s
C_1	$\begin{pmatrix} \phi_1(C_1, D_1) \\ \vdots \\ \phi_n(C_1, D_1) \end{pmatrix}$	$\ldots$	$\begin{pmatrix} \phi_1(C_1, D_s) \\ \vdots \\ \phi_n(C_1, D_s) \end{pmatrix}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_q	$\begin{pmatrix} \phi_1(C_q, D_1) \\ \vdots \\ \phi_n(C_q, D_1) \end{pmatrix}$	$\ldots$	$\begin{pmatrix} \phi_1(C_q, D_s) \\ \vdots \\ \phi_n(C_q, D_s) \end{pmatrix}$

Table 1: A schematic presentation of the problem of comparing different classifiers over multiple data sets with respect to multiple quality criteria simultaneously.

challenges when comparing classifiers:

Level 1: In the case of multiple quality criteria, two classifiers can generally not be trivially compared already on *one single* data set. For instance, consider a situation with conflicting quality criteria such as $\phi_1(C_1, D) > \phi_1(C_2, D)$ but at the same time $\phi_2(C_1, D) < \phi_2(C_2, D)$. This problem is also strongly connected with multi-objective optimization, where one can find and analyze the Pareto-optimal points (Müssel et al., 2012; Deb, 2014). However, without further assumptions, no decision between the classifiers can be made in such situations: The component-wise dominance relation is only a *partial order* (compare Section 2 for details). Another solution is the introduction of measures that combine and trade off multiple dimensions into one single number (e.g. Brazdil et al. (2003) or Marler and Arora (2010)).

Level 2: Even if the problem in Level 1 can be circumvented somehow (for instance if we indeed happen to have component-wise dominance), the rank order of classifiers that holds over one fixed data set may change or even completely reverse over *another* data set. For instance it might hold that $\phi_i(C_1, D_1) > \phi_i(C_2, D_1)$ for all $i \in \{1, \ldots, n\}$ but there exists some $i_0 \in \{1, \ldots, n\}$ such that $\phi_{i_0}(C_1, D_2) < \phi_{i_0}(C_2, D_2)$. This makes the comparison of

classifiers a decision problem *under uncertainty* about the data sets and, of course, this uncertainty should be adequately included in any further analysis of the problem.

Level 3: Since both the set of all relevant data sets and their probability distribution will in general be unknown, it is often impossible to analyze the decision problem from Level 2 in practice. Instead, one can only analyze an empirical analogon of the problem over a sample of data sets. This means that even if one has found ways to meaningfully solve the problems of Levels 1 and 2 and thus could define a meaningful order of classifiers for the concrete sample of data sets, a different order of classifiers could occur as soon as another sample of data sets is considered. The solution of such an empirical decision problem is subject to *statistical uncertainty*. It is desirable to be able to control this statistical uncertainty by constructing a suitable statistical test. For single criteria evaluation tests have been proposed in Demšar (2006), see also Corani et al. (2017) for a Bayesian variant.

1.3 Sketch of our Proposal and Overview

In the present paper, we propose a framework that allows a comparison of classifiers with respect to multiple quality criteria over multiple data sets *simultaneously* (therefore accounting for the problems of Levels 1 and 2). For also accounting for the statistical uncertainty arising from the specific selection of the benchmark data sets (as described in Level 3), we also propose a permutation-based statistical test for our criterion. Our comparison criterion is based on a generalized notion of first-order stochastic dominance and, in general, provides only an incomplete ranking of the classifiers. Here, however, the ordering power of our dominance relation can be explicitly modelled by a parameter whose increase attenuates each quality dimension to the same extent instead of being limited to only one dimension.

Generally, addressing problems of multiplicity by stochastic dominance appears promising. However, while approaches in this spirit are common in several fields ranging from biometrics (e.g., Davidov and Peddada (2013)) to econometrics (e.g., Whang (2019)), the authors are not aware of such contributions in machine learning.

To derive and discuss our framework, the paper is organized as follows: Section 2 recalls the required mathematical definitions. Section 3 introduces the concept of δ-dominance between classification algorithms, while Section 4 gives an algorithm for detecting δ-dominance and discusses how to test for it if only a sample of data sets is available. Sections 5 and 6 demonstrate the ideas presented on simulated data and with standard benchmark data sets. Section 7 concludes by elaborating on some promising perspectives for future research.

2. Preliminiaries

Throughout the paper, we will consider *binary relations* at several points, be it on the set of all quality vectors as in Equation (1), on a binary relation itself as in Equation (2), or on the set of all classifiers as in Definition 6. We, therefore, begin with a compilation of some important concepts in this context. First, recall that a binary relation R on a non-empty set M is a subset of the Cartesian product of the set with itself, that is $R \subseteq M \times M$. Several (potential) properties of binary relation occur in what follows: $R \subseteq M \times M$ is called

- *reflexive*, if $(m, m) \in R$,

- *transitive*, if $(m_1, m_2), (m_2, m_3) \in R$ implies $(m_1, m_3) \in R$,
- *antisymmetric*, if $(m_1, m_2), (m_2, m_1) \in R$ implies $m_1 = m_2$,
- *complete*, if $(m_1, m_2) \in R$ or $(m_2, m_1) \in R$ (or both)

for arbitrary elements $m, m_1, m_2, m_3 \in M$. A *preference relation* is a binary relation that is complete and transitive; a *pre-order* is a binary relation that is reflexive and transitive; a *linear order* is a preference relation that is antisymmetric; a *partial order* is a pre-order that is antisymmetric.

Equipped with these concepts, we can now define the central ordering structure for us, so-called *preference systems*. With the help of these systems, it is possible to model ordered sets on which the scale of measurement can vary locally: The (potential partial) ordinal part of the system is modelled by a pre-order on the set itself, while the (potential partial) metric part of the system is modelled by a pre-order on the ordinal relation, which covers those parts of the set for which a strength of order can also be specified. The following Definitions 1, 2, and 3 have been introduced in a decision-theoretic context by Jansen et al. (2018) and are also discussed in Jansen et al. (2022). As these form the basis of our generalized stochastic dominance concept, they are listed here for further reference.

Definition 1 *Let A be a non-empty set and let $R_1 \subseteq A \times A$ denote a pre-order on A. Moreover, let $R_2 \subseteq R_1 \times R_1$ denote a pre-order on R_1. Then the triplet $\mathcal{A} = [A, R_1, R_2]$ is called a* ***preference system*** *on A.*

Originally stemming rather from a decision-theoretic context, the relations R_1 and R_2 are usually given a behavioristic interpretation: If $(a, b) \in R_1$, this is interpreted as *a being at least as desirable as b*, whereas if $((a, b), (c, d)) \in R_2$, this is interpreted as *exchanging b by a being at least as desirable as exchanging d by c*. However, such an interpretation is far from mandatory: As will be shown in Section 3, preference systems are also perfectly suited to analyze multidimensional data structures with both ordinal and metric dimensions. Note that the specific preference system based on the comparison of multiple quality measures for classifiers simultaneously will be defined in Equations (1), (2) and (3).

Since the formal definition of a preference system does not restrict the interaction of the involved relations R_1 and R_2 in any way, usually, an additional consistency criterion is introduced. Here, for a pre-order $R \subseteq M \times M$ on a set M, we denote by $P_R \subseteq M \times M$ its *strict part* and by $I_R \subseteq M \times M$ its *indifference part*, respectively defined by

$$(m_1, m_2) \in P_R \Leftrightarrow (m_1, m_2) \in R \wedge (m_2, m_1) \notin R,$$

$$(m_1, m_2) \in I_R \Leftrightarrow (m_1, m_2) \in R \wedge (m_2, m_1) \in R.$$

This leads to the following definition.

Definition 2 *The preference system $\mathcal{A} = [A, R_1, R_2]$ is* ***consistent*** *if there exists a function $u : A \to [0, 1]$ such that for all $a, b, c, d \in A$ we have:*

i) If $(a, b) \in R_1$, then $u(a) \geq u(b)$ with $=$ iff $(a, b) \in I_{R_1}$.

ii) If $((a, b), (c, d)) \in R_2$, then $u(a) - u(b) \geq u(c) - u(d)$ with $=$ iff $((a, b), (c, d)) \in I_{R_2}$.

The set of all such ***representations*** *u satisfying i) and ii) is denoted by $\mathcal{U}_{\mathcal{A}}$.*

In the case of a consistent preference system that possesses R_1-minimal and R_1-maximal elements in A, it may be useful to consider only representations that measure the utility of consequences on the same scale. This proves of particular importance for the parameter δ from Definitions 5 and 6 to have a meaningful interpretation in terms of regularization. We have the following definition.

Definition 3 *Let $\mathcal{A} = [A, R_1, R_2]$ be a consistent preference system containing $a_*, a^* \in A$ such that $(a^*, a) \in R_1$ and $(a, a_*) \in R_1$ for all $a \in A$. Then*

$$\mathcal{N}_{\mathcal{A}} := \left\{ u \in \mathcal{U}_{\mathcal{A}} : u(a_*) = 0 \ \wedge \ u(a^*) = 1 \right\}$$

is called the ***normalized representation set*** *of $\mathcal{A}$. Further, for a number $\delta \in [0, 1)$, we denote by $\mathcal{N}_{\mathcal{A}}^{\delta}$ the set of all $u \in \mathcal{N}_{\mathcal{A}}$ satisfying*

$$u(a) - u(b) \geq \delta \quad \wedge \quad u(c) - u(d) - u(e) + u(f) \geq \delta$$

*for all $(a, b) \in P_{R_1}$ and for all $((c, d), (e, f)) \in P_{R_2}$. We call $\mathcal{A}$ δ-**consistent** if $\mathcal{N}_{\mathcal{A}}^{\delta} \neq \emptyset$.*

Finally, we will define the notion of an automorphism in the context of preference systems, as we need this notion later in Proposition 1 and Proposition 2.

Definition 4 *Let $\mathcal{A} = [A, R_1, R_2]$ be a preference system. A mapping $T : A \longrightarrow A$ is called* ***automorphism*** *if it is bijective and if furthermore for arbitrary $a, b, c, d \in A$ we have*

$$(a, b) \in R_1 \Leftrightarrow (T(a), T(b)) \in R_1 \text{ and}$$
$$((a, b), (c, d)) \in R_2 \Leftrightarrow ((T(a), T(b)), (T(c), T(d)) \in R_2.$$

3. Comparing Classifiers by Generalized Stochastic Dominance

We now propose a criterion with which it is possible to compare classification algorithms with respect to multiple quality measures on multiple data sets *simultaneously*. The basic idea of our criterion is based on the concept of *first-order stochastic dominance* for random variables over partially ordered sets, which (essentially) states: a variable X is stochastically greater or equal than another variable Y if the expectation of the variable $u \circ X$ is greater or equal than the expectation of the variable $u \circ Y$ for every real-valued representation u of the underlying partial order (see, e.g., Mosler and Scarsini (1991) for more details). Intuitively spoken, this means that X has a greater or equal expected value in every metric interpretation of the underlying space that does not contradict the given ordering structure.

In the context of comparing classification algorithms, the variables under consideration are functions associated with the various classifiers that assign a vector of quality values, or short *quality vector*, to each possible data set. The set of all these quality vectors is then partially ordered by the component-wise greater or equal relation (see Equation (1) below). However, since some of the considered quality measures may also be interpretable on a metric scale (e.g. if one considers *prediction accuracy* as a quality measure), there is even more structure. To be able to exploit this partial cardinal structure, we suitably define a preference system

on the set of quality vectors (see Equation (2) below). A natural generalization of stochastic dominance is then to require the expectation dominance mentioned above no longer for all representations of the component-wise partial order, but only for all representations of the constructed preference system in the sense of Definitions 2 and 3.

3.1 Generalized Stochastic Dominance

Before turning to the construction just described in detail, we give the following central definition of generalized stochastic dominance over preference systems for arbitrary random variables. This relation can be viewed as a generalization of first-order stochastic dominance in two respects: First, as just discussed, the relation R_2 can also include partial metric information. Second, the parameter δ allows to explicitly model from which threshold on a difference in utility should be included in the analysis of the random variables.

