On the VC-Dimension of the Choquet Integral

Eyke Hüllermeier and Ali Fallah Tehrani

Department of Mathematics and Computer Science
University of Marburg, Germany
{eyke,fallah}@mathematik.uni-marburg.de

Abstract. The idea of using the Choquet integral as an aggregation operator in machine learning has gained increasing attention in recent years, and a number of corresponding methods have already been proposed. Complementing these contributions from a more theoretical perspective, this paper addresses the following question: What is the VC dimension of the (discrete) Choquet integral when being used as a binary classifier? The VC dimension is a key notion in statistical learning theory and plays an important role in estimating the generalization performance of a learning method. Although we cannot answer the above question exactly, we provide a first interesting result in the form of (relatively tight) lower and upper bounds.

1 Introduction

While being widely used and well-known as a flexible aggregation operator in different research fields and application areas, such as multiple criteria decision making [1–3], the Choquet integral is much less common in machine learning so far. Nevertheless, the interest in using the Choquet integral as a mathematical tool in machine learning is increasing, and several papers on its use for problems like classification and regression have been published recently [4–8].

Indeed, the Choquet integral exhibits a number of properties that appear to be appealing from a machine learning point of view. For example, the authors in [9] especially advocate the Choquet integral as a tool for learning monotone nonlinear models. They specifically emphasize the fact that, by its very nature as an integral, the Choquet integral is a monotone operator and hence assures a monotone behavior in each individual attribute; this property is often desirable and sometimes even requested by the application [10–12]. At the same time, however, the Choquet integral also allows for modeling interactions between different attributes. Moreover, thanks to the existence of natural measures for quantifying the influence of individual and the interaction between groups of features, it provides important insights into the model, thereby supporting interpretability.

The use of the Choquet integral in the context of machine learning begs an interesting theoretical question, namely the question concerning its "expressive power" or, say, "flexibility" for modeling functional dependencies (in regression) and decision boundaries (in classification). This question is closely connected to the notion of the *capacity* of a model class (hypothesis space) in machine

learning, which is a key notion in statistical learning theory and plays an important role in estimating the generalization performance of a learning method [13]. An important measure of the capacity of a model class is the so-called Vapnik–Chervonenkis (VC) dimension. In this paper, we therefore address the following question: What is the VC dimension of the (discrete) Choquet integral when being used as a binary classifier? Although we cannot answer this question exactly, we provide a first interesting result in the form of (relatively tight) lower and upper bounds.

The rest of this paper is organized as follows. In the next section, we briefly recall the basic definition of the (discrete) Choquet integral and some related notions. In Section 3, we sketch the idea of using the Choquet integral for modeling decision boundaries in the setting of binary classification. In Section 4, we introduce the definition of VC dimension. Our main result in then presented in Section 5, prior to concluding the paper with a few remarks in Section 6.

2 The Discrete Choquet Integral

To make the paper self-contained and, moreover, to recall the basic mathematical notation to be used later on, this section starts with a brief introduction to the (discrete) Choquet integral.

2.1 Non-Additive Measures

Let $X = \{x_1, \ldots, x_m\}$ be a finite set and $\mu(\cdot)$ a measure $2^X \to [0, 1]$. For each $A \subseteq X$, we interpret $\mu(A)$ as the *weight* or, say, the *importance* of the set of elements A. A standard assumption on a measure $\mu(\cdot)$, which is, for example, at the core of probability theory, is additivity: $\mu(A \cup B) = \mu(A) + \mu(B)$ for all $A, B \subseteq X$ such that $A \cap B = \emptyset$. Unfortunately, additive measures cannot model any kind of interaction between elements: Extending a set of elements A by a set of elements B always increases the weight $\mu(A)$ by the weight $\mu(B)$, regardless of A and B.

Non-additive measures, also called capacities or fuzzy measures, are simply normalized and monotone [14]:

$$\mu(\emptyset) = 0,\ \mu(X) = 1 \quad \text{and} \quad \mu(A) \leq \mu(B) \text{ for all } A \subseteq B \subseteq X\ . \tag{1}$$

A useful representation of non-additive measures, that we shall explore later on, is in terms of the *Möbius transform*:

$$\mu(B) = \sum_{A \subseteq B} \boldsymbol{m}_\mu(A) \tag{2}$$

for all $B \subseteq X$, where the Möbius transform $\boldsymbol{m}_\mu$ of the measure μ is defined as follows:

$$\boldsymbol{m}_\mu(A) = \sum_{B \subseteq A} (-1)^{|A|-|B|} \mu(B)\ . \tag{3}$$

The value $\boldsymbol{m}_\mu(A)$ can be interpreted as the weight that is *exclusively* allocated to A, instead of being indirectly connected with A through the interaction with other subsets.

