Depth Functions for Partial Orders with a Descriptive Analysis of Machine Learning Algorithms

Hannah Blocher HANNAH.BLOCHER@STAT.UNI-MUENCHEN.DE
Georg Schollmeyer GEORG.SCHOLLMEYER@STAT.UNI-MUENCHEN.DE
Christoph Jansen CHRISTOPH.JANSEN@STAT.UNI-MUENCHEN.DE
Malte Nalenz MALTE.NALENZ@STAT.UNI-MUENCHEN.DE
Department of Statistics, Ludwig–Maximilians–Universität München, Munich, Germany

Abstract

We propose a framework for descriptively analyzing sets of partial orders based on the concept of depth functions. Despite intensive studies of depth functions in linear and metric spaces, there is very little discussion on depth functions for non-standard data types such as partial orders. We introduce an adaptation of the well-known simplicial depth to the set of all partial orders, the union-free generic (ufg) depth. Moreover, we utilize our ufg depth for a comparison of machine learning algorithms based on multidimensional performance measures. Concretely, we analyze the distribution of different classifier performances over a sample of standard benchmark data sets. Our results promisingly demonstrate that our approach differs substantially from existing benchmarking approaches and, therefore, adds a new perspective to the vivid debate on the comparison of classifiers.

Keywords: partial orders, data depth, benchmarking, algorithm comparison, outlier detection, non-standard data

1. Introduction

Partial orders – and the systematic incomparabilities of objects encoded in them – occur naturally in a variety of problems in a wide range of scientific disciplines. Examples range from decision theory, where the agents under consideration might be unable to arrange the consequences of their actions into total orders (see, e.g., [38, 22]) or have partial cardinal preferences (see, e.g., [18, 20]), over social choice theory, where a fair aggregate order might only be possible by incorporating systematic incomparabilities (see, e.g., [31, 19]), to finance, where risky assets do not always have to be comparable (see, e.g., [24, 7]). Of course, many other relevant examples exist.

In the specific context of statistics and machine learning, the incompleteness of the considered orders often originates from the fact that the objects to be ordered are to be compared with respect to several criteria and/or on several instances *simultaneously*: only if there is unanimous dominance of one object over another, this order is included in the corresponding relation. Quite a number of research papers recently have been devoted to such comparison in the specific context of classification algorithms, either with respect to multiple quality metrics (e.g., [12, 21]) or across multiple data sets (e.g., [10, 3]) or with respect to genuinely multidimensional performance criteria like receiver operating characteristic (ROC) curves (e.g., [8]). Another source of partial incomparability of classifiers is the case of classifiers that make only imprecise predictions, like for example the naive credal classifier (cf., [46]) or credal sum-product networks (cf., [27]). In this case the imprecision in the predictions may take over to incomparabilities of the then possibly interval-valued performance measures[1]

Within the application field of machine learning and statistics, one further aspect is of special importance: Since the instances generally depend on chance, the same is true for the partial orders considered. Consequently, instead of a single partial order, random variables must then be analyzed that map into the set of *all possible* partial orders on the set of objects under consideration. For example, in the aforementioned comparison of classification algorithms, the concrete order obtained depends on the random instantiation of the data set on which they are applied. In this paper, we are interested in exactly this situation: we discuss ways to *descriptively* analyze samples of such partial order-valued (or short: *poset-valued*[2]) random variables.

Before starting, it is worth taking the time to distinguish, right at the beginning, those works that we believe are closest to ours. The main difference from the analysis in [21], which addresses a similar setting, is that in our case, in addition to the emphasis on the descriptive rather than the inferential aspects, the random orders do not (necessarily) arise retrospectively from the pairwise comparisons of the

[1] For example one could think of comparing classifiers with utility-discounted predictive accuracy, cf., [47, p. 1292] under the usage of a whole range $[\underline{a}, \overline{a}]$ for the coefficient of risk aversion.

[2] Note that in fact we speak here about random variables which have posets (on a common underlying ground space) as outcomes. This should not be confused with random variables which have values in a partially ordered set.

individual algorithms. Rather, they are conceived as abstract random objects. Further, the goal of our paper is also very different: while [21] is interested in exploiting the available information in the best possible way to give one *global* partial order over the classifiers under consideration, the present paper aims to analyze the *distribution* of the partial classifier orders over a given set of data sets. The two methods are therefore – despite similarity of the formal setting – very different and thus not directly comparable.

Additionally, we do not hold the view that there is an underlying true (random) total order together with a coarsening mechanism that generates the (random) partial order. Such views are termed *epistemic view* within the IP-community, cf., [9]. Applications of this view in the context of partial order data can be found for example in [23, 30]. Opposed to this view, within the nomenclature of [9], we see our approach more in the spirit of the *ontic view* that is usually applied to set-valued data and that states that such data are set-valued by nature and that there is no true but unobserved data point within the observed sets. Generally, this distinction between the ontic and the epistemic view is much discussed in the IP comunity and especially at the ISIPTA conferences.[3] However, since in our case the random objects are partial orders, the term *ontic* as used in [9] seems to fit not perfectly. Instead we understand poset-valued data as a special type of *non-standard data*.[4]

While the same view is taken in [5], the differences here are found more in the objective: while that paper focuses on theory for stochastic modeling of poset-valued random variables, the present paper can be viewed as a framework for analyzing data, e.g., sampled from one of these models.

Of course, a descriptive analysis of samples of partial orders which explicitly addresses the above interpretation requires a completely different – and so far to the best of our knowledge not existing – mathematical apparatus compared to an analysis of standard data. A suitable formal framework is by no means obvious here. Fortunately, it turns out that the concept of a *depth function*, which has so far mainly been applied to $\mathbb{R}^d$-valued random variables[5] (see, e.g., [39, 28]), can be promisingly adapted to poset-valued random variables. Generally speaking, (data) depth functions define a notion of centrality and outlyingness of observations with respect to the entire data cloud. Equipped with this adapted depth concept, some classical descriptive statistics can then be naturally adapted to this particular non-standard data type as well.

Our paper is organized as follows: In Section 2, we briefly discuss the required mathematical definitions and concepts. We give a formal definition of our depth function, the *ufg depth*, in Section 3 and discuss some of its properties in Section 4. The concrete theorems and proofs can be found in the appendix. While Section 5 prepares our application by providing the required background, Section 6 is devoted to applying our framework to a specific example, namely the analysis of the goodness of classification algorithms on different data sets. Section 7 concludes by elaborating on some promising perspectives for future research.

2. Preliminaries

Partial orders (posets) sort the elements of a set M, where we allow that two elements $y_1, y_2 \in M$ are incomparable. Formally stated: Let M be a fixed set. Then $p \subseteq M \times M$ defines a partial order (poset) on M if and only if p is *reflexive* (for each $y \in M$ holds $(y, y) \in p$), *antisymmetric* (if $(y_1, y_2), (y_2, y_1) \in P$ then $y_1 = y_2$ is true) and *transitive* (if $(y_1, y_2), (y_2, y_3) \in p$ then also $(y_1, y_3) \in p$). If p is also strongly connected (for all $y_1, y_2 \in M$ either $(y_1, y_2) \in p$ or $(y_2, y_1) \in p$), then p defines a *total/linear order*. For a fixed set M, various posets sort the set M. We are interested in all posets that can be obtained for the set M, where the cardinality $\#M$ is finite. We denote the set of all posets on M by $\mathcal{P}_M$ (or $\mathcal{P}$ for short). Sometimes it can be useful to consider only the *transitive reduction*, this means that for a poset p we delete all pairs (y_1, y_2) which can be obtained by a transitive composition of two other elements in p. Note that there exists a one-to-one correspondence between the transitive reduction of posets and the posets itself. We denote the transitive reduction of a poset p by $tr(p)$. This transitive reduction is often used to simplify the diagram used to represent the partial order. These diagrams are called *Hasse diagram*. They consist of edges and knots where the knots are the elements of M and the edges state which element lies below the other, e.g., see Figure 2. The reverse, where we add all pairs which follow from transitivity, is called *transitive hull*. We call $th(p)$ the transitive hull of a relation p. We refer to [15] for further readings on partial orders. From now on, let $\mathcal{P}$ be all posets for a fixed set M. We denote the elements of M by y.