Definition 5 *Let $\mathcal{A} = [A, R_1, R_2]$ be a δ-consistent preference system and let $[S, \sigma(S), \pi]$ be a probability space. Denote by*

$$\mathcal{F}_{(\mathcal{A},S)} := \Big\{ X \in A^S : u \circ X \text{ is } \sigma(S)\text{-}\mathcal{B}_{\mathbb{R}}([0,1])\text{-measurable for all } u \in \mathcal{U}_{\mathcal{A}} \Big\}$$

*For random variables $X, Y \in \mathcal{F}_{(\mathcal{A},S)}$, we say that X $(\mathcal{A}, \pi, \delta)$-**dominates** Y, abbreviated with $X \geq_{(\mathcal{A},\pi,\delta)} Y$, whenever it holds that*

$$\mathbb{E}_\pi(u \circ X) \geq \mathbb{E}_\pi(u \circ Y)$$

for all normalized representations $u \in \mathcal{N}_{\mathcal{A}}^{\delta}$ respecting the threshold δ.

Remark 1 *Consider again the situation of Definition 5. For the special case of $R_2 = \emptyset$ and $\delta = 0$ the relation $\geq_{(\mathcal{A},\pi,0)}$ essentially reduces to classical first-order stochastic dominance on partially ordered sets. Further, for the special case of $\delta = 0$ and relations R_1 and R_2 that are compatible in the sense of satisfying the axioms in Krantz et al. (1971, Definition 1, p. 147) and, thus, admitting a representation that is unique up to positive linear transformations, the relation $\geq_{(\mathcal{A},\pi,0)}$ essentially reduces to the classical principle of maximizing expected utility. Finally, again setting $\delta = 0$, the relation $\geq_{(\mathcal{A},\pi,0)}$ can be viewed as that special case of the relation $R_{\forall\forall}$ from Jansen et al. (2018, p. 123), where the there mentioned set of probability measures $\mathcal{M}$ is chosen to consist solely of π, that is $\mathcal{M} = \{\pi\}$ is a singleton.*

3.2 Utilizing Generalized Stochastic Dominance for Comparing Classifiers

As indicated at the beginning of the section, we now show how the relation $\geq_{(\mathcal{A},\pi,\delta)}$ can be utilized to compare classification algorithms with respect to multiple quality measures on multiple data sets simultaneously. This requires some additional notation. Let

- $\mathcal{D}$ denote the set of all *data sets* that are relevant for the classification task in question,
- $\mathcal{C}$ denote the set of all *classifiers* that intend to classify the data sets from $\mathcal{D}$,
- $\phi_i : \mathcal{C} \times \mathcal{D} \to Q_i$ denote a *criterion of classification quality* for every $i \in \{1, \ldots, n\}$,

- $\phi := (\phi_1, \ldots, \phi_n) : \mathcal{D} \times \mathcal{C} \to \mathcal{Q}$, where $\mathcal{Q} := Q_1 \times \cdots \times Q_n$ is the set of *quality vectors.*[1]

Specifically, for a data set $D \in \mathcal{D}$ and a classifier $C \in \mathcal{C}$, the (not necessarily numerical) *reward* $\phi_i(C, D)$ is interpreted as the quality of the classifier C for data set D with respect to the classification quality criterion ϕ_i. Importantly, note that the different reward sets $Q_1, \ldots, Q_n$ *are not* assumed to be of the same scale of measurement. In particular, this implies that some of the sets are of ordinal scale (i.e. are equipped with *a preference order but no metric*), while others allow for a metric interpretation (i.e. are equipped with *both a preference order and a metric*). However, all of them are assumed to be of *at least* ordinal scale and to possess minimal and maximal elements. For every $i \in \{1, \ldots, n\}$, the preference order of the space Q_i will be denoted by $\geq_i$. Note already now that these assumptions directly imply that the set $\mathcal{Q}$ possesses minimal and maximal elements w.r.t. R_1 from Equation (1), ensuring the normalized representation set (see Definition 3) of the preference system $\mathbb{C}$ from Equation (3) to be well-defined.

Without loss of generality, we assume the quality criteria $(\phi_1, \ldots, \phi_n)$ to be arranged such that there exists $k \in \{1, \ldots, n\}$ for which the sets $Q_1, \ldots, Q_k$ are of metric scale, equipped with metrics $d_i : Q_i \times Q_i \to \mathbb{R}$, $i = 1, \ldots, k$, respectively, whereas the remaining sets are of ordinal scale not allowing for any meaningful metric interpretation. We then define a preference system on the set $\mathcal{Q}$ of all quality vectors by setting

$$R_1 := \Big\{ (q, p) \in \mathcal{Q} \times \mathcal{Q} : q_i \geq_i p_i \ \text{ for all } \ i = 1, \ldots, n \Big\} \tag{1}$$

$$R_2 := \Big\{ ((q, p), (r, s)) \in R_1 \times R_1 : d_i(q_i, p_i) \geq d_i(r_i, s_i) \ \text{ for all } \ i = 1, \ldots, k \Big\} \tag{2}$$

We denote the preference system which is composed of the set $\mathcal{Q}$ and the two relations just defined by $\mathbb{C}$, i.e., we have that

$$\mathbb{C} = [\mathcal{Q}, R_1, R_2]. \tag{3}$$

The two relations R_1 and R_2 can be given the following natural interpretation:

Interpretation of R_1: Assume we have $D \in \mathcal{D}$ and $C_i, C_j \in \mathcal{C}$ such that $\phi(C_i, D) = q$ and $\phi(C_i, D) = p$. Then $(q, p) \in R_1$ means that classifier C_i has at least as high quality as classifier C_j for *every* considered quality measure, when evaluated on data set D.

Interpretation of R_2: Assume we have $D \in \mathcal{D}$ and $C_i, C_j, C_k, C_l \in \mathcal{C}$ such that $\phi(C_i, D) = q$ and $\phi(C_i, D) = p$ and $\phi(C_k, D) = r$ and $\phi(C_l, D) = s$. Then $((q, p), (r, s)) \in R_2$ means that, when evaluated on data set D, the dominance of C_i over C_j is *at least as strong* as the the dominance of C_k over C_l. This is due to the fact that there is component-wise dominance in both cases and, additionally, the quality differences of C_i and C_j are at least as high as the quality differences of C_k and C_l for those quality measures that allow for a metric interpretation.

To take the final step of transferring the dominance criterion from Definition 5 to the comparison of classifiers, we still need to be clear about the random component in this context. This is obviously the randomness over the data sets since we are after all concerned

1. Since $\mathcal{D}$ may or may not contain labels, it is not necessary to distinguish between different types of classification tasks (such as, e.g., supervised or unsupervised) within the proposed framework.

with the expected classification quality. So, if now $[\mathcal{D}, \sigma(\mathcal{D}), \pi]$ is a suitable probability space, we can use $(\mathbb{C}, \delta)$-dominance to compare classifiers with respect to all quality criteria simultaneously. To stress the crucial role of the concept, this special case deserves a separate definition for further reference.

Definition 6 *Assume $\mathbb{C}$ to be δ-consistent. For $C_i, C_j \in \mathcal{C}$, with $\mathcal{C}$ chosen such that $\{\phi(C, \cdot) : C \in \mathcal{C}\} \subseteq \mathcal{F}_{(\mathbb{C},\mathcal{D})}$, we say that C_i δ-**dominates** C_j, abbreviated with $C_i \succsim_\delta C_j$, whenever it holds that*

$$\phi(C_i, \cdot) \geq_{(\mathbb{C},\pi,\delta)} \phi(C_j, \cdot)$$

In other words, it holds that $C_i \succsim_\delta C_j$, whenever

$$\mathbb{E}_\pi(u \circ \phi(C_i, \cdot)) \geq \mathbb{E}_\pi(u \circ \phi(C_j, \cdot))$$

for all normalized representations $u \in \mathcal{N}_\mathbb{C}^\delta$ respecting the threshold δ.

From a decision-theoretic point of view, the threshold parameter δ can be motivated by the concept of just noticeable differences discussed in the seminal work of Luce (1956): It quantifies the minimal utility difference the decision maker can notice/finds relevant given utility is measured on a $[0, 1]$-scale. Translated to the context of comparing classifiers, it rather can be seen as a regularization device: If some of the classifiers remain incomparable for a threshold of $\delta = 0$, then increasing δ provides the opportunity to strengthen the ordering power of the dominance relation while attenuating the influence of *all* quality measures used to the *same degree*. This proves particularly useful for the statistical test for δ-dominance discussed in Section 4.2: Already very small values for δ can cause a remarkable improvement of the power of the respective test, although the basic order is only marginally changed (see also the discussion in Remark 3 and Footnote 6).

3.3 Some Useful Properties of the δ-Dominance Relation

The following proposition lists some important properties of the binary relation $\succsim_\delta$ just introduced. Despite their elementary character, some of these properties will play an important role when applying the concepts in Sections 5 and 6.

Proposition 1 *Consider the same situation as in Definition 6. The following holds:*

i) For every $\xi \in [0, \delta]$, the relation $\succsim_\xi$ defines a pre-order on $\mathcal{C}$.

ii) The relations are nested with increasing δ, i.e., we have $\succsim_{\xi_1} \subseteq \succsim_{\xi_2}$ for $\xi_1 \leq \xi_2 \in [0, \delta]$.

iii) Let $T : A \to A$ be an automorphism w.r.t. $\mathbb{C}$. Then we have that $C_i \succsim_\delta C_j$ if and only if $C_i^T \succsim_\delta C_j^T$, where, for $p \in \{i, j\}$, C_p^T represents classifier C_p, but evaluated not in the space A, but in the space $T[A]$, i.e., where $\phi(C_p, \cdot)$ is replaced by $T \circ \phi(C_p, \cdot)$.

Proof i) Reflexivity is trivially true. To verify transitivity, assume that $C_i \succsim_\xi C_j$ and $C_j \succsim_\xi C_k$. Choose $u \in \mathcal{N}_\mathbb{C}^\xi$ arbitrarily (this is always possible, since δ-consistency obviously implies ξ-consistency). Then, by assumption and definition, it holds that

$$\mathbb{E}_\pi(u \circ \phi(C_i, \cdot)) \geq \mathbb{E}_\pi(u \circ \phi(C_j, \cdot)) \quad \text{and} \quad \mathbb{E}_\pi(u \circ \phi(C_j, \cdot)) \geq \mathbb{E}_\pi(u \circ \phi(C_k, \cdot))$$

directly implying

$$\mathbb{E}_\pi(u \circ \phi(C_i, \cdot)) \geq \mathbb{E}_\pi(u \circ \phi(C_k, \cdot)).$$

As u was chosen arbitrarily, this implies $C_i \succsim_\xi C_k$.

ii) Assume it holds $\xi_1 \leq \xi_2 \in [0, \delta]$. By definition and ξ-consistency for all $\xi \in [0, \delta]$ (see i)), this implies $\emptyset \neq \mathcal{N}_{\mathbb{C}}^{\xi_2} \subseteq \mathcal{N}_{\mathbb{C}}^{\xi_1}$. Assume it holds that $C_i \succsim_{\xi_1} C_j$. By definition, this implies

$$\mathbb{E}_\pi(u \circ \phi(C_i, \cdot)) \geq \mathbb{E}_\pi(u \circ \phi(C_j, \cdot))$$

for all $u \in \mathcal{N}_{\mathbb{C}}^{\xi_1}$ and, due to the super set relation, also for all $u \in \mathcal{N}_{\mathbb{C}}^{\xi_2}$. Thus $C_i \succsim_{\xi_2} C_j$.

iii) Because of $\{u \mid u \in \mathcal{N}_{\mathbb{C}}^{\delta}\} = \{u \circ T \mid u \in \mathcal{N}_{\mathbb{C}}^{\delta}\}$ we have that

$$\forall u \in \mathcal{N}_{\mathbb{C}}^{\delta} : \mathbb{E}_\pi(u \circ \phi(C_i, \cdot)) \geq \mathbb{E}_\pi(u \circ \phi(C_j, \cdot))$$

is equivalent to

$$\forall u \in \mathcal{N}_{\mathbb{C}}^{\delta} : \mathbb{E}_\pi(u \circ T \circ \phi(C_i, \cdot)) \geq \mathbb{E}_\pi(u \circ T \circ \phi(C_j, \cdot)),$$

which shows the claim. ■

Remark 2 *Benavoli et al. (2016) convincingly questioned the idea of comparing classifiers by comparing their average ranks over multiple data sets, where the ranks are computed with respect to some single quality measure ϕ. The main problem of such an approach is that the comparison of two classifiers may depend on other classifiers that are irrelevant for the problem under consideration: Specifically, if only two classifiers C_1 and C_2 are considered, C_1 may get a higher average rank than C_2, but this relation is exactly reversed if a third classifier C_3 is considered, although the quality values of C_1 and C_2 are not changed. A simple example for such a situation is given in Table 2.*

$\phi(C_i, D_j)$	D_1	D_2	D_3	D_4	D_5
C_1	0.8	0.8	0.8	0.6	0.6
C_2	0.6	0.6	0.6	0.8	0.8

$\phi(C_i, D_j)$	D_1	D_2	D_3	D_4	D_5
C_1	0.8	0.8	0.8	0.6	0.6
C_2	0.6	0.6	0.6	0.8	0.8
C_3	0.9	0.9	0.9	0.7	0.7

Table 2: On the left, C_1 receives a rank sum of 8, dominating C_2 with a rank sum of 7. However, adding a third classifier C_3 (right table) dominating C_1 and C_2 on D_1, D_2 and D_3 and lying between C_2 and C_1 for D_4 and D_5, gives C_2 and C_1 rank sums of 9 and 8, respectively. The ordering of C_1 and C_2 is reversed.