A measure μ is said to be k-order additive, or simply k-additive, if k is the smallest integer such that $\boldsymbol{m}_\mu(A) = 0$ for all $A \subseteq X$ with $|A| > k$. This property is interesting for several reasons. First, as can be seen from (2), it means that a measure μ can formally be specified by significantly fewer than 2^m values, which are needed in the general case. Second, k-additivity is also interesting from a semantic point of view: This property simply means that there are no interaction effects between subsets $A, B \subseteq X$ whose cardinality exceeds k.

2.2 The Choquet Integral

So far, the criteria $x_i \in X$ were simply considered as binary features, which are either present or absent. Mathematically, $\mu(A)$ can thus also be seen as an *integral* of the indicator function of A, namely the function f_A given by $f_A(x) = 1$ if $x \in A$ and $= 0$ otherwise. Now, suppose that $f : X \to \mathbb{R}_+$ is any non-negative function that assigns a *value* to each criterion x_i. An important question, then, is how to *aggregate* the evaluations of individual criteria, i.e., the values $f(x_i)$, into an overall evaluation, in which the criteria are properly weighted according to the measure μ. Mathematically, this overall evaluation can be considered as an integral $C_\mu(f)$ of the function f with respect to the measure μ.

Indeed, if μ is an additive measure, the standard integral just corresponds to the *weighted mean*

$$C_\mu(f) = \sum_{i=1}^{m} w_i \cdot f(x_i) = \sum_{i=1}^{m} \mu(\{x_i\}) \cdot f(x_i) \;, \tag{4}$$

which is a natural aggregation operator in this case. A non-trivial question, however, is how to generalize (4) in the case where μ is non-additive. This question is answered by the Choquet integral, which, in the discrete case, is formally defined as follows:

$$C_\mu(f) = \sum_{i=1}^{m} \big(f(x_{(i)}) - f(x_{(i-1)})\big) \cdot \mu(A_{(i)}) \;,$$

where $(\cdot)$ is a permutation of $\{1, \ldots, m\}$ such that $0 \leq f(x_{(1)}) \leq f(x_{(2)}) \leq \ldots \leq f(x_{(m)})$ (and $f(x_{(0)}) = 0$ by definition), and $A_{(i)} = \{x_{(i)}, \ldots, x_{(m)}\}$. In terms of the Möbius transform $\boldsymbol{m} = \boldsymbol{m}_\mu$ of μ, the Choquet integral can also be expressed

as follows:

$$\begin{aligned} C_\mu(f) &= \sum_{i=1}^{m} \big(f(x_{(i)}) - f(x_{(i-1)})\big) \cdot \mu(A_{(i)}) \\ &= \sum_{i=1}^{m} f(x_{(i)}) \cdot (\mu(A_{(i)}) - \mu(A_{(i+1)})) \\ &= \sum_{i=1}^{m} f(x_{(i)}) \sum_{R \subseteq T_{(i)}} \boldsymbol{m}(R) \\ &= \sum_{T \subseteq X} \boldsymbol{m}(T) \times \min_{i \in T} f(x_i) \end{aligned} \tag{5}$$

where $T_{(i)} = \big\{S \cup \{(i)\} \,|\, S \subset \{(i+1), \ldots, (m)\}\big\}$.

3 The Choquet Integral as a Tool for Classification

As mentioned earlier, the Choquet integral has been used as a tool for different types of machine learning problems. In the following, we focus on the setting of binary classification, where the goal is to predict the value of an output (response) variable $y \in \mathcal{Y} = \{0, 1\}$ for a given instance

$$\boldsymbol{x} = (x_1, \ldots, x_m) \in \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_m$$

represented in terms of a feature vector. More specifically, the goal is to learn a classifier $\mathcal{L} : \mathcal{X} \to \mathcal{Y}$ from a given set of (i.i.d.) training data

$$\mathcal{D} = \Big\{(\boldsymbol{x}^{(i)}, y^{(i)})\Big\}_{i=1}^{n} \subset (\mathcal{X} \times \mathcal{Y})^n \tag{6}$$

so as to minimize the risk

$$R(\mathcal{L}) = \int_{\mathcal{X} \times \mathcal{Y}} \ell(\mathcal{L}(\boldsymbol{x}), y)\, d\mathbf{P}_{XY}(\boldsymbol{x}, y)\ , \tag{7}$$

where $\ell(\cdot)$ is a loss function (e.g., the simple 0/1 loss given by $\ell(\hat{y}, y) = 0$ if $\hat{y} = y$ and $= 1$ if $\hat{y} \neq y$).