The analysis concept for poset-valued observations presented here is based on a *closure operator* on $\mathcal{P}$, see Section 3. In general, a closure operator $\gamma_\Omega : 2^\Omega \to 2^\Omega$ on set Ω is an operator which is *extensive* (for $A \subseteq \Omega$, $A \subseteq \gamma_\Omega(A)$ holds), *increasing*, (for $A, B \subseteq \Omega$ with

[3] Concrete applications can be found e.g. in [33] for the ontic, and in [32] for the epistemic view. A case where the ontic and the epistemic view coincide is discussed in [34]. Beyond this, in the field of partial identification in the context of generalizing confidence intervals to confidence regions for the so-called identified set, the question about an ontic vs. an epistemic view is also implicitly asked (without reference to these terms) by asking if such a confidence set should cover the true parameter or the whole identified set with prespecified probability, cf., [40].

[4] Note that we do not want to generally rule out epistemic treatment, but this is not the focus of this paper. Such a treatment could use that every partial order can be described by the set of all its linear extensions. For further discussions, c.f., [5].

[5] Exceptions are, e.g., the definition of depth functions for functional data in [26] and on lattices in [34, 35].

$A \subseteq B$, $\gamma_\Omega(A) \subseteq \gamma_\Omega(B)$ is true) and *idempotent* (for $A \subseteq \Omega$, $\gamma_\Omega(A) = \gamma_\Omega(\gamma_\Omega(A))$ holds). The set $\gamma_\Omega(2^\Omega)$ is called the *closure system*. Note that every closure operator (and therefore the closure system) can be uniquely described by an implicational system. An *implicational system $\mathcal{I}$* on Ω is a subset of $2^\Omega \times 2^\Omega$. The implicational system corresponding to the closure operator γ_Ω is defined by all pairs $(A, B) \in 2^\Omega \times 2^\Omega$ satisfying $\gamma_\Omega(A) \supseteq \gamma_\Omega(B)$. For short, we denote this by $A \longrightarrow B$. For more details on closure operators and implicational systems, see [4].

We aim to define a centrality and outlyingness measure on the set of all posets $\mathcal{P}$ based on a fixed and finite set M. In general, functions that measure centrality of a point with respect to an entire data cloud or an underlying distribution are called *(data) depth functions*. Depth functions on $\mathbb{R}^d$ have been studied intensively by [39] and [29], and various notions of depth have been defined, such as Tukeys' depth, see [42], and simplicial depth, see [25]. The idea behind the ufg depth introduced here is an adaptation of the *simplicial depth* on $\mathbb{R}^d$ to posets, which uses the concept of a closure operator. The simplicial depth on $\mathbb{R}^d$ is based on the convex closure operator which is defined as follows:

$$\gamma_{\mathbb{R}^d}\colon \begin{array}{l} 2^{\mathbb{R}^d} \longrightarrow 2^{\mathbb{R}^d} \\ A \mapsto \left\{ x \in \mathbb{R}^d \,\middle|\, \begin{array}{l} x = \sum_{i=1}^k \lambda_i a_i \text{ with } a_i \in A, \\ \lambda_i \in [0,1], \sum_{i=1}^k \lambda_i = 1, k \in \mathbb{N} \end{array} \right\}. \end{array}$$

For the simplicial depth, we consider only input sets A with cardinality $d+1$ which form a $(d+1)$-simplex (when no duplicates occur). Then, the simplicial depth of a point $x \in \mathbb{R}^d$ is the probability that x lies in the codomain of the convex closure operator of $d+1$ points randomly drawn from the underlying (empirical) distribution. The set of all sets A with cardinality $d+1$-simplices is a proper subset of $2^{\mathbb{R}^d}$. By using Carathéodorys' Theorem, see [11], we obtain that any set B of $d+1$ unique points is the smallest set, for which there exists no family of proper subsets $(A_i)_{i\in\{1,\ldots,\ell\}}$ with $A_i \subsetneq B$ such that $\bigcup_{i\in\{1,\ldots,\ell\}} \gamma_{\mathbb{R}^d}(A_i) = \gamma_{\mathbb{R}^d}(B)$. Thus, these simplices still characterizes the corresponding closure system. For $\mathcal{M}$ being the set of all probability measures on $\mathbb{R}^d$ the simplicial depth is then given by

$$D\colon \begin{array}{l} \mathbb{R}^d \times \mathcal{M} \longrightarrow [0,1], \\ (x, \nu) \mapsto \nu(x \in \gamma_{\mathbb{R}^d}\{X_1, \ldots, X_{d+1}\}), \end{array}$$

where $X_1, \ldots, X_{d+1} \overset{iid}{\sim} \nu$. When we consider a sample $x_1, \ldots, x_n \in \mathbb{R}^d$; $n \in \mathbb{N}$, we use the empirical probability measure instead of a probability measure ν. Thus, for a sample $x_1, \ldots, x_n \in \mathbb{R}^d$ with empirical measure ν_n we obtain as empirical simplicial depth

$$D_n\colon \begin{array}{l} \mathbb{R}^d \longrightarrow [0,1], \\ x \mapsto \binom{n}{d+1} \sum_{1\le i_1<\ldots<i_{d+1}\le n} 1_{\gamma_{\mathbb{R}^d}\{x_{i_1},\ldots,x_{i_{d+1}}\}}(x). \end{array}$$

Hence, if $x_1, \ldots, x_n$ are affine independent, then the depth of a point x is the proportion of $(d+1)$-simplices given by $x_1, \ldots, x_n$ that contain x.

3. Union-Free Generic Depth on Posets

Now, we introduce the union-free generic (ufg) depth function for posets which is in the spirit of the simplicial depth function, see Section 2. To define the depth function, we start, similar to the simplicial depth, with defining a closure operator on $\mathcal{P}$. The definition of the closure operator uses formal concept analysis, see [15], and the formal context introduced in [5]. This gives us

$$\gamma\colon \begin{array}{l} 2^{\mathcal{P}} \longrightarrow 2^{\mathcal{P}} \\ P \mapsto \left\{ p \in \mathcal{P} \mid \bigcap_{\tilde{p}\in P} \tilde{p} \subseteq p \subseteq \bigcup_{\tilde{p}\in P} \tilde{p} \right\}. \end{array}$$

This closure operator maps a set of posets P onto the sets of posets where each poset is a superset of the intersection of P and a subset of the union. In other words, any p lying in the closure of P satisfies the following condition: First, every pair $(y_1, y_2) \in M \times M$ that lies in every poset in P is also contained in p, and second, for every pair (y_1, y_2) that lies in p, there exists at least one $\tilde{p} \in P$ such that $(y_1, y_2) \in \tilde{p}$. Note that while the intersection of posets defines a poset again, this does not hold for the union. Analogously to the definition of the simplicial depth, we now only consider a subset of $2^{\mathcal{P}}$ and define

$$\mathcal{S} = \{P \subseteq \mathcal{P} \mid \text{Condition } (C1) \text{ and } (C2) \text{ hold }\}$$

with Conditions (C1) and (C2) given by:

(C1) $P \subsetneq \gamma(P)$,

(C2) There does not exist a family $(A_i)_{i\in\{1,\ldots,\ell\}}$ such that for all $i \in \{1, \ldots, \ell\}$, $A_i \subsetneq P$ and $\bigcup_{i\in\{1,\ldots,\ell\}} \gamma(A_i) = \gamma(P)$.[6]

$\mathcal{S}$ is a proper subset of $2^{\mathcal{P}}$, see Theorem 2 for details, which reduces $2^{\mathcal{P}}$ by redundant elements in the following sense: First, all subsets $P \subseteq \mathcal{P}$ with $\gamma(P) = P$ are trivial and therefore not included. Second, if there exists a proper subset $\tilde{P} \subsetneq P$ with $\gamma(\tilde{P}) = \gamma(P)$, then P is also not in $\mathcal{S}$. This follows by setting $\ell = 1$ and $A_1 = \tilde{P}$, which defines a family contradicting Condition (C2). These two properties can be generalized to arbitrary closure systems, and referring to [2], we call a set fulfilling these properties *generic*. The final reduction is to delete also all sets P where P can be decomposed by a family of proper subsets $(A_i)_{i\in\{1,\ldots,\ell\}}$ of P. Further, in this case, the union of $(\gamma(A_i))_{i\in\{1,\ldots,\ell\}}$ equals $\gamma(P)$. Note that due to extensitivity, the assumption $\cup_{i\in\{1,\ldots,\ell\}}\gamma(A_i) \subseteq \gamma(P)$ is always true. We call sets

[6] In formal concept analysis this is sometimes called *proper*.

respecting this third part *union-free*. Thus $\mathcal{S}$ consists of elements which are *union-free and generic*.