In the case where also multiple quality criteria $\phi_1, \ldots, \phi_5$ are considered, a similar situation may already occur on one specific fixed data set D. For a simple example, one can reinterpret the columns in Table 2 as the quality values $\phi_k(C_i, D)$ of the respective classifier with respect to the respective quality criteria on the fixed data set D: Comparing the average ranks of the classifiers across the quality criteria gives again reversed rankings of C_1 and C_2 for the tables on the left and on the right. Obviously, for the case of multiple classifiers and

multiple quality criteria both problems may occur at the same time, thereby even increasing the problem. Note that this fact is well-known in social choice theory: the Borda rule from voting theory (see Section 1) does not satisfy Arrow's axiom of independence of irrelevant alternatives (see, e.g., Brams and Fishburn (2002)).

Generally, it seems that any method that uses ranks (and also any cardinal relative criterion like that used in Webb (2000) and discussed in Demšar (2006)) is akine to violating independence of irrelevant alternatives. Note that in Demšar (2006) the rationale behind computing ranks is to make the quality values for different data sets commensurable. This implicitly assumes that beforehand the quality values for different data sets cannot be compared at all. If, at the same time, one also does not want to compare on one data set at least the values/ranks of different classifiers to avoid violating independence of irrelevant alternatives, then one effectively says that one quality value of one classifier for one data set cannot be compared to any other quality value at all. This obviously will lead to an unsolvable undertaking. In our approach to ranking classifiers, we do not use ranks at all and instead demand that the quality values for different data sets are commensurable or have been made commensurable beforehand. We strongly think that this is doable by specifying – in a decision-theoretic rigorous manner – an adequate loss function that provides one with adequate corresponding commensurable quality criteria.[2]

Against this background, one major advantage of comparing classifiers with respect to our dominance relation $\succsim_\delta$ instead of applying rank-based approaches, is that, in fact, independence of irrelevant alternatives is guaranteed: If it holds that $C_i \succsim_\delta C_j$, then this statement is independent of how the space $\mathcal{C} \setminus \{C_i, C_j\}$ looks like, i.e. of how many classifiers are considered in the analysis besides C_i and C_j. In this way our dominance relation circumvents one major issue which has recently been raised in the context of rank-based comparisons.

Another nice structural property of the relation $\succsim_\delta$ is that iii) in Proposition 1 is still valid for a random automorphism, as long as this random automrphism is applied independently of the process that generates the data sets. We will concretize this property in the following

Proposition 2 *Let I be an index set, let $\{T_z \mid z \in I\}$ be a family of automorphisms w.r.t. $\mathbb{C}$ and let $Z : \Omega \to I$ be an indexing random variable with law P that is independent of the process that generates the data sets D. Let, for the moment, $\mathbb{E}$ be a shorthand notation for the expectation w.r.t. the product law $\pi \otimes P$. Let now $T(\omega) := T_{Z(\omega)}$ be a random automorphism. Furthermore, assume that for every $u \in \mathcal{N}_{\mathbb{C}}^{\delta}$ the conditional expectations $\mathbb{E}(u \circ T \circ \phi(C_i, \cdot) \mid Z = z)$ and $\mathbb{E}(u \circ T \circ \phi(C_j, \cdot) \mid Z = z)$ exist. Then we have*

$$\forall u \in \mathcal{N}_{\mathbb{C}}^{\delta} : \mathbb{E}(u \circ \phi(C_i, \cdot)) \geq \mathbb{E}(u \circ \phi(C_j, \cdot)) \tag{4}$$

if and only if

$$\forall u \in \mathcal{N}_{\mathbb{C}}^{\delta} : \mathbb{E}(u \circ T \circ \phi(C_i, \cdot)) \geq \mathbb{E}(u \circ T \circ \phi(C_j, \cdot)). \tag{5}$$

2. Otherwise, decision theory would be 'science about nothing': If a decision maker is not able to quantify a gain/loss of a specific method in a specific situation, then she should not consult decision theory.

Proof Let (4) hold. Then for every **fixed** automorphism T_z and every arbitrary $u \in \mathcal{N}_{\mathbb{C}}^{\delta}$ we have $\mathbb{E}_{\pi}(u \circ T_z \circ \phi(C_i, \cdot)) \geq \mathbb{E}_{\pi}(u \circ T_z \circ \phi(C_j, \cdot))$ and therefore $\mathbb{E}(u \circ T_z \circ \phi(C_i, \cdot)) \geq \mathbb{E}(u \circ T_z \circ \phi(C_j, \cdot))$ (compare Proposition 1). This implies

$$\begin{aligned}\mathbb{E}(u \circ T \circ \phi(C_i, \cdot)) &= \int \mathbb{E}(u \circ T \circ \phi(C_i, \cdot) \mid Z = z)\, dP(z) \\ &= \int \mathbb{E}(u \circ T_z \circ \phi(C_i, \cdot))\, dP(z) \geq \int \mathbb{E}(u \circ T_z \circ \phi(C_j, \cdot))\, dP(z) \\ &= \mathbb{E}(u \circ T \circ \phi(C_j, \cdot))\end{aligned}$$

and therefore (5) holds for every arbitrary $u \in \mathcal{N}_{\mathbb{C}}^{\delta}$. The implication (5) $\implies$ (4) follows analaguously by applying the corresponding inverse (random) automorphism. ■

Proposition 2 has a nice implication: If the quality values are observed with some additional noise that can be described by a random automorphism, then the dominance criterion will not change. Note that especially a random intercept or a random scaling of the cardinal dimensions will not influence the notion of dominance. This particularly implies that in our simulation study (see Section 5) we do not need to implement such random effects.

3.4 The GSD-δ Method

As $\succsim_{\delta}$ defines a pre-order on the set $\mathcal{C}$ of all considered classifiers (see Proposition 1), it naturally induces an ordering structure on this set. The method of obtaining this ordering structure by relying on generalized stochastic dominance as the underlying relation, will be referred to as GSD-δ in the following (with GSD-0 abbreviated by GSD). In order to make this method applicable in practice, two substantial questions have to be addressed. First, the question on how to efficiently check for δ-dominance arises. Second, a test must be developed for judging if in-sample differences between classifiers are statistically significant.

4. Testing for Dominance

In this section, we first establish a linear program for checking δ-dominance between two classifiers if the set $\mathcal{D}$ of data sets is finite and the true probability law π over this set is known. Taking into account the problem described in Level 3 from Section 1, i.e., the fact that both π and the set $\mathcal{D}$ will in general be inaccessible, we then describe how to adapt this linear program to check for δ-dominance in its empirical version, i.e., in the concrete sample of data sets drawn from the distribution π. Afterwards, we discuss how the optimal value of this adapted linear program can be reinterpreted as a test statistic for a statistical test for distributional equality of the two competing classifiers and discuss how to extend this test to the complete ordering structure between all considered classifiers. Finally, in preparation for the comparative study carried out in Section 5, we briefly review the rank-based test proposed in Demšar (2006) and suggest ways to extend it to more than one quality criterion.

4.1 A Linear Program for Checking δ-Dominance

We begin by discussing a linear program for checking δ-dominance in the finite case. For that, consider again the preference system $\mathbb{C}$ as defined in Equation (3), however, with the

additional assumption that the sets $\mathcal{C}$ (the classifiers under consideration) and $\mathcal{D}$ (the data sets relevant for the comparison) are both finite. Without loss of generality, we can then assume $\mathcal{Q} = \{q_1, \ldots, q_d\}$ to be finite and that q_1 and q_2 are minimal and maximal elements of $\mathcal{Q}$ with respect to R_1, respectively.[3] Moreover, we assume $\delta \in [0, 1)$ to be chosen such that $\mathbb{C}$ is δ-consistent. A vector $(u_1, \ldots, u_d) \in [0, 1]^d$ then contains the images of a utility function $u : \mathcal{Q} \to [0, 1]$ from $\mathcal{N}_{\mathbb{C}}^{\delta}$ if it satisfies the system of linear (in-)equalities given by

- $u_1 = 0$ and $u_2 = 1$,
- $u_i = u_j$ for every pair $(q_i, q_j) \in I_{R_1}$,
- $u_i - u_j \geq \delta$ for every pair $(q_i, q_j) \in P_{R_1}$,
- $u_k - u_l = u_r - u_t$ for every pair of pairs $((q_k, q_l), (q_r, q_t)) \in I_{R_2}$ and
- $u_k - u_l - u_r + u_t \geq \delta$ for every pair of pairs $((q_k, q_l), (q_r, q_t)) \in P_{R_2}$.

Denote by $\nabla_{\mathbb{C}}^{\delta}$ the set of all vectors $(u_1, \ldots, u_d) \in [0, 1]^d$ satisfying all these (in)equalities. We then have the following proposition on how to check δ-dominance.

Proposition 3 *Consider the same situation as described above. For $C_i, C_j \in \mathcal{C}$, we consider the linear programming problem*

$$\sum_{\ell=1}^{d} u_\ell \cdot [\pi(\phi(C_i, \cdot)^{-1}(\{q_\ell\})) - \pi(\phi(C_j, \cdot)^{-1}(\{q_\ell\}))] \longrightarrow \min_{(u_1, \ldots, u_d) \in \mathbb{R}^d} \tag{6}$$

with constraints $(u_1, \ldots, u_d) \in \nabla_{\mathbb{C}}^{\delta}$. Denote by opt_{ij} the optimal value of this programming problem. It then holds that $C_i \succsim_{\delta} C_j$ if and only if $opt_{ij} \geq 0$.

Proof First, let $opt_{ij} \geq 0$. Choose $u \in \mathcal{N}_{\mathbb{C}}^{\delta}$ arbitrarily and let $g : \mathbb{R}^d \to \mathbb{R}$ denote the objective function of the linear program. We then have

$$D(u) := \mathbb{E}_\pi(u \circ \phi(C_i, \cdot)) - \mathbb{E}_\pi(u \circ \phi(C_j, \cdot)) = g(u(q_1), \ldots, u(q_d)) \geq 0 \tag{7}$$

where the equation follows by simple manipulations of the expected values and the lower bound of 0 follows since, by definition, $(u(q_1), \ldots, u(q_d)) \in \nabla_{\mathbb{C}}^{\delta}$. Since $u \in \mathcal{N}_{\mathbb{C}}^{\delta}$ was chosen arbitrarily, this implies $C_i \succsim_{\delta} C_j$.

Conversely, let $opt_{ij} < 0$. Choose $(u_1^*, \ldots, u_d^*) \in \nabla_{\mathbb{C}}^{\delta}$ to be an optimal solution yielding opt_{ij} and define $u : \mathcal{Q} \to [0, 1]$ by setting $u(q_i) := u_i^*$ for all $i = 1, \ldots, d$. We then have to distinguish two different cases:

Case 1: $\delta > 0$. One then easily verifies that $u \in \mathcal{N}_{\mathbb{C}}^{\delta}$ and

$$D(u) = g(u_1^*, \ldots, u_d^*) = opt_{ij} < 0 \tag{8}$$

Thus, u is a function from $\mathcal{N}_{\mathbb{C}}^{\delta}$ with $\mathbb{E}_\pi(u \circ \phi(C_i, \cdot)) < \mathbb{E}_\pi(u \circ \phi(C_j, \cdot))$. Thus $\neg(C_i \succsim_{\delta} C_j)$.

3. As the sets $\mathcal{C}$ and $\mathcal{D}$ are finite, it makes no difference in what follows, if we replace $\mathcal{Q}$ by the finite set $\phi(\mathcal{C} \times \mathcal{D})$. If $\phi(\mathcal{C} \times \mathcal{D})$ does not contain minimal and maximal elements, we define new vectors q_1 and q_2 containing an minimal or maximal element of Q_i in every dimension i, respectively. By re-indexing the remaining vectors and considering the finite set $\phi(\mathcal{C} \times \mathcal{D}) \cup \{q_1, q_2\}$ we are done.