In this context, the predictor variables (features) play the role of the criteria in decision making. The Choquet integral can be used in order to model nonlinear dependencies between these variables and the response, thus taking interactions between predictors into account while preserving monotonicity in each individual feature. This can be done in different ways. In [15], for example, the authors propose a model that can be seen as an extension of logistic regression. The basic idea of this approach is to model the log-odds ratio between the positive ($y = 1$) and the negative ($y = 0$) class as a function of the Choquet integral of the input attributes. This leads to expressing the (posterior) probability of the positive class (and hence of the negative class) as follows:

$$\mathbf{P}(y = 1 \,|\, \boldsymbol{x}) = \Big(1 + \exp\big(-\gamma\,(\mathcal{C}_\mu(\boldsymbol{x}) - \beta)\big)\Big)^{-1}\ , \tag{8}$$

where $\mathcal{C}_\mu(\boldsymbol{x})$ is the Choquet integral (with respect to the measure μ) of the function

$$f_{\boldsymbol{x}} : \{c_1, \ldots, c_m\} \to [0, 1] \tag{9}$$

that maps each attribute c_i to a normalized feature value $x_i = f_{\boldsymbol{x}}(c_i) \in [0, 1]$; $\beta, \gamma \in \mathbb{R}$ are constants.

The (machine) learning problem itself can then be stated as follows: Given a set of training data (6), find a fuzzy measure μ and parameters β, γ, such that the corresponding model (8) generalizes well in terms of the risk (7).

4 The VC Dimension

In machine learning, it is well-known that the generalization performance of a learning algorithm strongly depends on the *capacity*[1] or, say, flexibility of the underlying model class $\mathcal{H}$, also called the hypothesis space. In fact, if $\mathcal{H}$ is not flexible enough, the true underlying dependency between predictor variables and response cannot be captured in a sufficiently accurate way; correspondingly, the training data will typically be "under-fitted". For example, if two classes are separated by a quadratic discriminant function, it is not enough to fit only a linear decision boundary (i.e., to define $\mathcal{H}$ as the set of all linear discriminant functions). On the other hand, if the flexibility of $\mathcal{H}$ is too high, there is a strong danger of "over-fitting" the training data. The notion of "over-fitting" refers to situations in which the learned model fails to produce good predictions for instances not seen so far, although it is able to reproduce the training data quite accurately.

The question of how to choose a model class $\mathcal{H}$ having the right capacity can be approached in different ways, both theoretically and empirically. From a theoretical point of view, it is convenient to have a measure that allows one to quantify the capacity of a model class. One of the most important measures of that kind, which is often used to estimate the generalization performance of a learning algorithm, is the so-called Vapnik–Chervonenkis (VC) dimension [13].

Definition 1. *The VC dimension of a model class $\mathcal{H} \subset 2^{\mathcal{X}}$ is defined as the maximum number of instances $\boldsymbol{x} \in \mathcal{X}$ that can be* shattered*:*

$$VC(\mathcal{H}) \;=\; \max\{|\mathcal{D}| \mid \mathcal{D} \subseteq \mathcal{X} \text{ and } \mathcal{D} \text{ can be shattered by } \mathcal{H}\}$$

A set of instances $\mathcal{D}$ can be shattered by $\mathcal{H}$ if, for each subset $\mathcal{P} \subseteq \mathcal{D}$, there is a model $H \in \mathcal{H}$ such that $H(\boldsymbol{x}) = 1$ for all $\boldsymbol{x} \in \mathcal{P}$ and $H(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathcal{D} \setminus \mathcal{P}$.

In light of the aforesaid, advocating the Choquet integral as a novel tool for machine learning immediately begs the interesting theoretical question regarding the capacity of the corresponding model class. In fact, since the Choquet integral in its general form or, more specifically, the underlying fuzzy measure μ (not restricted to the k-additive case) has a rather large number of parameters, one

[1] Not to be confused with the use of same term for a non-additive measure.

may expect it to be quite flexible and, therefore, to have a high capacity. On the other hand, the parameters cannot be chosen freely. Instead, they are highly constrained due to the monotonicity properties that need to be satisfied by μ.