Example 1 *As a concrete example, consider the set $\mathcal{S}$ based on all posets on $\{y_1, y_2, y_3\}$. Let p_1, p_2 and p_3 be posets given by the transitive hull of $\{(y_1, y_2)\}$, $\{(y_1, y_2), (y_1, y_3)\}$ and $\{(y_1, y_3), (y_2, y_3)\}$. One can show that the closure of the family $\{p_1, p_3\}$ gives the same closure as $\{p_1, p_2, p_3\}$. Thus, $\{p_1, p_2, p_3\}$ contradicts Condition (C2). For a single poset p we can immediately prove that the closure contains only itself. Therefore, any set consisting of only one poset does not satisfy Condition (C1). In contrast, $\{p_2, p_3\}$ satisfies both Condition (C1) and Condition (C2), since it implies the trivial poset $p_\Delta := \{(y, y) \mid y \in M\}$, consisting only of the reflexive part. Thus, $\{p_1, p_2\}$ is an element of $\mathcal{S}$.*

Now, we define the *union-free generic (ufg) depth* of a poset p to be the weighted probability that p lies in a randomly drawn element of $\mathcal{S}$. Let $\mathcal{M}$ be the set of probabilities on $\mathcal{P}$. The *union-free generic (ufg for short) depth on posets* is given by

$$D\colon \begin{array}{l} \mathcal{P}\times\mathcal{M} \to [0,1] \\ (p,\nu) \mapsto \begin{cases} 0, & \text{if for all } S\in\mathcal{S}\colon \prod_{\tilde p\in S}\nu(\tilde p)=0 \\ c\sum_{S\in\mathcal{S}\colon p\in\gamma(S)}\prod_{\tilde p\in S}\nu(\tilde p), & \text{otherwise,}\end{cases}\end{array}$$

with $c = \left(\sum_{S\in\mathcal{S}}\prod_{\tilde p\in S}\nu_n(\tilde p)\right)^{-1}$[7]. These two cases are needed because c is not defined in the first case. Note that if there exists an $S \in \mathcal{S}$ with $\prod_{\tilde p\in\mathcal{S}}\nu(\tilde p) \neq 0$, then $D \not\equiv 0$. The case that $D \equiv 0$ only occurs in two specific situations which result from the structure of the probability mass, see Property (P2) and Corollary 3 for details. In contrast to the simplicial depth where only sets of cardinality $d + 1$ are considered, the elements of $\mathcal{S}$ differ in their cardinality. Thus, different approaches on how to include the different cardinalities are possible, i.e., by weighting. Here, we used weights equal to one.

The empirical version of the ufg depth uses the empirical probability measure ν_n given by an iid sample of posets $\underline{p} = (p_1, \ldots, p_n)$, $n \in \mathbb{N}$ instead. We obtain as *empirical union-free generic (ufg) depth*

$$D_n\colon \begin{array}{l} \mathcal{P} \to [0,1] \\ p \mapsto \begin{cases} 0, & \text{if for all } S\in\mathcal{S}\colon \prod_{\tilde p\in S}\nu_n(\tilde p)=0 \\ c_n\sum_{S\in\mathcal{S}, p\in\gamma(S)}\prod_{\tilde p\in S}\nu_n(\tilde p), & \text{else,}\end{cases}\end{array}$$

with $c_n = \left(\sum_{S\in\mathcal{S}}\prod_{\tilde p\in S}\nu_n(\tilde p)\right)^{-1}$. The empirical ufg depth of a poset p is therefore the normalized weighted sum of drawn sets $S \in \mathcal{S}$ which imply p. Note that when restricting $\mathcal{S}$ to the set $\{S \cap \{p_1, \ldots, p_n\} \mid S \in \mathcal{S}\}$, this does not change the depth value. This holds since for other elements $S \in \mathcal{S}$, the empirical measure for at least one $p \in S$ is zero.

Example 2 *Returning to Example 1, suppose that we observe (p_1, p_2, p_3). Then for the trivial poset p_Δ, the empirical depth is $D_n(p_\Delta) = 1/2$. For the set p_4 given by the transitive hull of $\{(y_3, y_1)\}$, the value of the empirical depth is zero. For p_{total} given by the transitive hull of $\{(y_1, y_3), (y_3, y_2)\}$, the empirical depth value is again zero.*

4. Properties of the UFG Depth and $\mathcal{S}$

For a better understanding of the ufg depth, we now discuss some properties of D_n and $\mathcal{S}$. The properties of D_n describe the mutual influence between the (empirical) measure and the ufg depth while the properties of $\mathcal{S}$ can be used to improve the computation time.

4.1. Properties of the (Empirical) UFG Depth

The following statements are given for D_n. Those properties which focus on the empirical measure and not on the concrete sample values can be transferred to D. The first observation is that the ufg depth

(**P1**) *considers the orders as a whole, not just pairwise comparisons.*

More precisely, the ufg depth cannot be represented as a function of the sum-statistics

$$\left(w_{(a,b)} := \#\{i \in \{1, \ldots, n\} \mid (a, b) \in p_i\}\right)_{(a,b)\in M\times M}$$

of the pairwise comparisons,[8] see Theorem 7.

Remarkably, this concretization of this property formalizes precisely the analogy to the ontic notion of non-standard data mentioned at the beginning: Computing the depth of a partial order cannot be broken down via simple sum-statistics, but requires the partial order as a holistic entity. This is due to the fact that the involved set operations within the closure operator γ rely on the partial orders as a whole.

In Section 3, we defined the ufg depth in terms of two cases. If there exists at least one element $S \in \mathcal{S}$ such that every $p \in S$ has a positive empirical measure, then $D_n \not\equiv 0$. In Corollary 3 we specify this

(**P2**) *non-triviality property*.

[7] Note that Condition (C1) and (C2) can be applied to the convex closure operator on $\mathbb{R}^d$, see Section 2, and we obtain an adapted $\mathcal{S}_{convex}$. Then, $\mathcal{S}_{convex}$ together with $\mathcal{M}_{convex}$ the set of measures which are absolute continuous to the Lebesgue measure, leads to the simplicial depth.

[8] Note that many classical approaches rely only on the sum-statistics. For example within the Bradly-Terry-Luce model (cf., [6, p. 325]) or the Mallows Φ model (cf., [13, p. 360]), the likelihood function that is maximized depends only on the data through the sum-statistics.

We claim that $D_n \equiv 0$ occurs only when either the entire (empirical) probability mass lies on one poset or when the (empirical) probability mass is on two posets where the transitive reduction differ only in one pair.

The next observation relates to how the sampled posets affect the ufg depth value. For example, let us recall Example 1 and Example 2. From the structure of the sample, we can immediately see that p_Δ has a nonzero depth and that p_{total} must have a depth of zero. For this

(P3) *implications of the sample on D_n property,*

let $\underline{p} = (p_1, \ldots, p_n)$ be a sample from $\mathcal{P}$. Let $(y_1, y_2) \in M \times M$ such that for all $i \in \{1, \ldots, n\}$, $(y_1, y_2) \notin p_i$. Then for every $p \in \mathcal{P}$ with $(y_1, y_2) \in p$, we get $D_n(p) = 0$. This means that if a pair does not occur in any poset of the sample, then every poset which contains this pair needs to have zero empirical depth. Reverse, when looking at non-pairs, a similar statement is true. Let $p \in \mathcal{P}$ with $(y_1, y_2) \notin p$ but for all $i \in \{1, \ldots, n\}$, $(y_1, y_2) \in p_i$ holds. Then, $D_n(p) = 0$. This follows from Corollary 4. The influence of duplicates on the value of the empirical ufg depth D_n is immediately apparent by using the empirical measure ν_n. Thus, each element in $\mathcal{S}$ is weighted by the number of duplicates in the sample $\{p_1, \ldots, p_n\}$.