Case 2: $\delta = 0$. If $u \in \mathcal{N}_{\mathbb{C}}^0$, then the same argument as in the first case applies. Thus, assume that $u \notin \mathcal{N}_{\mathbb{C}}^0$. Then, since $(u_1^*, \ldots, u_d^*) \in \nabla_{\mathbb{C}}^0$, we still know that u is monotone but we no longer have *strict* monotonicity with respect to the relations R_1 and R_2 of $\mathbb{C}$ (meaning that properties i) and ii) from Definition 2 are still valid but without the *iff* condition). Now, choose $u^+ \in \mathcal{N}_{\mathbb{C}}^0$ arbitrarily (this is always possible, since we assume 0-consistency). If $D(u^+) < 0$, then $\mathbb{E}_\pi(u^+ \circ \phi(C_i, \cdot)) < \mathbb{E}_\pi(u^+ \circ \phi(C_j, \cdot))$. This yields $\neg(C_i \succsim_\delta C_j)$. If $D(u^+) \geq 0$, then we have

$$0 \leq \xi := \frac{D(u^+)}{D(u^+) - D(u)} < 1$$

and we can choose $\alpha \in (\xi, 1)$. One then easily verifies that $u_\alpha := \alpha u + (1-\alpha)u^+ \in \mathcal{N}_{\mathbb{C}}^0$ and that $\mathbb{E}_\pi(u_\alpha \circ \phi(C_i, \cdot)) < \mathbb{E}_\pi(u_\alpha \circ \phi(C_j, \cdot))$. This again yields that $\neg(C_i \succsim_\delta C_j)$, thereby completing the proof. ■

4.2 A Statistical Test for δ-Dominance

Typically, the setting discussed in Section 4.1 will be heavily idealized as actually we are in the situation described in Level 3 from Section 1: The true probability law π on the set $\mathcal{D}$ as well as the set $\mathcal{D}$ itself will be unknown and inaccessible and, thus, the algorithm for checking δ-dominance from Proposition 3 will not be directly applicable. Instead of knowing the true components, we thus usually will have to work with an i.i.d. sample $D_1, \ldots, D_s \sim \pi$ of data sets from $\mathcal{D}$ in such cases. Accordingly, for defining an empirical version of the algorithm, i.e., an algorithm for checking δ-dominance in the observed sample, we set $\hat{\mathcal{D}}_s := \{D_1, \ldots, D_s\}$ and then consider the empirical law given by

$$\hat{\pi}(\mathcal{W}) := \frac{1}{s} \cdot |\{j : j \in \{1, \ldots, s\} \ \wedge \ D_j \in \mathcal{W}\}| \tag{9}$$

for all $\mathcal{W} \in 2^{\hat{\mathcal{D}}_s}$. We then can simply run Proposition 3 with $\mathcal{D}$ replaced by $\hat{\mathcal{D}}_s$ and π replaced by $\hat{\pi}$. Of course, the result of this empirical version of Proposition 3 is then subject to statistical uncertainty: even if the optimal value indicates δ-dominance within the observed sample of data sets, this might not generalize to the true space $\mathcal{D}$. Conversely, it might also happen that there is δ-dominance in the true space $\mathcal{D}$, however, this dominance cannot be detected in the observed sample.

In order to control the probability of an erroneous conclusion, an appropriate statistical test should be carried out. A statistical test for the similar setup of classical stochastic dominance between random variables with values in partially ordered sets is discussed in Schollmeyer et al. (2017) and based on the two-sample observation-randomization test to be found, e.g., in Pratt and Gibbons (2012, Chapter 6). We now demonstrate how such test can be transferred to our setting: As already emphasized, for the empirical version, classifier C_i dominates classifier C_j if and only if the optimal value opt_{ij} of the linear program (6) is greater than or equal to zero. Therefore, it is natural to calculate opt_{ij} in the observed sample and reject the null hypothesis

$$H_0 : C_j \succsim_\delta C_i \tag{10}$$

if this value is larger than a critical value c. Since the distribution of the statistic opt_{ij} under the null hypothesis is difficult to handle, we use a permutation test that randomly swaps the labels of the classifiers for every data point, i.e., for every data set in the sample. In this way we can analyze the distribution of the test statistic under the most extreme hypothesis in H_0, i.e., the situation where the quality vectors of C_i and C_j are identically distributed. Then one can reject the null hypothesis if the value of opt_{ij} for the actually observed data sets is larger than the $(1-\alpha)$-quantile of the values obtained under the resampling scheme.

Importantly, note that we are actually interested in a statistical test that is only sensitive for deviations from H_0 in the direction of δ-dominance in the sense of $C_i \succ_\delta C_j$. Therefore it would be desirable to take as the null hypothesis the negation of $C_i \succ_\delta C_j$, however, under this null-hypothesis, the analysis of the distribution of opt_{ij} seems to be difficult. Additionally, at least for $R_2 = \emptyset$ and $\delta = 0$, which corresponds to classical first-order stochastic dominance, a consistent test seems to be unreachable,[4] cf., Whang (2019, p.106) and also Garcia-Gomez et al. (2019).

The concrete procedure for evaluating the distribution of opt_{ij} has the following five steps:

Step 1: Use the sampled data sets to produce two separate samples $(x_1, \ldots, x_s)$ and $(y_1, \ldots, y_s)$ from $\mathcal{Q}$, one for each classifier under consideration. Thereby, we used the notations $x_l := \phi(C_i, D_l)$ and $y_l := \phi(C_j, D_l)$ for all $l = 1, \ldots, s$.

Step 2: Take the pooled sample $z = (x_1, \ldots, x_s, y_1, \ldots, y_s)$.

Step 3: Take all index sets $I \subseteq \{1, \ldots, 2s\}$ of size s and compute the optimal outcome opt^I_{ij} of the linear program (6) that would be obtained if C_i would have produced the quality vectors $(z_i)_{i \in I}$ and if C_j would have produced the quality vectors $(z_i)_{i \in \{1,\ldots,2s\} \setminus I}$.

Step 4: Sort all opt^I_{ij} in increasing order.

Step 5: Reject H_0 if opt_{ij} is greater than the $\lceil (1-\alpha) \cdot \binom{2s}{s} \rceil$-th value of the increasingly ordered values opt^I_{ij}, where α is the envisaged confidence level.

If $\binom{2s}{s}$ is too large, instead of computing opt^I_{ij} for all index sets I, one can alternatively compute opt^I_{ij} only for a large enough number N of randomly drawn index sets I.

Remark 3 *Three important points should be added:*

i) If the statistical test described in (10) is to be used to test the entire order structure on the set $\mathcal{C}$ instead of just a single pairwise comparison, it must be performed for $n \cdot (n-1)$ pairs. Then, it must be corrected for multiple testing to guarantee the specified global significance level.

ii) As already discussed at the end of Section 3.2, the parameter δ acts as a regularizer. This becomes even clearer in the context of statistical testing: For $\delta = 0$, the maximum value of the linear program (6) is exactly zero in the dominance case and strictly less

4. A promising line of future research could be to reflect on whether the introduction of $\delta \neq 0$ and/or $R_2 \neq \emptyset$ indeed leads to a consistent test for the now 'regularized' null- and alternative hypotheses.

than zero otherwise. Thus, in this case, it is impossible to compare the extent of dominance for two different dominance situations using our test statistic. If, on the other hand, we choose a value $\delta > 0$ for the test, the maximum value of the linear program (6) in the dominance case can also assume values strictly greater than zero. In this way, different dominance situations can also be compared with each other in this case: The further the maximum value is above zero, the greater the extent of dominance. This potentially increases the power of the test, since situations can also be distinguished in which dominance is present in the resample, but a greater degree of dominance is present in the sample.

iii) As the parameter δ changes, of course, the hypotheses of the statistical test (10) also change. Thus, strictly speaking, a different statistical test is performed for each δ. However, it is important to note here that the extreme case of distributional equality of the two competing classifiers for any δ belongs to the null hypothesis. Thus, the test from (10) for arbitrary choices of δ is suitable for detecting systematic differences in the distributions of the classifiers. Furthermore, it can be argued that a very small value of $\delta > 0$, changes the order little to nothing compared to $\succsim_0$ (and thus the hypotheses of the associated statistical tests). However, it is shown (not least in the simulation study from Section 5) that even such a very small value of $\delta > 0$ can have a clearly visible positive effect on the power of the associated test.

4.3 Alternative Statistical Tests

As alternative methods to GSD-δ, we now briefly review the rank-based test for comparing competing classifiers as proposed in Demšar (2006) and suggest ways to extend it to more than one quality criterion. In the original test, classifiers are ranked on each data set based on their quality, typically estimated via cross-validation. Note that ranking is only straight-forward for one criterion, with no obvious way to extend it to multiple dimensions.

Ranks are then averaged over all data sets. The rationale is that data sets vary in difficulty and therefore ranking is a way to bring the different data sets on the same scale and avoid normality assumptions. The Friedman test can be applied to test for overall differences in mean-ranks. If significant differences are detected, post-hoc tests, such as the Nemenyi test can be used to determine which pairs of classifiers are significantly different. In the second step some form of correction for multiple testing is required to hold the overall α-level.

As the test by Demšar (2006) only accounts for differences between classifiers with respect to *one single* quality criterion, it must first be adapted to the setting with multiple quality criteria in order to allow a meaningful comparison with our approach. To reach a decision for multiple quality criteria we propose the following two intuitive heuristics:

all-test: Classifier C_i is considered better than C_j if it performs significantly better on each quality criterion.

one-test: Classifier C_i is considered better than classifier C_j if C_i performs significantly better in at least one dimension and if, additionally, in any other dimension classifier C_j does not perform significantly better than classifier C_i.

It should be noted that when using the all-test or one-test heuristics the α-level of the one-

dimensional tests is no longer preserved. In case of the one-test the true type 1 error will exceed α and is therefore no longer a valid α-level test and instead becomes over sensitive (cf. Section 5 and Appendix A3 for an example of this effect on simulated data). The all-test will often lead to a type 1 error much lower than α, as all dimensions need to be significant and can be therefore considered as a very conservative α-level test. We do not adjust the α level, but instead note, that the two approaches are perhaps the most intuitive ways to combine tests on several quality criteria.

5. A Simulation Study

In this section, we perform a simulation study to compare the proposed statistical test for δ-dominance from (10) for two different choices of δ with the adapted rank-based heuristics (the all-test and the one-test) as discussed in Section 4.3. In addition, we compare how well the relation $\succsim_\delta$ can reproduce the order structure in the groundtruth of the simulation when evaluated only in the sample, and again contrast the rank-based sample orders comparatively. Further, we shed some light on the role of δ for the performance of our test.

5.1 Design of the Simulation Study

The simulation study is designed as follows: Seven simulated classifiers $C_1, \dots, C_7$ with expected performance $\theta_i \in [0,1]^2$ on two dimensions (i.e. on two quality criteria) are compared. Both quality criteria are assumed to be interpretable on a metric scale in the sense that both contribute to the construction of the relation R_2 from Equation (2). The structure among the classifiers $C_1, \dots, C_7$ in the groundtruth is then induced by the recursive graph shown in Figure 1, where a *separation parameter* η controls the expected difference in performance. A specific example for such a recursive graph is given in Figure 2.

The performances x_{ij} of classifier C_i on data set D_j, where $j = 1, \cdots, s$, are i.i.d. drawn from a normal distribution, i.e., $x_{ij} \sim \mathcal{N}_2(\theta_i, \Sigma_\epsilon)$, where $\Sigma_\epsilon = \sigma_\epsilon I$ and σ_ϵ is a noise term, which, together with η controls the difficulty in unravelling the underlying dominance structure. Note that the θ_i's do not depend on the data set and therefore the difficulty is set to be the same for each data set. Due to Proposition 2, the setting of varying difficulties is also implicitly covered for many relevant situations: In particular, a random intercept or a random scaling of the cardinal dimensions will not influence the notion of dominance, making the simulation setup quite general (see also the discussion at the end of Section 3.3).