5 The VC Dimension of the Choquet Integral

We consider a setting in which the Choquet integral is used to classify instances represented in the form of m-dimensional vectors $\boldsymbol{x} = (x_1, x_2, \ldots, x_m) \in \mathbb{R}_+^m$, where $x_i = f(c_i)$ can be thought of as the evaluation of the criterion c_i. More specifically, we consider the model class $\mathcal{H}$ consisting of all threshold classifiers of the form

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_m) \mapsto \mathbb{I}\Big(\mathcal{C}_\mu(\boldsymbol{x}) > \beta\Big) \ , \tag{10}$$

where $\mathbb{I}$ maps truth degrees $\{\text{false}, \text{true}\}$ to $\{0, 1\}$ as usual, μ is a fuzzy measure, $\mathcal{C}_\mu(\boldsymbol{x})$ is the Choquet integral of the (normalized) attribute values $x_1, x_2, \ldots, x_m$, and $\beta \in [0, 1]$ is a threshold value. Note that the class $\mathcal{H}$ is parametrized by μ and β. In terms of the VC dimension, the model (10) is equivalent to most other models based on the Choquet integral that have been used in the literature so far, including (8).

Theorem 1. *For the model class $\mathcal{H}$ as defined above, $VC(\mathcal{H}) = \Omega(2^m/\sqrt{m})$. That is, the VC dimension of $\mathcal{H}$ grows asymptotically at least as fast as $2^m/\sqrt{m}$.*

Proof. In order to prove this claim, we construct a sufficiently large data set $\mathcal{D}$ and show that, despite its size, it can be shattered by $\mathcal{H}$. In this construction, we restrict ourselves to binary attribute values, which means that $x_i \in \{0, 1\}$ for all $1 \leq i \leq m$. Consequently, each instance $\boldsymbol{x} = (x_1, \ldots, x_m) \in \{0, 1\}^m$ can be identified with a subset of indices $S_{\boldsymbol{x}} \subseteq X = \{1, 2, \ldots, m\}$, namely its *indicator set* $S_{\boldsymbol{x}} = \{i \,|\, x_i = 1\}$.

In combinatorics, an *antichain* of $X = \{1, 2, \ldots, m\}$ is a family of subsets $\mathcal{A} \subset 2^X$ such that, for all $A, B \in \mathcal{A}$, neither $A \subseteq B$ nor $B \subseteq A$. An interesting question related to the notion of an antichain concerns its potential size, that is, the number of subsets in $\mathcal{A}$. This number is obviously restricted due to the above non-inclusion constraint on pairs of subsets. An answer to this question is given by a well-known result of Sperner [16], who showed that this number is

$$\binom{m}{\lfloor m/2 \rfloor} \ . \tag{11}$$

Moreover, Sperner has shown that the corresponding antichain $\mathcal{A}$ is given by the family of all q-subsets of X with $q = \lfloor m/2 \rfloor$, that is, all subsets $A \subset X$ such that $|A| = q$.

Now, we define the data set $\mathcal{D}$ in terms of the collection of all instances $\boldsymbol{x} = (x_1, \ldots, x_m) \in \{0, 1\}^m$ whose indicator set $S_{\boldsymbol{x}}$ is a q-subset of X. Recall that, from a decision making perspective, each attribute can be interpreted as a criterion. Thus, each instance in our data set satisfies exactly q of the m

criteria, and there is not a single "dominance" relation in the sense that the set of criteria satisfied by one instance is a superset of those satisfied by another instance. Intuitively, the instances in $\mathcal{D}$ are therefore maximally incomparable. This is precisely the property we are now going to exploit in order to show that $\mathcal{D}$ can be shattered by $\mathcal{H}$.

Recall that a set of instances $\mathcal{D}$ can be shattered by a model class $\mathcal{H}$ if, for each subset $\mathcal{P} \subseteq \mathcal{D}$, there is a model $H \in \mathcal{H}$ such that $H(\boldsymbol{x}) = 1$ for all $\boldsymbol{x} \in \mathcal{P}$ and $H(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathcal{D} \setminus \mathcal{P}$. Now, take any such subset $\mathcal{P}$ from our data set $\mathcal{D}$ as constructed above, and recall that the Choquet integral in (10) can be written as

$$\mathcal{C}_\mu(\boldsymbol{x}) = \sum_{T \subseteq C} \boldsymbol{m}(T) \times f_T(\boldsymbol{x}) \ ,$$

where $f_T(\boldsymbol{x}) = 1$ if $T \subseteq S_{\boldsymbol{x}}$ and $f_T(\boldsymbol{x}) = 0$ otherwise. We define the values $\boldsymbol{m}(T)$, $T \subseteq C$, of the Möbius transform as follows:

$$\boldsymbol{m}(T) = \begin{cases} |\mathcal{P}|^{-1} & \text{if } T = S_{\boldsymbol{x}} \text{ for some } \boldsymbol{x} \in \mathcal{P} \\ 0 & \text{otherwise} \end{cases} \ .$$