Conversely to Property (P3), in some cases, structure in the sample can be inferred by the ufg depth values. In Example 1 and Example 2, knowing only the values of the depth function gives us some insight into the observed posets. For example, we know that there must be at least one pair (y_i, y_j) that is an element of p_{total}, but which is not given by any observed poset. Moreover, the fact that p_Δ has nonzero depth implies that there exists no pair (y_i, y_j) that every observed poset has. We call this property

(P4) *implications of the outliers on the sample.*

More precisely, the depth value of the trivial poset, which consists only of the reflexive part, as well as the values of the total orders, can provide further information about the sample. Therefore, let p_Δ be the trivial poset, and p_{total} be a total order. By Corollary 4 we obtain that if $D_n(p_\Delta) = 0$, then there exists at least one pair (y_1, y_2) which is in every poset of the sample. The knowledge of p_{total} leads to an statement about the non-edges. So, if $D_n(p_{\text{total}}) = 0$ is true, then there exists at least one pair $(y_1, y_2) \in p_{\text{total}}$ which is in no poset of the sample.

The last properties have summarized how the structure of a sample is reflected in the ufg depth and vice versa. Finally, we have

(P5) *consistency of the empirical ufg depth D_n.*

This means that D_n converges uniformly to D almost surely under the assumption of observing i.i.d. samples, see Theorem 6.

4.2. Properties of $\mathcal{S}$

In this subsection, we introduce some properties of $\mathcal{S}$, which we use to improve the computation. The first one is

(P6) *a lower bound for all $S \in \mathcal{S}$,*

which is given by $\#S \geq 2$. This fact is already discussed in Example 1. For the upper bound we use a complexity measure of $\mathcal{S}$, the Vapnik-Chervonenkis dimension (VC dimension for short), see [44]. The VC dimension of a family of sets $\mathscr{C}$ is the largest cardinality of a set A, such that A can still be shattered into the power set of A by $\mathscr{C}$.[9] With this, we obtain

(P7) *an upper bound for all $S \in \mathcal{S}$*

is given by $\#S \leq vc$, with vc the VC dimension of the closure system $\gamma(2^{\mathcal{P}})$. The proof of the upper and lower bound can be found in Theorem 5. Note that in our case of posets, the VC dimension is small compared to the number of all posets.

We conclude with the observation that

(P8) *the elements of $\mathcal{S}$ are connected*

in the sense that for every $S \in \mathcal{S}$ with $\#S = m \geq 3$, there is an $\tilde{S} \in \mathcal{S}$ such that $\tilde{S} \subsetneq S$ and $\#\tilde{S} = m - 1$, see Theorem 2.

5. Comparing Machine Learning Algorithms

Before turning to our actual application, we first indicate, which possible contributions our methodology based on data depth in the context of poset-valued data is able to add to the general task of analyzing machine learning (ML) algorithms beyond pure benchmarking considerations. The basic task of performance comparison of algorithms is very common in machine learning (cf., [17] and the references therein). Our methodological contribution deviates from the typical benchmark setting with regard to at least two points:

(I) First, we compare algorithms not with respect to one unidimensional criterion like, e.g., balanced accuracy, but instead we look at a whole set of performance measures. We then judge one algorithm as at least as good as another one if it is not outperformed with respect to any of these performance measures. With this, for every data set, we get a partial order of algorithms and since we are not looking at only one, but a whole population/sample of data sets, we get a poset-valued random variable.

(II) Second, we are not interested in the question which algorithm is in some certain sense the best or competitive,

[9] To be more precise: The intersection between a set A and a family of sets $\mathscr{C}$ is defined by $A \cap C = \{A \cap C \mid C \in \mathscr{C}\}$. We say that a set A can be shattered (by $\mathscr{C}$) if $\#(A \cap \mathscr{C}) = 2^{\#A}$ holds. The VC dimension of $\mathscr{C}$ is now defined as $vc = \max\{\#A \mid (A \cap \mathscr{C}) = 2^{\#A}\}$.

etc. Instead, we are interested in the question how the relative performance of different algorithms is distributed over a population/sample of different data sets. Analyzing the distribution of performance relations is in our view a research question of its own statistical importance that may add further insights to analyses in the spirit of e.g., [21] which are of most importance when it comes to choosing between different machine learning algorithms.

These both deviations can have very different motivations: The analysis of a multidimensional criteria (of performance, here) is already motivated in the fact that in a general analysis, different performance measures are in the first place conceptually on an equal footing (at least, if one has no further concrete, e.g., decision-theoretic desiderata at hand). Therefore it appears natural to take more than one measure at the same time into account. Beyond this, there are far more possible motivations for dealing with multidimensional criteria of performance: For example for classification, if one accounts for the impact of distributional shifts within covariates, then one aspect to consider is that for different covariate distributions, the class balance of the class labels will vary, which can naturally be captured by either looking at different weightings of the true positives and the true negatives within the construction of a classical performance measure[10] or alternatively by taking into account different discrimination thresholds for the classifiers simultaneously, which would correspond to looking at a whole region of the receiver operating characteristic.

Also the motivation for the second point can be manifold: Generally, it seems somehow naive to search for one best algorithm per se. For example, the scope of application of an algorithm can vary very strongly and therefore, for different situations, different algorithms could be the best, or in certain situations different algorithms can be comparable in its performance, or, on the other hand, incomparable if one looks at different performance measures at the same time. Generally, it can be of high interest, how the conditions between different algorithms change over the distribution of different data sets or application scenarios. For example, if in one very narrowly described data situation the performances of different algorithms vary extremely from case to case, but not so much across algorithms, then, at some point it would become more or less hopeless to search for a best algorithm in the training phase, because one knows that in the prediction setting, the situation is too different to the training situation.

Another aspect is outlier detection: If one knows that in a large, maybe automatically generated benchmark suite there are data sets that have some bad data quality (for example if some covariates are meaningless because of some data formatting error etc.), then, it would be reasonable to try to exclude such outlying data sets from a benchmark analysis beforehand. Candidates of such outliers are then naturally data sets with a low depth value.

6. Application on Classifier Comparison

After this motivation, we now apply our ufg depth on poset-valued data: each poset arises from comparison of classifiers based on multiple performance measures on a data set.

6.1. Implementation

Let $(p_1, \dots, p_n)$ be a sample of posets. There are two difficulties in computing D_n. First, going through each subset of $\{p_1, \dots, p_n\}$ is very time-consuming, especially since the subsets that are an element of $\mathcal{S}$ can be very sparse in $2^{\{p_1, \dots, p_n\}}$. Second, it is difficult to test whether a subset is an element of $\mathcal{S}$ or if it is not an element.

The first part can be improved by using the lower and upper bound on the cardinality of $S \in \mathcal{S}$, see Section 4. Here we use a binary integer linear programming formulation described in [37, p.33f] to compute the VC dimension. Further, we use the connectedness of the elements $S \in \mathcal{S}$, see Property (P8) in Section 4. With this, we do not have to go through every subset that lies between the lower and upper bounds, but can stop the search earlier.

To check whether a subset $P \subseteq \{p_1 \dots, p_n\}$ is an element of $\mathcal{S}$, we begin with two observations: First, Condition (C2) implies that there must exist a poset p that does not lie in any closure operator output of any proper subset of P. (This follows from the extensivity of the closure operator γ.) Second, Condition (C2) implies Condition (C1) for $\ell \geq 2$, since only for $\ell = 1$ we cannot define a family $(A_i)_{i \in \{1, \dots, \ell\}}$ consisting of proper subsets of P such that for every $p \in P$ there exists an $i \in \{1, \dots, n\}$ with $p \in A_i$.

By Property (P6), we know that for any $S \in \mathcal{S}$, $\#S \geq 2$ is true. So we only need to check if Condition (C2) is true. Thus, we want to find a poset p which is given only by the entire set P. Suppose that p is such a poset. Then $\bigcap_{\tilde{p} \in P} \tilde{p} \subseteq p \subseteq \bigcup_{\tilde{p} \in P} \tilde{p}$ and for every $\tilde{p} \in P$ at least one of the following statements is true:

$(T1)_{\tilde{p}}$ There exists a pair $(y_1, y_2) \in \tilde{p}$ with $(y_1, y_2) \in p$. But for all other $\hat{p} \in P \setminus \tilde{p}$, $(y_1, y_2) \notin \hat{p}$ is true.