In the groundtruth just described, independent of the choice of $\eta > 0$, there are ten pairs of classifiers between which there is component-wise dominance in expectation, whereas all other pairwise comparisons are set to be not dominated, meaning that each classifier is preferable on one dimension. It is important to note that under the assumption of independence and constant variances, the component-wise dominance in expectation in the groundtruth also implies each of the three types of dominance discussed earlier, namely δ-dominance as well as the dominance w.r.t. the average rank of each classifier along the data sets under each quality criterion. Thus, in the groundtruth of the simulation there are also ten pairs of classifiers which are in relation with respect to the orders underlying the test for δ-dominance, the all-test and the one-test, respectively.

As the tests proposed by Demšar (2006) are only one-dimensional, we use the all-test and one-test heuristic described in Section 4.3 as our best effort to generalize the test to multiple dimensions for a meaningful comparison.

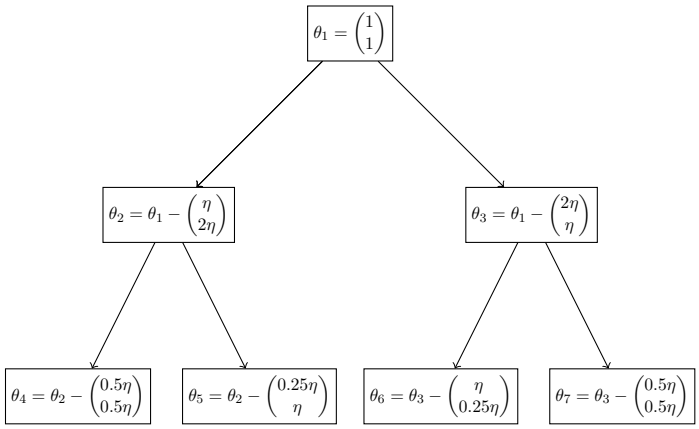

Figure 1: Simulation setting specifying the dominance between the simulated classifiers.

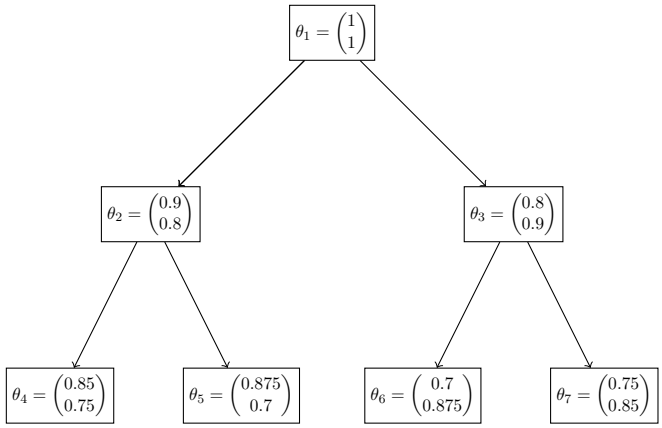

Figure 2: Example for $\Delta = 0.1$. Each entry in θ is the expected value on the corresponding quality dimension drawn from normal distribution with fixed variance and normalized to $[0, 1]$.

5.2 Results

In this simulation setup, we consider a total of twelve different simulation scenarios, namely all combinations of $\eta \in \{0.01, 0.05, 0.1\}$ (i.e. varying the separation parameter) and $s \in \{7, 10, 15, 18\}$ (i.e. varying the number of sampled data sets). Within each simulation scenario, we carry out a total of 25 simulation runs. In each of the simulation runs, the number of resamples drawn for the corresponding resample test is chosen to increase with the number d of sampled data sets. Concretely, we are interested in the following questions:

How well does the order structure found in sample reproduce the groundtruth?

To answer this question, we proceed as follows: In each simulation run, we compute the order structure of the three orders underlying the tests in the sample and compare it to the true dominance structure in the ground truth. Specifically, we compute δ-dominance in the sample by the empirical variant of the linear program from Proposition 3 discussed earlier. For receiving the (coinciding) orders underlying the all-test and the one-test, we compute the average rank of each classifier along the data sets under each quality criterion, and then define a classifier to dominate another one, whenever its average rank is superior in both quality dimensions. To measure the similarity of the orders in groundtruth and in sample, we use the F-score, i.e., a trade-off measure between non-detected dominances (*false negatives, FN*) and falsely detected dominances (*false positives, FP*).[5]

The results of the analyses carried out in the samples for the twelve simulation scenarios are visualized in Figure 3. The results show a balanced picture with regard to the different methods: All methods reproduce the order in the ground truth about equally well. As expected, the F-score of the methods tends to improve with increasing separation parameter η and increasing sample size d, with some random fluctuations which are especially visible for the lowly separated simulation scenarios.

How well do the statistical significance tests reproduce the groundtruth?

To answer this question, under each simulation run in each scenario, we perform four different statistical tests: the δ-dominance test for $\delta = 0$, the δ dominance test for $\delta = 10^{-5}$, the all-test, and the one-test.[6] It is important to note that the one-test was included only for the sake of completeness: As already described in Section 4.3, this test in general will not adhere to (and often drastically exceed) the specified α-level. Thus, the comparison with significance tests at this level is of course extremely problematic.

To measure the similarity of the order in the ground truth and the order given by the significant edges, we again use the F-score. The used global confidence level is $\alpha = 0.05$. The method used for correcting for multiple testing is the (very conservative) Bonferroni-correction for all four tests. The results of the analyses carried out at the test level for the twelve simulation scenarios are visualized in Figure 4.

Here, some remarkable observations suggest themselves: Under each simulation scenario, both tests for δ-dominance reveal the order structure at a global significance level of $\alpha = 0.05$ at least as well as the all-test heuristic. This dominance becomes increasingly clear as the separation parameter δ and the number d of simulated data sets increase. As expected, the F-score of all methods tends to improve with increasing separation parameter η and increasing sample size d. Interestingly, both tests for δ-dominance outperform also the one-test heuristic for separation parameters $\eta \geq 0.05$ and at least 15 data sets, although this

5. The F-Score is defined as $F = \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$, where TP denotes the number of correctly detected dominances. We chose this measure due to its popularity, however, an analysis under other measures (such as the Jaccard index, see Jaccard (1912)) essentially yields the same results.
6. The idea behind choosing an extremely small value such as 10^{-5} for δ in the second test, is to change the hypotheses of the test as little as possible compared to the test for $\delta = 0$, but still benefit from the gain in power that a strictly positive δ brings (cf. Remark 3 for further details).

heuristic exceeds the given first-type error probability of 0.05 (cf. A3).[7]

Furthermore, the comparison of the two tests for δ-dominance for $\delta = 0$ and $\delta > 0$ confirms the gain in power for the latter case already theoretically indicated in Remark 3 iii): The F-Score of the dominance test for $\delta > 0$ exceeds the one of the dominance test for $\delta = 0$ for every simulation scenario. Especially remarkable is the fact that this effect already occurs for a very small value of $\delta = 10^{-5}$. Since for such a small δ the order $\succsim_\delta$ is presumably changed only very marginally compared to $\succsim_0$, this suggests once more that the parameter δ, in addition to its decision-theoretic interpretation, also has a pure regularization component and helps to make the hypotheses more separable.

Summary of the results: We have shown that the proposed statistical test for δ-dominance reveals the ordering structure in the groundtruth more adequately than the all-test heuristic in each of the considered simulation scenarios. Further, we demonstrated that in the scenarios with at least medium separation ($\eta \geq 0.05$) and enough data sets available ($d \geq 15$), the tests for δ-dominance also outperform the one-test heuristic, even if this heuristic does not guarantee the global α-level. Finally, it turned out that the test for δ-dominance with $\delta > 0$ reproduces the order structure in all simulation scenarios at least as well as the test with $\delta = 0$, even for very small choices of the parameter δ.

6. Application

We now showcase on standard benchmark data sets how the relation $\succsim_\delta$ and the resulting GSD-δ method can be used to rank classifiers based on their performance on multiple data sets with respect to multiple quality criteria. In addition, we use the statistical test proposed in Section 4.2 to investigate which of the orderable pairs of classifiers found in the sample may also be assumed to be statistically significant. As in Section 5, we again compare our results with those obtained under the adapted rank-based tests from Section 4.3. Finally, we examine how our results change when the analyses are based on classical stochastic dominance instead of the generalized stochastic dominance order $\succsim_\delta$.

6.1 Experimental Setup

For comparison we use 16 binary classification benchmark data sets. All data sets are taken from the UCI machine learning repository (Dua and Graff, 2017). The data sets strongly vary in size, dimensionality and class imbalance.

For the classifier comparison, we consider the three well established metrics *accuracy*, *area under the curve* and *Brier score*. On each data set 10-fold cross-validation is performed and results are averaged for each metric and classifier separately. Importantly, note that in following analyses all three of these quality criteria are considered to be of metric scale and,

7. We note that much better trade-offs (reflected in F-score) can be found for GSD and GSD-δ if we do not enforce an overall but instead individual α-level of 0.05 (cf. A3 for the evaluation without Bonferroni correction, where we can see that GSD uniformly outperforms the one-test). Presumably, this could also be achieved using a more efficient multiple-testing correction strategy.

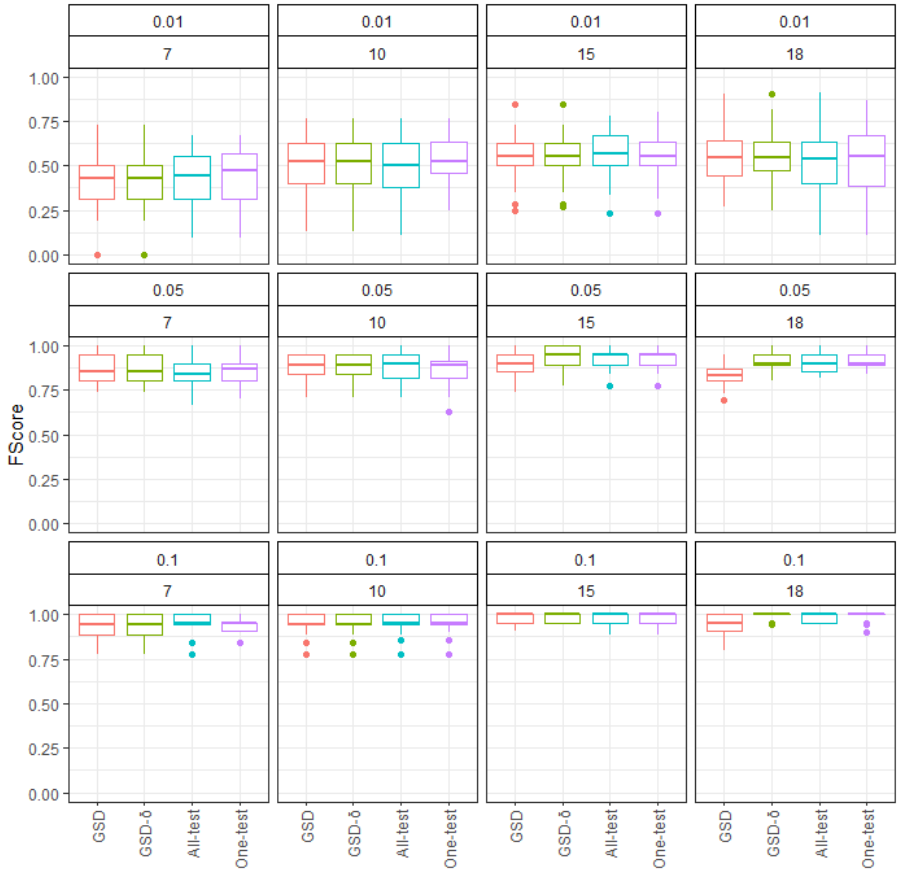

Figure 3: The figure shows the empirical distribution of the F-scores (larger is better) *in the samples* (without statistical test) of the different methods along the 25 simulation runs separately for the twelve different scenarios. The F-score is computed by counting the number of TPs, FPs, and FNs in the respective *sample* order compared with groundtruth and then evaluating the formula from Footnote 5.

accordingly, all equally contribute to the construction of the relation R_2 as most generally defined in Equation (2).[8] We compare two groups of algorithms:

- For decision tree based classifiers we included *classification and regression trees* (CART) (Breiman et al., 1983), *random forests* (RF) (Breiman, 2001), *gradient boosted trees* (GBM) (Friedman, 2002) and *boosted decision stumps* (BDS) (trees with depth 2).

- As examples of more traditional models we included *generalized linear models* (GLM), *lasso regression* (LASSO) (Tibshirani, 1996), *elastic net* (EN) (Zou and Hastie, 2005) and *ridge regression* (RIDGE), implemented in the `glmnet` R-package.