Obviously, this definition of the Möbius transform is feasible and yields a proper fuzzy measure μ: The sum of masses is equal to 1, and since all masses are non-negative, monotonicity is guaranteed right away. Moreover, from the construction of $\boldsymbol{m}$ and the fact that, for each pair $\boldsymbol{x} \neq \boldsymbol{x}' \in \mathcal{D}$, neither $S_{\boldsymbol{x}} \subseteq S_{\boldsymbol{x}'}$ nor $S_{\boldsymbol{x}'} \subseteq S_{\boldsymbol{x}}$, the Choquet integral is obviously given as follows:

$$\mathcal{C}_\mu = \begin{cases} |\mathcal{P}|^{-1} & \text{if } \boldsymbol{x} \in \mathcal{P} \\ 0 & \text{otherwise} \end{cases} \ .$$

Thus with $\beta = 1/(2|\mathcal{P}|)$, the classifier (10) behaves exactly as required, that is, it classifies all $\boldsymbol{x} \in \mathcal{P}$ as positive and all $\boldsymbol{x} \notin \mathcal{P}$ as negative.

Noting that the special case where $\mathcal{P} = \emptyset$ is handled correctly by the Möbius transform $\boldsymbol{m}$ such that $\boldsymbol{m}(C) = 1$ and $\boldsymbol{m}(T) = 0$ for all $T \subsetneq C$ (and any threshold $\beta > 0$), we can conclude that the data set $\mathcal{D}$ can be shattered by $\mathcal{H}$. Consequently, the VC dimension of $\mathcal{H}$ is at least the size of $\mathcal{D}$, whence (11) is a lower bound of $VC(\mathcal{H})$.

For the asymptotic analysis, we make use of Sterling's approximation of large factorials (and hence binomial coefficients). For the sequence $(b_1, b_2, \ldots)$ of the so-called central binomial coefficients b_n, it is known that

$$b_n = \binom{2n}{n} = \frac{(2n)!}{(n!)^2} \geq \frac{1}{2} \frac{4^n}{\sqrt{\pi \cdot n}} \ . \tag{12}$$

Thus, the fact that $VC(\mathcal{H})$ grows asymptotically at least as fast as $2^m/\sqrt{m}$ immediately follows by setting $n = m/2$ and ignoring constant terms.

Remark 1. Recall the expression (5) of the Choquet integral in terms of its Möbius transform. This expression shows that the Choquet integral corresponds

to a linear function, albeit a constrained one, in the *feature space* spanned by the set of features $\{f_T \,|\, T \subseteq \{1, 2, \ldots, m\}\}$, where each feature is a min-term

$$f_T = f_T(x_1, \ldots, x_m) \,=\, \min_{i \in T} x_i \ . \tag{13}$$

The dimensionality of this feature space is $2^m - 1$. Thus, it follows immediately that $VC(\mathcal{H}) \leq 2^m$ (the class of linear hyperplanes in $\mathbb{R}^n$ has VC-dimension $n+1$). Together with the lower bound $2^m/\sqrt{m}$, which is not much smaller (despite the restriction to binary attribute vectors), we thus dispose of a relatively tight approximation of $VC(\mathcal{H})$.

Remark 2. Interestingly, the proof of Theorem 1 does not exploit the full non-additivity of the Choquet integral. In fact, the measure we constructed there is $\lfloor m/2 \rfloor$-additive, since $\boldsymbol{m}(T) = 0$ for all $T \subseteq C$ with $|T| > \lfloor m/2 \rfloor$. Consequently, the estimation of the VC-dimension still applies to the restricted case of k-additive measures, provided $k \geq \lfloor m/2 \rfloor$. For smaller k, it is not difficult to adapt the proof so as to show that

$$VC(\mathcal{H}) \geq \binom{m}{k} \ . \tag{14}$$

6 Concluding Remarks

Our result shows that the VC dimension of the Choquet integral, when being used as a threshold classifier, grows almost exponentially with the number of attributes. Due to the strong monotonicity constraints on the underlying fuzzy measure, this level of flexibility was not necessarily expected. Anyway, it suggests that learning with the Choquet integral may come with the danger of over-fitting the training data.