$(T2)_{\tilde{p}}$ There exists a pair $(y_1, y_2) \notin \tilde{p}$ with $(y_1, y_2) \notin p$. But for all other $\hat{p} \in P \setminus \tilde{p}$, $(y_1, y_2) \in \hat{p}$ is true.

Thus, none of these $\tilde{p}$'s can be deleted since then $P \setminus \tilde{p}$ does not contain p in the closure output of γ anymore. After this analysis of the candidate p. we are now interested in the construction of the candidate p. First, observe that

[10] Note that usual performance measures are more or less simple transformations of the vector of the true positives and the true negatives (and the class balance).

for every $\tilde{p} \in \mathcal{P}$ and $i \in \{1, 2\}$ one can collect all pairs $(y_1^{(Ti)_{\tilde{p}}}, y_2^{(Ti)_{\tilde{p}}})$ which can be used to ensure that $(Ti)_{\tilde{p}}$ holds. Set $p_M = \cap_{\tilde{p} \in P} \tilde{p}$. For every element $\tilde{p} \in P$ choose one pair $(y_1^{(Ti)_{\tilde{p}}}, y_2^{(Ti)_{\tilde{p}}})$. Now, add those pairs to p_M with $i = 1$. Compute $th(p_M)$ and for every $\tilde{p}$ for which we chose $(y_1^{(Ti)_{\tilde{p}}}, y_2^{(Ti)_{\tilde{p}}})$ with $i = 2$, check that $\tilde{p}$ is necessary to obtain p_M in the output of the closure operator. If this holds for any $\tilde{p}$ where a pair with $i = 2$ is chosen, and $th(p_M) \subseteq \cup_{\tilde{p} \in P} \tilde{p}$, then p_M is a poset that ensures Condition (C2). Thus, by going through all combinations of $(y_1^{(Ti)_{\tilde{p}}}, y_2^{(Ti)_{\tilde{p}}})$, we can check whether such a poset exists. It is sufficient to pick for each $\tilde{p}$ precisely one $(y_{\tilde{p}_1}, y_{\tilde{p}_2})$ since $th(\hat{p}) \subseteq th(\tilde{p})$ is true if $\hat{p} \subseteq \tilde{p}$.

All in all, we improved the computation compared to the naive approach by using the knowledge provided in Section 4. Now, we can specify a worst and best case by the bounds. By further including the improved testing of Condition (C1) and (C2) and the connectedness property, we could decrease the computation time, although we currently cannot calculate the exact amount in general as this depends on the complexity of the data set used. Note that even the upper bound is not fixed, but depends on the structure of the data set. In our application, see Section 6.2 and 6.3, we consider 80 posets. The naive approach is not computable in reasonable time since one would have to compute and test each subset of the 80 posets. The approach above then leads to a computation time of approximately four hours.

6.2. Data Set

To showcase the application of the ufg depth on machine learning algorithms we use openly available data from the OpenML benchmarking suite [43].

In our comparison we include the following supervised learning methods: *Random Forests* (RF, implemented in the R-package `ranger` [45]), *Decision Tree* (CART, implemented via the `rpart` library [41]), *Logistic regression* (LR), *L1-penalized logistic regression* (Lasso, implemented through the `glmnet` library [14]) and *k-nearest neighbors* (KNN, through the `kknn` library [16]). As stated in the OpenML experiment-documentation all methods are run with default settings of the corresponding libraries. Hence, our application analyzes the behavior of methods using default settings and does not necessarily extend to general statements about the performance of hyperparameter-tuned versions of the respective algorithms. The algorithms were chosen as a selection of widely used supervised learning methods that perform reasonably without much tuning, in contrast to methods such as neural networks or boosting, which require considerable tuning to perform well.

From the available data sets for which results for all above algorithms are available in the OpenML database, we limit our analysis to binary classification data sets with more than 450 and less than 10000 observations, leading us to a total of 80 data sets for comparison. The data sets come from a variety of domains and strongly vary in their class balance as well as their overall difficulty. Included in our multidimensional criteria comparison are the measures *area under the curve*, *F-score*, *predictive accuracy* and *Brier score*. These performance measures capture different aspects of performance, especially in the case of unbalanced data sets.

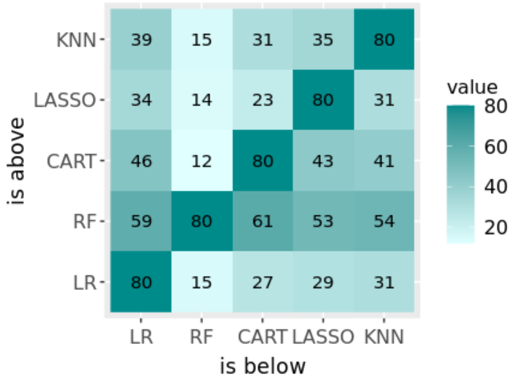

Figure 1: Heatmap representing the sum-statistics, see Footnote 9.

The corresponding posets then result from the multidimensional criteria comparison. It should be noted that rescaling the performance measures does not change the posets. This follows from the fact that the posets defined here do not depend on the absolute differences but on whether or not the multiple performance measures are better in all dimensions.

6.3. Analysis

The resulting poset-valued set consists of 80 posets, 58 of which are unique. Each of the 58 unique posets have a different depth value. The sum-statistics, see Footnote 4, which count for each pair the number of occurrences along the 80 posets, can be seen in Figure 1. It shows that RF is very often above all other methods. So if one only looks at the sum-statistics RF is clearly the strongest method. The other methods are more balanced with respect to each other. Note that due to reflexivity the diagonal is always 80.

The most central poset with the maximum depth value is a total order and can be seen in Figure 2. Its depth value is 0.34. The poset with the highest depth value also has the most duplicates, meaning it is the most common pattern. As described in Section 5, we are interested in the distribution of the observed posets. Nevertheless, we can consider the poset with the highest depth value as the poset whose structure is the most common one. Or, in other

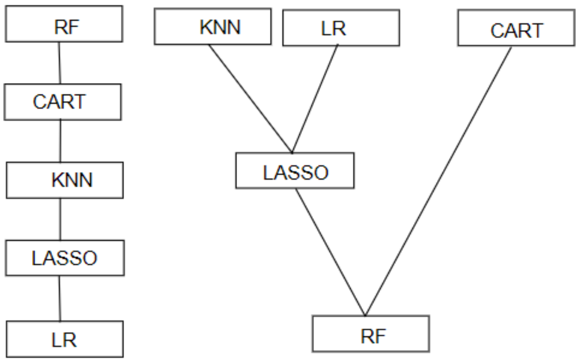

Figure 2: Observed poset with maximal (left) and minimal (right) ufg depth. On the left-hand side RF dominates every other algorithm and in contrast on the right-hand side RF is dominated by every other.

words, this poset is the one that is most supported based on all observations. Comparing this to Figure 1 or, e.g., the results in [21], we see many similarities, such as LR often has worse performance than the other algorithms, and RF dominates all other algorithms in many cases. In contrast to the sum-statistics which here give a representative poset, the strength of our method is that we not only obtain one single poset structure, but also a distribution over the set of posets. Note that in general the order given by the sum-statistics is not a poset, i.e. their might exist cycles.

Figure 3 describes which edges the posets with the $k \in \{1, \dots, 80\}$ highest depth values have in common. For example, one can observe that the dominance of RF over all classifiers based on all four performance measures holds for the 35 posets with the highest depth values. In particular, any other classifier dominance (like CART outperforms KNN according to all performance measures), does not hold for the 35 posets with the highest depth values. For example that CART outperforms KNN is only true for the 13 deepest posets. Note that the posets with the highest 46 depth values have nothing more in common. Conversely, it is of interest to see what non-edges the posets have in common. Since the poset with the highest depth value is the total order, this is immediately apparent in Figure 3. The posets with $k \in \{1, \dots, 80\}$ highest depth values have those non-edges in common, which are given by the inversely ordered poset of highest depth value intersecting with the inversely already deleted ones. For example, the posets with the nine highest depth values have in common that the RF is not dominated by CART, CART not by KNN and KNN not by Lasso, but they do not agree on LASSO being not dominated by LR since the poset with the 8th highest depth value does not agree on this.