8. An exception is of course given by Section 6.3, where a comparison with classical stochastic dominance is considered, and, thus, all three quality criteria are considered to be purely ordinal for this case.

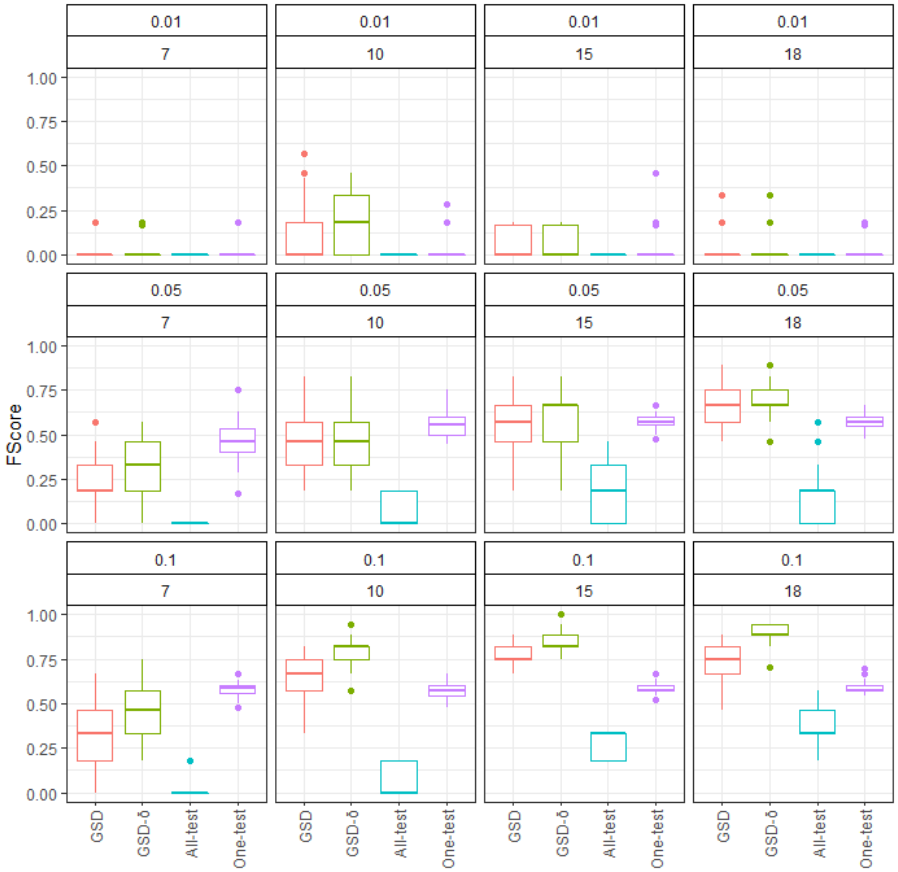

Figure 4: The figure shows the empirical distribution of the F-scores (larger is better) *of the significant orders* of the different methods along the 25 runs separately for the twelve scenarios. The F-score is computed by counting the number of TPs, FPs and FNs in the respective *significant order* compared with groundtruth. To account for multiple testing the Bonferroni-correction is used in all cases.

Generally, we expected the ensemble methods RF and GBM to dominate other methods, especially CART, whereas the ordering of the remaining methods is expected to be less clear. More details on data set selection, quality metrics and algorithm implementation can be found in Appendix A2.

6.2 Results

In the sample of data sets just described, evaluated and visualized in the three Hasse diagrams[9] in Figure 5, the following picture emerges: even though some of the classifiers

9. *Hasse diagrams* are graph representations of partial orders: Whenever two nodes can be connected by a path leading top down in the graph, then the upper node dominates the lower node with respect to the considered partial order. Nodes that cannot be connected by such a top down path are incomparable.

remain incomparable even for a maximum threshold of $\delta_{max} = 0.0077$, concretely BDS and RF as wells as EN and LASSO, a clear best classifier for this sample can be identified already for a minimum threshold of $\delta_{min} = 0$, namely GBM. In this case also two clear second-best, but incomparable to each other, classifiers can be seen, namely RF and BDS.

While an analysis under δ_{min} leaves the classifier pairs

(GLM,RIDGE), (GLM, EN), (GLM, LASSO), (RIDGE, LASSO), (RIDGE, EN)

incomparable to each other, raising the threshold to an intermediate level of $\delta = 0.004$ makes RIDGE dominant to both EN and LASSO. Raising the threshold even more to δ_{max}, the GLM classifier becomes dominant over EN, LASSO and RIDGE. Note that the order under threshold δ_{max} is the most structured relation we can hope for in this concrete sample: There exists no δ for which the relation $\succsim_\delta$ is a linear (or a preference) order. In particular, the classifiers BDS and RF as wells as EN and LASSO remain incomparable in this sample for no matter what threshold value is chosen.

Next, similar to what we did in the simulation study in Section 5, we perform the following three statistical tests for all pairwise comparisons: the test for δ-dominance from Equation (10) for δ_{min} and $\delta = 10^{-5}$, as well as the all-test as described in Section 4.3. The one-test is omitted since this test will in general drastically exceed the envisaged confidence level α, which was illustrated in the simulation study in Section 5. Also note that the choice of a threshold value of 10^{-5} can be exactly motivated as done in Section 5: Such small value will ensure to change the hypotheses of the test as little as possible compared to the test for δ_{min}, but still benefit from the gain in power that a strictly positive δ brings (compare in particular Footnote 6 and Remark 3). The results are as follows:

Dominance tests for δ_{min}: Here we find only one of the pairwise tests to be significant on a confidence level of $\alpha = 0.05$ (interpreted as a single test), namely the test of GBM over BDS (even enlarging the number of resamples from $N = 1000$ to $N = 10000$ does not change the situation). For all other pairwise comparison no significant distributional difference can be identified for any confidence level smaller or equal than 0.05. Note that under any correction procedure for multiple testing no significant ordering structure among the classifiers can be found using this test.

all-test: This test finds three pairwise comparison of classifiers to be significant on a confidence level of $\alpha = 0.05$: BDS over CART, GBM over CART and RF over CART. Again, for all other pairwise comparison, no significant distributional difference can be identified on this level. Interestingly, the pairs of identified significant pairwise comparisons of classifiers are disjoint for this test and the resample test for δ_{min}. Note that these three pairwise comparisons of classifiers still remain significant at a global α-level of 0.05 under any correction procedure for multiple testing (concretely, Bonferroni-correction was used here).

Dominance tests for $\delta = 10^{-5}$: The results for the resample tests for all pairwise comparisons of classifiers are given in Table 3. Concretely, for every pair (C_i, C_j) of classifiers, the table gives the share of resamples with test statistic strictly smaller than the test statistic in the original data, i.e., the value $\frac{1}{N}\sum_{I\in\mathcal{I}_N} \mathbb{1}_{\{opt^I_{ij} < opt_{ij}\}}$, with $\mathcal{I}_N$ the set of resampled index sets. A line symbolizes that this share was strictly below 0.95.

The table shows directly that – interpreted as one global test on the whole ordering struc-

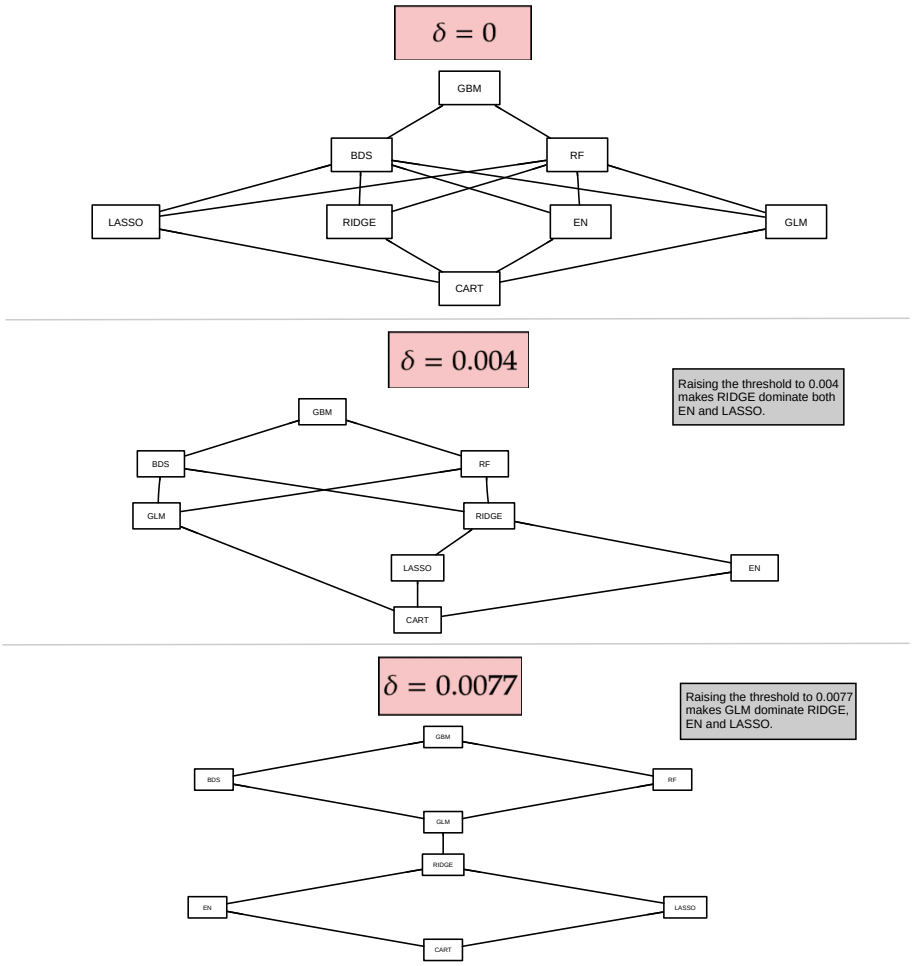

Figure 5: Hasse diagrams of $\succsim_\delta$ in the sample for the threshold values $\delta_{min} = 0$ (top), $\delta = 0.004$ (middle) and $\delta_{max} = 0.0077$ (bottom).

ture on the considered set $\mathcal{C}$ with global level $\alpha = 0.05$ – a whole series of significant pairwise comparisons emerge: First, all classifiers dominate the CART method significantly. Next, the GBM method dominates significantly all methods except GLM and RF. Furthermore, it can be seen that – now interpreted as single tests at individual level $\alpha = 0.05$ – BDS dominates the methods LASSO and RIDGE, as well as RF dominates the method EN. Here it is important to note that the latter three pairwise comparisons are no longer significant when corrected with any procedure for multiple testing.

Interpretation of the results: Table 3 shows that GBM, RF and GLM are not significantly different from each other and are also not significantly dominated by any other method. All other methods are, depending on the choice of α-level and correction method, dominated by some other model. CART is dominated by all other methods, including the linear model-based methods, indicating that for most data sets linear models work quite well, without

strong non-linearities or interaction effects. On the other hand, single CART models may overfit which is a well-known issue of decision trees.

It is also interesting to note that the regularized regression models do not perform significantly better than GLM. The reason might be that the regularization parameter is chosen via cross-validation, which might become unstable for smaller data sets. Also some data sets might not include irrelevant predictors, making plain GLM a better choice. This could also be the reason why GLM is not dominated by GBM, as the hyper-parameters in GBM are not chosen via cross-validation and therefore might be suboptimal. GBM as the more flexible model outperforms BDS for all reasonable significance levels, whereas GBM and RF are considered incomparable, which is in line with our expectations.

Generally speaking, using the proposed resample test for δ-dominance may lead to conservative results, however, still finds more structure that the all-test heuristic, which is the only competitor we could extract from the literature. Recall again that the one-test heuristic is omitted here, as it does not hold the α level as discussed in Section 4.3 and, therefore, cannot be meaningfully compared to statistical tests meeting this level. As the results found by GSD are very trustworthy, as shown in the simulation study, we are able to make more definite statements about the performance, such as that GBM outperforms LASSO and most often EN and RIDGE regression.