On the other hand, one should keep in mind that the notion of VC dimension is based on a kind of worst case scenario. In fact, there are many examples of machine learning algorithms with a very high (or even infinite) VC dimension that practically perform quite well, at least when being combined with suitable methods for regularization. Thus, it might be of interest to complement our result with an empirical study, for example along the line of [17]. Moreover, our result also shows that a restriction to k-additive measures provides a suitable means for capacity control. An interesting question in this regard concerns the choice of a proper k providing the right level of flexibility for the data at hand.

Theoretically, it might be interesting to further tighten our bound. Indeed, since our result also holds for the restriction to binary features, one may expect that it is actually not as tight as it could be. The question whether or not this is indeed the case will be addressed in future work.

References

1. M. Grabisch, T. Murofushi, and M. Sugeno, editors. *Fuzzy Measures and Integrals: Theory and Applications.* Physica, 2000.

2. M. Grabisch. Fuzzy integral in multicriteria decision making. *Fuzzy Sets and Systems*, 69(3):279–298, 1995.
3. V. Torra. Learning aggregation operators for preference modeling. In J. Fürnkranz and E. Hüllermeier, editors, *Preference Learning*, pages 317–333. Springer, 2011.
4. M. Grabisch. Modelling data by the Choquet integral. In V. Torra, editor, *Information Fusion in Data Mining*, pages 135–148. Springer, 2003.
5. M. Grabisch and J.-M. Nicolas. Classification by fuzzy integral: performance and tests. *Fuzzy Sets and Systems*, 65(2-3):255–271, 1994.
6. V. Torra and Y. Narukawa. *Modeling Decisions: Information Fusion and Aggregation Operators*. Springer, 2007.
7. S. Angilella, S. Greco, and B. Matarazzo. Non-additive robust ordinal regression with Choquet integral, bipolar and level dependent Choquet integrals. In J. Carvalho, D. Dubois, U. Kaymak, and J. da Costa Sousa, editors, *Proceedings of the Joint 2009 International Fuzzy Systems Association World Congress and 2009 European Society of Fuzzy Logic and Technology Conference*, pages 1194–1199. IFSA/EUSFLAT, 2009.
8. G. Beliakov and S. James. Citation-based journal ranks: the use of fuzzy measures. *Fuzzy Sets and Systems*, 167(1):101–119, 2011.
9. A. Fallah Tehrani, W. Cheng, K. Dembczynski, and E. Hüllermeier. Learning monotone nonlinear models using the choquet integral. In *Proceedings ECML/PKDD–2011, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, Athens, Greece, 2011.
10. A. Ben-David. Monotonicity maintenance in information-theoretic machine learning algorithms. *Machine Learning*, 19:29–43, 1995.
11. R. Potharst and A. Feelders. Classification trees for problems with monotonicity constraints. *ACM SIGKDD Explorations Newsletter*, 4(1):1–10, 2002.
12. A. Feelders. Monotone relabeling in ordinal classification. In G. Webb, B. Liu, C. Zhang, D. Gunopulos, and X. Wu, editors, *Proceedings of the 10th IEEE International Conference on Data Mining*, pages 803–808. IEEE Computer Society, 2010.
13. V.N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.
14. M. Sugeno. *Theory of Fuzzy Integrals and its Application*. PhD thesis, Tokyo Institute of Technology, 1974.
15. A. Tehrani, W. Cheng, and E. Hüllermeier. Choquistic regression: Generalizing logistic regression using the Choquet integral. In S. Galichet, J. Montero, and G. Mauris, editors, *Proceedings Eusflat–2011, 7th International Conference of the European Society for Fuzzy Logic and Technology*, pages 868–875, Aix-les-Bains, France, 2011.
16. E. Sperner. Ein Satz über Untermengen einer endlichen Menge. *Mathematische Zeitschrift*, 27(1):544–548, 1928.
17. M. Pirlot, H. Schmitz, and P. Meyer. An empirical comparison of the expressiveness of the additive value function and the Choquet integral models for representing rankings. In *Proceedings URPDM–2010, Mini-EURO Conference Uncertainty and Robustness in Planning and Decision Making*, Coimbra, Portugal, 2010.