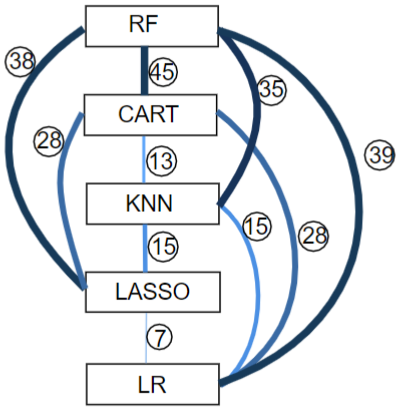

Figure 3: Represents what the observed posets with the k highest depth values have in common. Compare with Figure 2, where the poset with the highest depth value is plotted. Here each edge number kth indicates that the k deepest posets all contain this relation, but this is not true for the $k + 1$ deepest poset.

Unlike the posets with the highest depth values, the posets with low depth values do not have much in common. The posets with the tenth lowest depth values only agree on RF being dominated by another classifier. After that, no structure holds. All of these posets can be seen as outlier, or in other words, the corresponding data sets produce a performance structure on the classifiers which differ from the structure given by other data sets. The poset with the smallest depth value, which is 0.05, can be seen on the right side of Figure 2.

Finally, we want to give a notion of dispersion of the depth function. Therefore, we compute the depth function for every poset $p \in \mathcal{P}$ and compute the proportion of posets which lie in $\alpha \in [0, 1]$ deepest observed depth values. For $\alpha = 0.25, 0.5$ and 0.75 we get $0.02, 0.10$ and 0.26. Thus, the empirical ufg depth seems to be clustered on small parts of the set $\mathcal{P}$.

To summarize, the concept of depth functions allowed us to get valuable insight in the typical order of the analyzed classifiers. Further, it detects data sets where the structure of the classifiers, given by the performance measures, seems to have unusual structure.

7. Conclusion

In this paper, we have shown how samples of poset-valued random variables can be analyzed (descriptively) by utilizing a generalized concept of data depth. For this purpose, we first introduced an adaptation of the simplicial depth, the so-called ufg depth, and studied some of its properties.

Finally, we illustrated our framework with the example of comparing classifiers using multiple performance measures simultaneously. There are several promising avenues for future research, that include (but are not limited to):

Other ML Problems and Criteria: Here, we focused on the comparison of classifiers by a set of unidimensional performance criteria. For example, the performance of different optimization algorithms could be also of interest. Further, the analysis of classifiers with respect to other criteria could be an interesting modification. For example one could use ROC curves or criteria that do also take the fact into account that classical performance measures are only estimates of the true out of sample performance. Within our order-based approach this would be easily incooperateable.

Discussion on computation time: In Section 6.1 we briefly discussed the computation time and the difficulty of predicting it. For a deeper understanding further analyses, e.g. in form of a simulation study, would be helpful.

Inference: A first step towards inference for poset-valued random variables is already made by the consistency property in Section 4. Natural next tie-in points are provided by regression and statistical testing. Together with the results for modeling in [5], a complete statistical analysis framework for poset-valued random variables would then be achieved.

Other types of non-standard data: Our analysis framework is by no means limited to poset-valued random variables. Since the ufg depth is based on a closure operator, all non-standard data types for which a meaningful closure operator exists can be analyzed with it. As seen in [5] such closure operators are easily obtained by formal concept analysis, thus, there exists a natural generalization of the ufg depth for non-standard data.

Appendix A. Proofs of Section 4

The next part presents the proof of the properties given in Section 4. First, for a fixed $p \in \mathcal{P}$, we give a different representation of the sets $S \in \mathcal{S}$ with $p \in S$.

Lemma 1 *For $p \in \mathcal{P}$ we get*

$$\{S \in \mathcal{S} \mid p \in \gamma(S)\} \tag{1}$$

$$= \bigcap_{(y_i,y_j)\in p} \{S \in \mathcal{S} \mid \exists x \in S\colon (y_i, y_j) \in x\} \cap \tag{2}$$

$$\bigcap_{(y_i,y_j)\notin p} \{S \in \mathcal{S} \mid \exists x \in S\colon (y_i, y_j) \notin x\}. \tag{3}$$

Proof Let $p \in \mathcal{P}$. The proof is divided into two parts.
Part 1: We prove $\subseteq$. Let S be an element of (1). Since $p \in \gamma(S)$, we have $p \subseteq \cup_{\tilde{p}\in S}\, \tilde{p}$. So for every $(y_i, y_j) \in p$ there is a $\tilde{p} \in S$ such that $(y_i, y_j) \in \tilde{p}$. Therefore, S is an element of the intersection of (2). Also from $p \in \gamma(S)$ we get $\cap_{\tilde{p}\in S}\tilde{p} \subseteq p$ and thus we know that for every $(y_i, y_j) \notin p$ there exists a $\tilde{p} \in S$ such that $(y_i, y_j) \notin \tilde{p}$. Thus, S is an element of the intersection given by (3).
Part 2: We prove $\supseteq$. Therefore, let $S \in \mathcal{S}$ be an element of the right-hand side of the equation. We show that $p \in \gamma(S)$. Let S be in the intersection given by (2). Then we know that for every $(y_1, y_2) \in p$ there exists an $\tilde{p} \in S$ such that $(y_1, y_2) \in \tilde{p}$. Thus $p \subseteq \cup_{\tilde{p}\in S}\, \tilde{p}$. The second part of the intersection given by (3) analogously yields that $\cap_{\tilde{p}\in S}\tilde{p} \subseteq p$. Hence $p \in \gamma(S)$ and the second part is proven. The claim follows from Part 1 and Part 2. ■

The next theorem provides some information about the properties of the elements in $\mathcal{S}$.

Theorem 2 *The family of sets $\mathcal{S}$ given in Section 3 fulfills the following properties.*

1. *For every $p \in \mathcal{P}$, $\{p\} \notin \mathcal{S}$.*
2. *Let $\{p_1, p_2\} = S \in 2^{\mathcal{P}}$. Then $S \notin \mathcal{S}$ iff the transitive reductions $tr(p_1)$ and $tr(p_2)$ differ only on one (y_i, y_j) which is only contained in exactly either $tr(p_1)$ or $tr(p_2)$. This means that either $\#(tr(p_1) \setminus tr(p_2)) = 1$ or $\#(tr(p_2) \setminus tr(p_1)) = 1$ holds.*
3. *$\mathcal{S}$ is connected in the sense that for every set $S \in \mathcal{S}$ of size $m \geq 3$ there exists a subset $\tilde{S} \subsetneq S$ of size $m-1$ that is in $\mathcal{S}$, too.*

Proof Claim 1. follows directly from Condition (C1) of the definition of $\mathcal{S}$ as $\gamma(\{p\}) = \{p\}$ for every $p \in \mathcal{P}$.

Now, we show the second claim. Let us first assume that $\{p_1, p_2\} = S \notin \mathcal{S}$. Then there exists no $p \in \mathcal{P}$ such that $p \in \gamma(S) \setminus \{p_1, p_2\}$. Thus, the intersection must be either p_1 or p_2, (otherwise $p_1 \cap p_2 \in \gamma(S) \setminus S$). W.l.o.g., let $p_1 = p_1 \cap p_2$. Then p_2 must be a superset of p_1 where there is no poset lying between p_1 and p_2. Therefore, $\#\{tr(p_2) \setminus tr(p_1)\} = 1$ is true. Conversely, assume that $S \in \mathcal{S}$ and that p_1 is a superset of p_2. With this we obtain that $\gamma(S) = \{p \in \mathcal{P} \mid p_2 \subseteq p \subseteq p_1\}$. Further assume that $\#\{tr(p_1) \setminus tr(p_2)\} = 1$ holds. But then $\gamma(S) = S$ is true, since no $p \in \mathcal{P}$ can lie between p_1 and p_2, but this is a contradiction which proves the claim.

The proof of the last part uses that the closure operator γ stems from a formal context, which is a term from formal concept analysis. Since formal concept analysis is not part of this paper, we have outsourced the proof to [36]. ■

Using the theorem above, we can determine properties for ν such that $D \equiv 0$ is true.

Corollary 3 *$D(p) = 0$ for every $p \in \mathcal{P}$ iff the measure ν has either the entire positive probability mass on a single poset p or only on exactly two posets p_1 and p_2 where the transitive reduction differs only in a pair (y_1, y_2). More* *precisely, either* $\#\{tr(p_1) \setminus tr(p_2)\} = 1$ *or* $\#\{tr(p_2) \setminus tr(p_1)\} = 1$.