Finally, it should be mentioned that the advantages of a small value of $\delta > 0$ are also clearly shown in the concrete application: The test with $\delta = 10^{-5}$ finds remarkably more structure compared to the test with δ_{min}, although due to the very small value of δ the hypotheses of the underlying tests do hardly change.[10]

	BDS	CART	EN	GBM	GLM	LASSO	RF	RIDGE
BDS	–	1.000	0.976	–	–	0.967	–	0.951
CART	–	–	–	–	–	–	–	–
EN	–	0.998	–	–	–	–	–	–
GBM	0.998	1.000	0.998	–	–	0.999	–	0.997
GLM	–	1.000	–	–	–	–	–	–
LASSO	–	0.997	–	–	–	–	–	–
RF	–	1.000	0.953	–	–	–	–	–
RIDGE	–	0.999	–	–	–	–	–	–

Table 3: Results of the resample tests with $\delta = 10^{-5}$ and $N = 1000$ for all binary comparisons: For every pair (C_i, C_j) of classifiers, the table gives the share of resamples with test statistic strictly smaller than the test statistic in the original data. A line symbolizes that this share was strictly below 0.95.

10. As an indicator of how small this change of the underlying order actually is, one can name the fact that, restricted the observed sample, the orders $\succsim_0$ and $\succsim_{0.00001}$ do actually coincide.

6.3 Comparison with Classical Stochastic Dominance

To complete our study, we briefly compare the analysis results from Section 6.2 with the results that would be obtained under an analysis under classical (first-order) stochastic dominance. As already discussed in Remark 1 in Section 3, classical stochastic dominance arises as that special case of our dominance relation $\succsim_\delta$ where the relation R_2 from the preference system $\mathbb{C}$ defined in Equation (3) is empty, i.e., $R_2 = \emptyset$, and the threshold parameter δ is chosen to be 0. Importantly, note that the case $R_2 = \emptyset$ corresponds to the situation in which all the quality measures used are interpreted on a purely ordinal scale.

The detailed results of the analysis of the considered sample of data sets under classical stochastic dominance can be found in Appendix A1. In summary, we observe that an analysis under our dominance relation $\succsim_\delta$ (with non-empty R_2 and maybe even $\delta > 0$) allows for a much more structured comparison of the competing classifiers than is the case with an analysis under classical stochastic dominance. This is due to the fact that our dominance relation also allows us to fully exploit the information of the metrically interpretable quality criteria (here, all three), whereas stochastic dominance considers all dimensions of the quality vectors as purely ordinal. Thus, the application example suggests that ignoring available (partial) metric information indeed may lead to wasting relevant information about the underlying ordering structure.

7. Summary and Concluding Remarks

In this paper, we have developed a general framework for comparing classifiers with respect to different quality criteria on different data sets simultaneously. The basic idea of this comparison is based on a generalized version of classical multidimensional stochastic dominance, which also allows to adequately include the metric information of the quality measures used and can be regularized while attenuating the influence of the quality measures to the same extent. We have demonstrated how this dominance relation between classifiers can be detected by linear programming. Further, we showed how the optimal value of the linear program applied to a sample of data sets can be used as a statistic for statistically testing whether there is dominance between two competing classifiers. As the distribution of our test statistic is difficult to analyze, this test was performed by means of a permutation-based adapted two-sample observation-randomization test. Finally, we have illustrated the benefits of the proposed dominance concept over existing methods in a simulation study and applied it to real world data sets comparing eight classifiers with respect to three quality criteria. There are several promising directions for future research:

Incorporating classification difficulty: At this stage, the construction of our ordering $\succsim_\delta$ does not incorporate differences in the difficulty of classifying different data sets. However, as the heterogeneity of the considered space of data sets $\mathcal{D}$ increases, a co-consideration of these differences becomes more and more relevant. In principle, we believe that there would be two different ways to account for this: First, the quality vectors of the classifiers can be transformed before the dominance analysis with a loss function that depends on the data set. In this way, the challenges in comparing the difficulties would be outsourced to a pre-processing step. Of course, however, finding suitable loss functions is a research topic of its own. A second possibility is to incorporate the classification difficulty directly

into the modeling of the underlying preference system. For this purpose, the framework of *state-dependent* preference systems recently developed in Jansen and Augustin (2022) would presumably be directly transferable.

Reducing computational complexity for special cases: In its current form, the number of constraints of the linear program for checking $\succsim_\delta$-dominance given in Proposition 3 increases with complexity of at worst $\mathcal{O}(d^4)$, where d denotes the number of possible quality vectors (or the number of attained quality vectors in the observed sample, respectively). It certainly deserves further research how this worst-case complexity can be reduced if additional constraints on the considered preference system's metric relation R_2 are imposed.

Extension to multi-criteria decision making: The concepts presented here need by no means be limited to the comparison of classifiers. Thinking a bit more abstractly, any algorithms could be statistically compared with respect to different performance measures simultaneously in exactly the same way, including regression and even unsupervised learning settings, as long as meaningful quality criteria can be formulated. In principle, our framework could also be applied to general multi-criteria decision problems under uncertainty. An interesting aspect is that also multi-criteria decision problems can be analyzed with respect to purely ordinal as well as metrically scaled decision rules simultaneously.

Acknowledgments

Malte Nalenz and Georg Schollmeyer gratefully acknowledge the financial and general support of the LMU Mentoring program.

A. Appendix

A.1 Comparison with Classical Stochastic Dominance

If we analyze the same situation as described in Section 6.1, however, this time by means of classical stochastic dominance, we receive the results which are visualized in Figure 6. The first major difference from the analysis based on our dominance relation is the ordering under a threshold of $\delta = 0$: While a relatively structured picture emerged for the analysis under our dominance relation $\succsim_0$ already in this case (compare the top picture in Figure 5), the analysis based on stochastic dominance yields only two pairs of comparable classifiers, viz BDS over CART as well as GBM over CART. Based on stochastic dominance, no clear best and worst classifier can be identified within this concrete sample.

Considering the remaining five analyses under successively increasing threshold δ, two aspects in particular should be emphasized. First, as expected, the higher the threshold value, the more comparable the classifiers. In comparison to the analysis based on our dominance relation $\succsim_\delta$ under increasing δ, however, it is noticeable that even with threshold values that are higher by a factor of about ten, there still arise more weakly structured situations. Second, it is striking that even when analyzed with a relatively high threshold of $\delta = 0.06$, no clear best classifier can be identified: The methods RF and GBM remain incomparable here. Note that the order under threshold $\delta = 0.06$ is the best we can get: There exists no δ for which $\succsim_\delta$ possesses a superset of comparable pairs of classifiers.

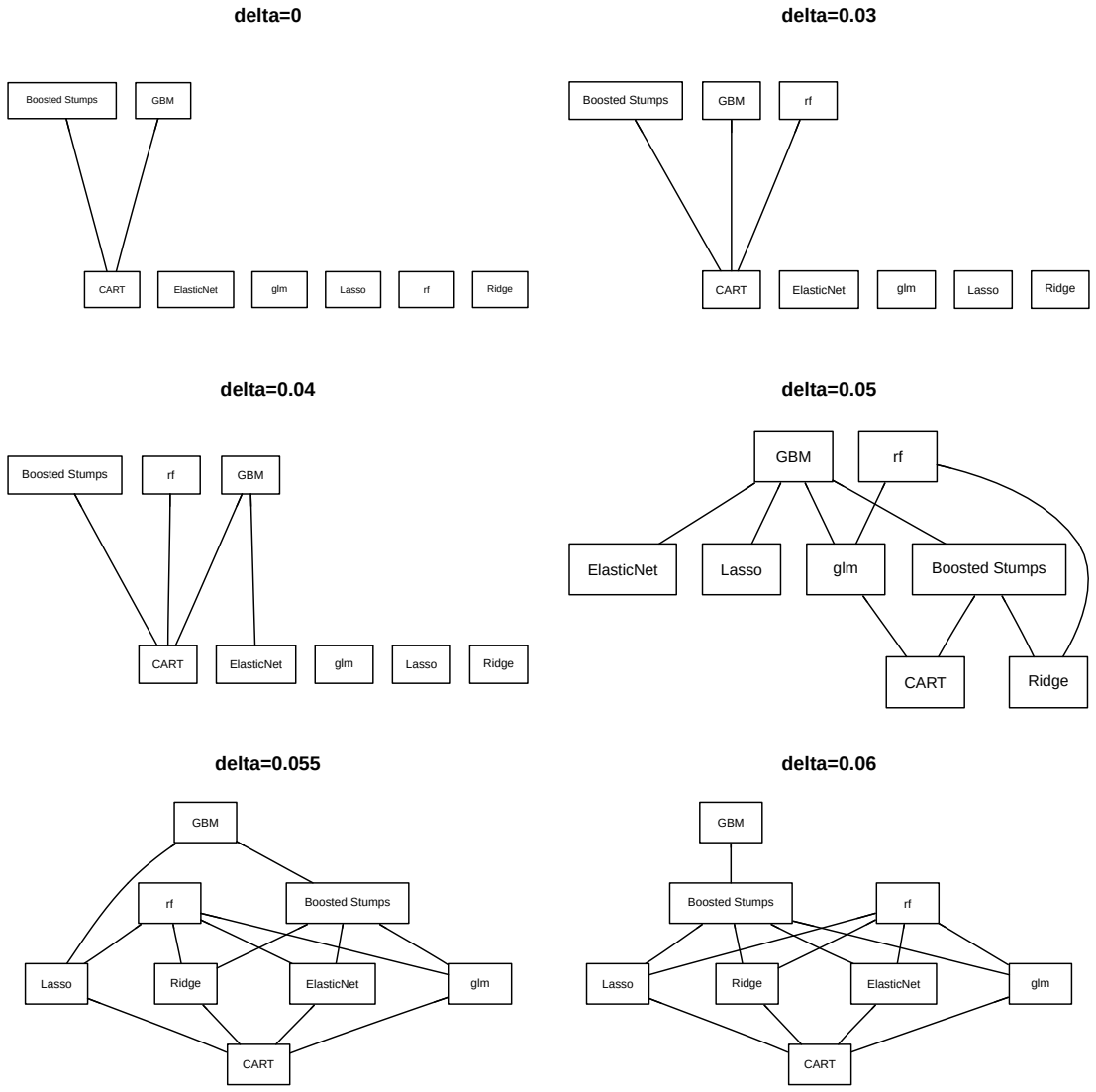

Figure 6: Hasse diagrams of $\succsim_\delta$ for six different threshold values δ with all quality criteria treated purely ordinal. The relation $\succsim_0$ coincides with stochastic dominance.

A.2 Data Set Selection and Implementation in Section 6.1

Data Set Selection. The data sets used for the experiments in Section 6.1 are taken from the UCI machine learning repository (Dua and Graff, 2017). The selection criteria are:

- Here, we only consider binary classification, but note that the method can be extended to any learning task where performance metrics of at least ordinal scale exist.
- We chose data sets with mostly numerical features or features with low cardinality.

- Only data sets with low number of missing values are considered.

Generally, we selected data sets, such that the need for pre-processing is minimal.

Algorithm Settings. We briefly describe the implementation of the compared methods:

- Ridge, Elastic Net and Lasso Regression are fit using the R-package(R Core Team, 2021) **glmnet** (Friedman et al., 2010). The optimal λ is determined via cross-validation. The mixing parameter in Elastic Net is set to 0.5.
- GBM and Gradient boosted decision stumps are fit using the **gbm** R-package. Gradient boosting uses 300 trees with a learning rate of 0.02 and a maximum depth of 3. The stumps use 500 trees and a learning rate of 0.05.
- Random Forest is fit using the **randomForest** R-package with default settings.
- For CART we use the **rpart** R-package with default settings.

Note that the results for all algorithms could likely be improved with parameter tuning, however, we used reasonable default values for the comparison. Our aim is solely to showcase our method, not to make any definite statements about the general performance of popular machine learning methods.

A.3 More Detailed Results

Figure 7: The figure shows the empirical distribution of the number of false positives.

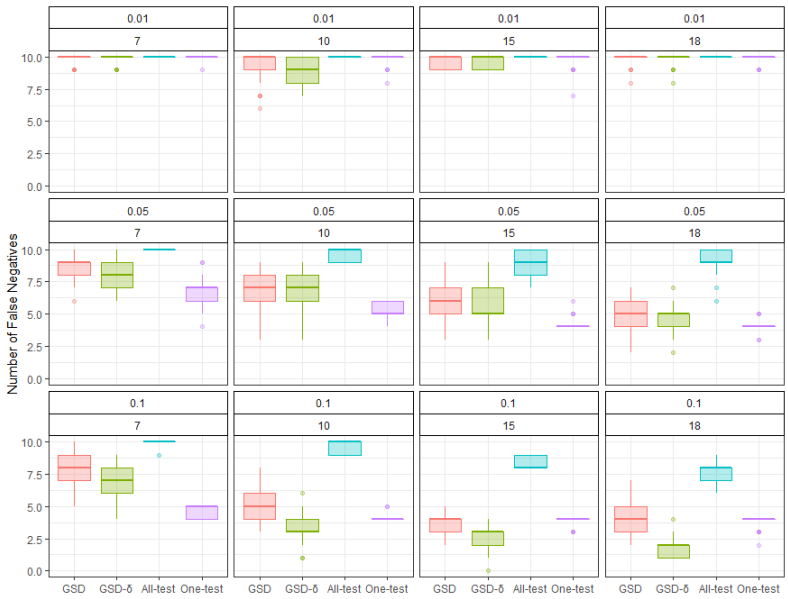

Figure 8: The figure shows the empirical distribution of the number of false negatives.