Proof Note that $D(p) = 0$ for every $p \in \mathcal{P}$ is true if for all $S \in \mathcal{S}$, $\prod_{\tilde{p} \in S} \nu(\tilde{p}) = 0$. Theorem 2 1. and 2. provide the cases when this holds which proves immediately the claim.

The converse follows analogously from Theorem 2. ■

We use Lemma 1 to prove the sample influence:

Corollary 4 *Let* $(p_1, \ldots, p_n)$ *be a sample of* $\mathcal{P}$. *Let* ν_n *be the empirical probability measure induced by* $(p_1, \ldots, p_n)$. *Furthermore, let* ν_n *be such a probability measure that* $D_n \not\equiv 0$. *Then for* D_n *defined in Section 3, it holds.*

1. *Assume that for all* $p_i \in \{p_1, \ldots, p_n\}$, $(y_1, y_2) \in p_i$ *is true. Then for every poset* $p \in \mathcal{P}$ *with* $(y_1, y_2) \notin p$, $D_n(p) = 0$ *follows.*
2. *Assume that for all* $p_i \in \{p_1, \ldots, p_n\}$, $(y_1, y_2) \notin p_i$ *holds. Then for every poset* $p \in \mathcal{P}$ *with* $(y_1, y_2) \in p$, $D_n(p) = 0$ *is true.*
3. *Let* p_{Δ} *be the poset consisting only of the reflexive part. If* $D_n(p_{\Delta}) = 0$, *then there exists a pair* (y_1, y_2) *such that for all* $p_i \in \{p_1, \ldots, p_n\}$, $(y_1, y_2) \in p_i$.
4. *Let* $p_{total} \in \mathcal{P}$ *be a total order. If* $D_n(p_{total}) = 0$, *then there exists a pair* $(y_1, y_2) \notin p_{total}$ *such that for all* $p_i \in \{p_1, \ldots, p_n\}$, $(y_1, y_2) \in p_i$ *is true.*

Proof First, note that for $S \in \mathcal{S}$, where there exists an $\tilde{p} \in S$ such that $\nu_n(\tilde{p}) = 0$, S contributes nothing to D_n. So one can replace $\mathcal{S}$ in the definition of D_n by $\tilde{\mathcal{S}} = \{S \cap \{p_1, \ldots, p_n\} \mid S \in \mathcal{S}\}$. The reduced set $\tilde{\mathcal{S}}$ is used to show the claims.

Claims 1., 2.,3. and 4. are analogous. Hence, here we prove only Claim 1. Let $(y_1, y_2) \in M \times M$ such that for all $i \in \{1, \ldots, n\}$ $(y_1, y_2) \in p_i$ and let $p \in \mathcal{P}$ such that $(y_1, y_2) \notin p$. Let $S \in \mathcal{S} \cap \{p_1, \ldots, p_n\}$ and take a closer look at (3) of Lemma 1. Since $(y_1, y_2) \notin p$, S cannot be an element of the intersection of (3). Thus, $\{S \cap \{p_1, \ldots, p_n\} \in \mathcal{S} \mid p \in \gamma(S)\}$ is empty and with the comment above we get that $D_n(p) = 0$. ■

The next theorem gives an upper and lower bound on the cardinality of the elements $S \in \mathcal{S}$.

Theorem 5 *For* $\mathcal{S}$, *as defined in Section 3,* $\#S \geq 2$ *and* $\#S \leq vc$ *is true for all* $S \in \mathcal{S}$, *where* vc *is the VC dimension of the set* $\gamma(2^{\mathcal{P}})$.

Proof Let $S \in \mathcal{S}$. The proof for $\#S \geq 2$ follows immediately from Theorem 2.

To prove $\#S \leq vc$ take an arbitrary subset $Q = \{p_1, \ldots, p_k\} \in \mathcal{S}$ of size $k > vc$. Then this subset is not shatterable and thus there exists a subset $R \subseteq Q$ that cannot be obtained as an intersection of Q and some $\gamma(S)$. In particular, with the extensivity of γ it follows $\gamma(R) \cap Q \supsetneq R$ which means that there exists an order p_i in $\gamma(R) \cap Q \backslash R$ for which the formal implication $R \rightarrow \{p_i\}$ holds. Thus, (because of the Armstrong rules, cf., [1, p. 581]) the order p_i is redundant in the sense of $Q \backslash \{p_i\} \rightarrow Q$ and thus Q is not minimal with respect to γ. Therefore, Q is not in $\mathcal{S}$ which completes the proof. ■

Finally, we show the consistency of D_n.

Theorem 6 D_n *converges almost surely uniformly to D for n to infinity.*

Proof Due to the i.i.d assumption and the law of large numbers, we know that for every $p \in \mathcal{P}$, $\|\nu_n(p) - \nu(p)\| \overset{n\to\infty}{\longrightarrow} 0$ almost surely (a.s). Since $\#\mathcal{P}$ is finite, we get that ν_n also converges a.s. uniformly to ν. Finally, we use that D_n and D are both the same finite composition of ν_n and ν, respectively, and we obtain $\sup_{p \in \mathcal{P}} \|D_n(p) - D(p)\| \overset{n\to\infty}{\longrightarrow} 0$ a.s. ■

The last theorem states a contradiction to the claim that D can be represented via pairwise comparisons.

Theorem 7 D_n *cannot be represented as a function of the sum-statistics* $w_{(a,b)}$.

Proof We simply give two data sets $\mathcal{D} = (p_1, p_2, p_3)$ and $\tilde{\mathcal{D}} = (\tilde{p}_1, \tilde{p}_2, \tilde{p}_3)$ on the basic set $M = \{y_1, y_2, y_3\}$ with the same sum-statistics but different associated depth functions: Let p_1, p_2 and p_3 be given as the transitive reflexive closures of $\{(y_1, y_2)\}$; $\{(y_1, y_2), (y_1, y_3)\}$ and $\{(y_2, y_3), (y_1, y_3)\}$, respectively. Let $\tilde{p}_1$, $\tilde{p}_2$ and $\tilde{p}_3$ be the transitive reflexive closure of $\{(y_1, y_2)\}$; $\{(y_1, y_3)\}$ and $\{(y_1, y_2), (y_2, y_3)\}$, respectively. Then both data sets have the same sum-statistics $w_{(y_1,y_2)} = w_{(y_1,y_3)} = 2$; $w_{(y_1,y_3)} = 1$ and $w_{(y_i,y_j)} = 0$ for all other $y_i \neq y_j$. But the ufg depth of $p_1 = \tilde{p}_1$ is $1/2$ w.r.t. the first data set but $7/10$ w.r.t the second data set. The corresponding code can be found at the link mentioned in Footnote 1. ■

Acknowledgments

We thank all four anonymous reviewers for valuable comments that helped to improve the paper. Hannah Blocher and Georg Schollmeyer gratefully acknowledge the financial and general support of the LMU Mentoring program.

Author Contributions

Hannah Blocher developed the idea of ufg depth. She wrote most of the paper. To be more precise: The introduction was written by Christoph Jansen. Hannah Blocher wrote the preliminaries and defined the (empirical) ufg depth. Furthermore, Hannah Blocher claimed and proved Lemma 1, Theorem 2, 1st and 2nd part, Corollary 3, Corollary 4, the lower bound of Theorem 5 and Theorem 6. Georg Schollmeyer made the claim and proofs of Theorem 2, 3rd part and the upper bound of Theorem 5. The claim of Theorem 7 was done by Georg Schollmeyer and Christoph Jansen. Georg Schollmeyer proved Theorem 7. Chapter 5 was written by Georg Schollmeyer. Hannah Blocher wrote Chapter 6.1 and implemented the test if a subset is an element of $\mathscr{S}$. Georg Schollmeyer contributed with the implementation of the connectedness property. Malte Nalenz provided the data set, performed the data preparation and wrote Chapter 6.2. Hannah Blocher analyzed the data set and wrote Chapter 6.3. Georg Schollmeyer, Christoph Jansen and Malte Nalenz supported the analysis with intensive discussions. Christoph Jansen provided the conclusion.