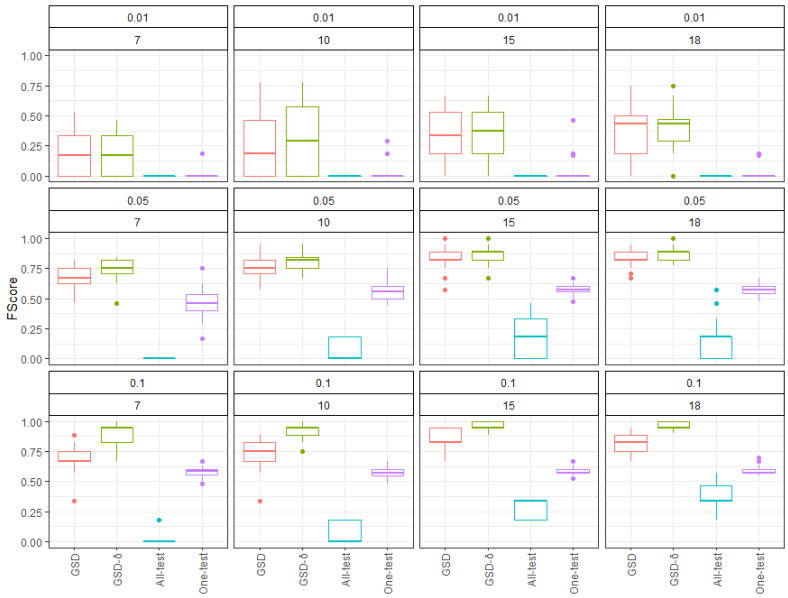

Figure 9: F-score for the different tests, without Bonferroni correction. A much better trade-off is achieved for GSD and GSD-δ.

Empirical Results. Tables 4, 5, and 6 show the raw performance values over the 16 analyzed data sets that the classifier comparison is based on.

data set	Boosted Stumps	CART	ElasticNet	GBM	glm	Lasso	rf	Ridge
australian	0.937	0.901	0.930	0.943	0.929	0.929	0.937	0.931
banknote	1.000	0.976	1.000	1.000	1.000	1.000	1.000	1.000
biodeg	0.926	0.847	0.917	0.926	0.922	0.917	0.937	0.916
blood_transfusion	0.738	0.722	0.751	0.736	0.752	0.752	0.670	0.751
diabetes	0.830	0.763	0.833	0.838	0.831	0.833	0.826	0.832
haberman	0.657	0.556	0.717	0.697	0.713	0.728	0.673	0.720
heart	0.906	0.823	0.904	0.902	0.912	0.906	0.904	0.910
ILPD	0.728	0.674	0.714	0.732	0.738	0.717	0.754	0.719
Ionosphere	0.972	0.915	0.910	0.973	0.866	0.904	0.980	0.913
liver	0.649	0.587	0.671	0.650	0.668	0.662	0.606	0.672
parkinsons	0.942	0.830	0.873	0.957	0.866	0.867	0.951	0.853
pop_failures	0.910	0.817	0.943	0.937	0.952	0.941	0.923	0.942
sonar	0.904	0.784	0.831	0.927	0.755	0.838	0.946	0.855
spambase	0.981	0.892	0.952	0.981	0.971	0.952	0.986	0.952
wbdc	0.993	0.960	0.994	0.992	0.962	0.994	0.991	0.993
wilt	0.990	0.956	0.970	0.989	0.977	0.970	0.989	0.963

Table 4: AUC for the different methods on the 16 data sets.

data set	Boosted Stumps	CART	ElasticNet	GBM	glm	Lasso	rf	Ridge
australian	0.865	0.845	0.859	0.864	0.859	0.859	0.867	0.859
banknote	0.995	0.965	0.975	0.989	0.987	0.975	0.993	0.977
biodeg	0.870	0.824	0.861	0.866	0.865	0.860	0.871	0.856
blood_transfusion	0.788	0.784	0.771	0.789	0.770	0.771	0.761	0.773
diabetes	0.751	0.734	0.769	0.763	0.773	0.771	0.768	0.772
haberman	0.735	0.719	0.735	0.729	0.742	0.735	0.725	0.732
heart	0.805	0.786	0.832	0.812	0.842	0.839	0.815	0.848
ILPD	0.700	0.671	0.710	0.712	0.724	0.708	0.705	0.712
Ionosphere	0.926	0.875	0.869	0.937	0.878	0.872	0.937	0.877
liver	0.583	0.569	0.629	0.606	0.615	0.626	0.554	0.629
parkinsons	0.902	0.841	0.876	0.922	0.860	0.876	0.907	0.861
pop_failures	0.926	0.928	0.915	0.943	0.957	0.915	0.924	0.915
sonar	0.818	0.755	0.756	0.842	0.736	0.737	0.842	0.794
spambase	0.944	0.893	0.884	0.939	0.927	0.884	0.954	0.884
wbdc	0.963	0.944	0.961	0.961	0.960	0.961	0.963	0.956
wilt	0.976	0.977	0.943	0.980	0.969	0.943	0.982	0.945

Table 5: Accuracy on the 16 data sets.

data set	Boosted Stumps	CART	ElasticNet	GBM	glm	Lasso	rf	Ridge
australian	0.095	0.119	0.106	0.091	0.101	0.106	0.096	0.106
banknote	0.011	0.032	0.034	0.011	0.009	0.034	0.006	0.041
biodeg	0.099	0.141	0.123	0.100	0.101	0.123	0.093	0.126
blood_transfusion	0.156	0.157	0.159	0.156	0.155	0.159	0.183	0.159
diabetes	0.163	0.195	0.163	0.158	0.158	0.163	0.161	0.163
haberman	0.191	0.203	0.184	0.187	0.183	0.183	0.194	0.184
heart	0.134	0.168	0.132	0.133	0.128	0.132	0.131	0.130
ILPD	0.182	0.224	0.187	0.179	0.176	0.187	0.172	0.187
Ionosphere	0.058	0.099	0.114	0.051	0.105	0.117	0.050	0.111
liver	0.248	0.278	0.234	0.242	0.228	0.235	0.253	0.233
parkinsons	0.070	0.126	0.112	0.061	0.111	0.114	0.072	0.117
pop_failures	0.052	0.058	0.061	0.043	0.033	0.061	0.055	0.061
sonar	0.120	0.204	0.181	0.110	0.264	0.176	0.124	0.169
spambase	0.046	0.094	0.112	0.048	0.059	0.112	0.039	0.111
wbdc	0.025	0.050	0.059	0.028	0.040	0.059	0.030	0.062
wilt	0.017	0.019	0.044	0.015	0.024	0.044	0.013	0.044

Table 6: Brier Score on the 16 data sets.

References

K. Arrow. A difficulty in the concept of social welfare. *Journal of Political Economy*, 58: 328–346, 1950.

A. Benavoli, G. Corani, and F. Mangili. Should we really use post-hoc tests based on mean-ranks? *The Journal of Machine Learning Research*, 17(1):152–161, 2016.

S. Brams and P. Fishburn. Voting procedures. In K. Arrow, A. Sen, and K. Suzumura, editors, *Handbook of Social Choice and Welfare, Vol. 1*, pages 173–236. North-Holland, 2002.

P. Brazdil, C. Soares, and J. Da Costa. Ranking learning algorithms: Using IBL and meta-learning on accuracy and time results. *Machine Learning*, 50(3):251–277, 2003.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and regression trees*. reprinted: 2017, Taylor & Francis, 1983.

G. Corani, A. Benavoli, J. Demšar, F. Mangili, and M. Zaffalon. Statistical comparison of classifiers through Bayesian hierarchical modelling. *Machine Learning*, 106(11):1817–1837, 2017.

O. Davidov and S. Peddada. Testing for the multivariate stochastic order among ordered experimental groups with application to dose–response studies. *Biometrics*, 69(4):982–990, 2013.

J. de Borda. *Memoire sur les elections au scrutin*. Historie de l'Academie Royale des Sciences, 1781.

N. de Condorcet. *Essai sur l'application de l'analyse a la probabilite des decisions rendues a la pluralite des voix*. Paris, 1785.

K. Deb. Multi-objective optimization. In E. Burke and G. Kendall, editors, *Search Methodologies*, pages 403–449. Springer, 2014.

J. Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30, 2006.

D. Dua and C. Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

M. Eugster, T. Hothorn, and F. Leisch. Domain-based benchmark experiments: Exploratory and inferential analysis. *Austrian Journal of Statistics*, 41(1):5–26, 2012.

M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1):3133–3181, 2014.

J. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38: 367–378, 2002.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22, 2010.

C. Garcia-Gomez, A. Perez, and M. Prieto-Alaiz. A review of stochastic dominance methods for poverty analysis. *Journal of Economic Surveys*, 33:178–191, 2019.

M. Graczyk, T. Lasota, Z. Telec, and B. Trawiński. Nonparametric statistical analysis of machine learning algorithms for regression problems. In R. Setchi, I. Jordanov, R. Howlett, and L. Jain, editors, *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 111–120. Springer, 2010.

H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4): 917–963, 2019.

P. Jaccard. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50, 1912.

C. Jansen and T. Augustin. Decision making with state-dependent preference systems. In D. Ciucci, I. Couso, D. Medina, J.and Slezak, D. Petturiti, B. Bouchon-Meunier, and R. Yager, editors, *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 729–742. Springer, 2022.

C. Jansen, G. Schollmeyer, and T. Augustin. Concepts for decision making under severe uncertainty with partial ordinal and partial cardinal preferences. *International Journal of Approximate Reasoning*, 98:112–131, 2018.

C. Jansen, H. Blocher, T. Augustin, and G. Schollmeyer. Information efficient learning of complexly structured preferences: Elicitation procedures and their application to decision making under uncertainty. *International Journal of Approximate Reasoning*, 144:69–91, 2022.

D. Krantz, R. Luce, P. Suppes, and A. Tversky. *Foundations of Measurement. Volume I: Additive and Polynomial Representations.* Academic Press, San Diego and London, 1971.

N. Lavesson and P. Davidsson. Evaluating learning algorithms and classifiers. *International Journal of Intelligent Information and Database Systems*, 1:37–52, 2007.

R. Luce. Semiorders and a theory of utility discrimination. *Econometrica*, 24:178–191, 1956.

R. Marler and J. Arora. The weighted sum method for multi-objective optimization: new insights. *Structural and Multidisciplinary Optimization*, 41(6):853–862, 2010.

O. Mersmann, M. Preuss, H. Trautmann, B. Bischl, and C. Weihs. Analyzing the BBOB results by means of benchmarking concepts. *Evolutionary Computation*, 23:161–185, 2015.

K. Mosler and M. Scarsini. Some theory of stochastic dominance. In K. Mosler and M. Scarsini, editors, *Stochastic Orders and Decision under Risk*, pages 203–212. Institute of Mathematical Statistics, Hayward, CA, 1991.

C. Müssel, L. Lausser, M. Maucher, and H. Kestler. Multi-objective parameter selection for classifiers. *Journal of Statistical Software*, 46:1–27, 2012.

J. Pratt and J. Gibbons. *Concepts of Nonparametric Theory*. Springer, 2012.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

G. Schollmeyer, C. Jansen, and T. Augustin. Detecting stochastic dominance for poset-valued random variables as an example of linear programming on closure systems, 2017. URL `https://epub.ub.uni-muenchen.de/40416/13/TR_209.pdf`. Technical Report 209, Department of Statistics, LMU Munich.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

G. Webb. Multiboosting: A technique for combining boosting and wagging. *Machine Learning*, 40:159–196, 2000.

Y-J. Whang. *Econometric analysis of stochastic dominance: Concepts, methods, tools, and applications*. Cambridge University Press, 2019.

B. Yu and K. Kumbier. Veridical data science. *Proceedings of the National Academy of Science*, 117(8):3920–3929, 2020.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Methodological)*, 67(2):301–320, 2005.