Georg Schollmeyer and Christoph Jansen also helped with discussions about the definition of ufg depth and all properties. Malte Nalenz, Christoph Jansen, and Georg Schollmeyer also contributed by providing detailed proofreading and help with the general structure of the paper.

Code & Data Availability

Reproducible implementation and data analysis are available at: www.github.com/hannahblo/23_Performance_Analysis_ML_Algorithms.

References

[1] W. Armstrong. Dependency structures of data base relationships. *International Federation for Information Processing Congress*, 74:580–583, 1974.

[2] Y. Bastide, N. Pasquier, R. Taouil, G. Stumme, and L. Lakhal. Mining minimal non-redundant association rules using frequent closed itemsets. In J. Lloyd, V. Dahl, U. Furbach, M. Kerber, K. Lau, C. Palamidessi, L. Pereira, Y. Sagiv, and P. Stuckey, editors, *Computational Logic — CL 2000*, pages 972–986. Springer, 2000.

[3] 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.

[4] K. Bertet, C. Demko, J. Viaud, and C. Guérin. Lattices, closures systems and implication bases: A survey of structural aspects and algorithms. *Theoretical Computer Science*, 743:93–109, 2018.

[5] H. Blocher, G. Schollmeyer, and C. Jansen. Statistical models for partial orders based on data depth and formal concept analysis. In D. Ciucci, I. Couso, J. Medina, D. Slezak, D. Petturiti, B. Bouchon-Meunier, and R. Yager, editors, *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 17–30. Springer, 2022.

[6] R. Bradley and M. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[7] C. Chang, J. Jiménez-Martín, E. Maasoumi, and T. Pérez-Amaral. A stochastic dominance approach to financial risk management strategies. *Journal of Econometrics*, 187(2):472–485, 2015.

[8] L. Chang. Partial order relations for classification comparisons. *Canadian Journal of Statistics*, 48(2): 152–166, 2020.

[9] I. Couso and D. Dubois. Statistical reasoning with set-valued information: Ontic vs. epistemic views. *International Journal of Approximate Reasoning*, 55 (7):1502–1518, 2014.

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

[11] J. Eckhoff. Helly, Radon, and Carathéodory type theorems. In *Handbook of Convex Geometry*, pages 389–448. Elsevier, 1993.

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

[13] M. Fligner and J. Verducci. Distance based ranking models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(3):359–369, 1986.

[14] J. Friedman, T. Hastie, R. Tibshirani, B. Narasimhan, K. Tay, N. Simon, and J. Qian. Package 'glmnet'. *CRAN R Repositary*, 2021.

[15] B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer, 2012.

[16] K. Hechenbichler and K. Schliep. Weighted k-nearest-neighbor techniques and ordinal classification. Technical Report, LMU, 2004. URL http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-1769-9.

[17] T. Hothorn, F. Leisch, A. Zeileis, and K. Hornik. The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics*, 14(3):675–699, 2005.

[18] 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.

[19] C. Jansen, G. Schollmeyer, and T. Augustin. A probabilistic evaluation framework for preference aggregation reflecting group homogeneity. *Mathematical Social Sciences*, 96:49–62, 2018.

[20] 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.

[21] C. Jansen, M. Nalenz, G. Schollmeyer, and T. Augustin. Statistical comparisons of classifiers by generalized stochastic dominance. Arxiv Preprint, 2022. URL https://arxiv.org/abs/2209.01857.

[22] D. Kikuti, F. Cozman, and R. Filho. Sequential decision making with partially ordered preferences. *Artificial Intelligence*, 175:1346 – 1365, 2011.

[23] G. Lebanon and Y. Mao. Non-parametric modeling of partially ranked data. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

[24] H. Levy and A. Levy. Ordering uncertain options under inflation: A note. *The Journal of Finance*, 39 (4):1223–1229, 1984.

[25] R. Liu. On a notion of data depth based on random simplices. *The Annals of Statistics*, 18:405–414, 1990.

[26] S. López-Pintado and J. Romo. On the concept of depth for functional data. *Journal of the American statistical Association*, 104(486):718–734, 2009.

[27] D. Mauá, F. Cozman, D. Conaty, and C. Campos. Credal sum-product networks. In A. Antonucci, G. Corani, I. Couso, and S. Destercke, editors, *International Symposium on Imprecise Probability: Theories and Applications*, volume 62, 10–14 Jul 2017.

[28] K. Mosler. *Multivariate Dispersion, Central Regions, and Depth: The Lift Zonoid Approach*. Springer, 2002.

[29] K. Mosler and P. Mozharovskyi. Choosing among notions of multivariate depth statistics. *Statistical Science*, 37:348–368, 2022.

[30] K. Nakamura, K. Yano, and F. Komaki. Learning partially ranked data based on graph regularization. *arXiv preprint arXiv:1902.10963*, 2019.

[31] M. Pini, F. Rossi, K. Venable, and T. Walsh. Incompleteness and incomparability in preference aggregation: Complexity results. *Artificial Intelligence*, 175 (7):1272–1289, 2011.

[32] J. Plass, T. Augustin, M. Cattaneo, and G. Schollmeyer. Statistical modelling under epistemic data imprecision: some results on estimating multinomial distributions and logistic regression for coarse categorical data. In T. Augustin, S. Doria, E. Miranda, and E. Quaeghebeur, editors, *ISIPTA '15, Proceedings of the Ninth International Symposium on Imprecise Probability: Theories and Applications*, 2015.

[33] J. Plass, P. Fink, N. Schöning, and T. Augustin. Statistical modelling in surveys without neglecting the undecided: Multinomial logistic regression models and imprecise classification trees under ontic data imprecision. In T. Augustin, S. Doria, E. Miranda, and E. Quaeghebeur, editors, *ISIPTA '15, Proceedings of the Ninth International Symposium on Imprecise Probability: Theories and Applications*, 2015.

[34] G. Schollmeyer. Lower quantiles for complete lattices. Technical Report, LMU, 2017. URL http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-40448-7.

[35] G. Schollmeyer. Application of lower quantiles for complete lattices to ranking data: Analyzing outlyingness of preference orderings. Technical Report, LMU, 2017. URL http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-40452-9.

[36] G. Schollmeyer and H. Blocher. A note on the connectedness property of union-free generic sets of partial orders. Arxiv Preprint, 2023. URL https://www.foundstat.statistik.uni-muenchen.de/personen/mitglieder/blocher/index.html.

[37] G. Schollmeyer, C. Jansen, and T. Augustin. Detecting stochastic dominance for poset-valued random variables as an example of linear programming on closure systems. Technical Report, LMU, 2017. URL http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-40416-0.

[38] T. Seidenfeld, J. Kadane, and M. Schervish. A representation of partially ordered preferences. *Annals of Statistics*, 23:2168–2217, 1995.

[39] R. Serfling and Y. Zuo. General notions of statistical depth function. *The Annals of Statistics*, 28(2):461 – 482, 2000.

[40] J. Stoye. Statistical inference for interval identified parameters. In T. Augustin, F. Coolen, S. Moral, and M. Troffaes, editors, *ISIPTA '09, Proceedings of the Sixth International Symposium on Imprecise Probabilities: Theories and Applications*, 2009.

[41] T. Therneau, B. Atkinson, and B. Ripley. Package 'rpart', 2015. URL http://cran.ma.ic.ac.uk/web/packages/rpart/rpart.pdf. [Accessed: 15.02.2023].

[42] J. Tukey. Mathematics and the picturing of data. In R. James, editor, *Proceedings of the International Congress of Mathematicians Vancouver*, pages 523–531, Vancouver, 1975. Mathematics-Congresses.

[43] J. Vanschoren, J. van Rijn, B. Bischl, and L. Torgo. Openml: Networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013.

[44] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In V. Vovk, H. Papadopoulos, and A. Gammerman, editors, *Measures of Complexity: Festschrift for Alexey Chervonenkis*, pages 11–30. Springer, 2015.

[45] M. Wright and A. Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1): 1–17, 2017.

[46] M. Zaffalon. The naive credal classifier. *Journal of Statistical Planning and Inference*, 105(1):5–21, 2002.

[47] M. Zaffalon, G. Corani, and D. Mauá. Evaluating credal classifiers by utility-discounted predictive accuracy. *International Journal of Approximate Reasoning*, 53(8):1282–1301, 2012.