Efficient Learning of Classifiers Based on the 2-Additive Choquet Integral[*]

Eyke Hüllermeier and Ali Fallah Tehrani

Abstract. In a recent work, we proposed a generalization of logistic regression based on the Choquet integral. Our approach, referred to as *choquistic regression*, makes it possible to capture non-linear dependencies and interactions among predictor variables while preserving two important properties of logistic regression, namely the comprehensibility of the model and the possibility to ensure its monotonicity in individual predictors. Unsurprisingly, these benefits come at the expense of an increased computational complexity of the underlying maximum likelihood estimation. In this paper, we propose two approaches for reducing this complexity in the specific though practically relevant case of the 2-additive Choquet integral. Apart from theoretical results, we also present an experimental study in which we compare the two variants with the original implementation of choquistic regression.

1 Introduction

The Choquet integral is well-known as a flexible aggregation function and, as such, has been used in various fields of application [14, 11, 21]. In machine learning, it is less common so far, although the interest in using the Choquet integral as a mathematical tool for tackling problems like classification, regression and ranking is increasing [12, 13, 22, 1, 2, 9].

In [8], we proposed a method called "choquistic regression", which is a generalization of logistic regression based on the Choquet integral. Choquistic regression has a number of appealing properties. Most notably, it combines three features in a non-trivial way, namely monotonicity, nonlinearity and interpretability. As for the

Eyke Hüllermeier · Ali Fallah Tehrani
Department of Mathematics and Computer Science, University of Marburg,
35032 Marburg, German
e-mail: {eyke, fallah}@mathematik.uni-marburg.de

[*] Dedicated to Professor Rudolf Kruse on the occasion of his 60^{th} birthday.

C. Moewes et al. (Eds.): Computational Intelligence in Intelligent Data Analysis, SCI 445, pp. 17–29.
springerlink.com

first, a monotone dependence between the input and output attributes is often desirable in a classification setting and sometimes even requested by the application [3, 19, 10]. At the same time, the Choquet integral also allows for modeling interactions between different attributes in a flexible, nonlinear way. Last but not least, thanks to the existence of natural measures for quantifying the influence of individual (e.g., the Shapley value) and the interaction between groups of features (e.g., the interaction index), it provides important insights into the model, thereby supporting interpretability [7].

Compared to standard logistic regression, these benefits are coming at the expense of an increased computational complexity of the underlying learning algorithm, which solves a maximum likelihood estimation problem. This is mainly caused by the large number of parameters of the fuzzy measure on which the Choquet integral is based, and the complicated dependency between these parameters. In this paper, we propose two approaches for reducing this complexity in the specific though practically relevant case of the 2-additive Choquet integral. To this end, we shall try to optimally exploit the simplified structure of a 2-additive measure in comparison to a non-additive measure in the general case.

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 binary classification and recall the basics of choquistic regression. In Section 4, we develop two alternative formulations of the learning (likelihood maximization) problem, both pursuing the same goal of complexity reduction. In Section 5, we present an experimental study in which we compare the two variants with the original implementation of choquistic regression, prior to concluding the paper with a few remarks in Section 6.

2 The Discrete Choquet Integral

In this section, we start with a brief recapitulation of the (discrete) Choquet integral and, along the way, introduce the main mathematical notation used throughout the paper.

Let $C = \{c_1, \ldots, c_m\}$ be a finite set and $\mu : 2^C \rightarrow [0,1]$ a measure. For each $A \subseteq C$, the value $\mu(A)$ can be interpreted 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 $A, B \subseteq C$ 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 the "context" A.

This lack of expressivity motivates the use of non-additive measures, also called capacities or fuzzy measures, which are simply normalized and monotone but not necessarily additive [20]:

$$\begin{aligned} &\mu(\emptyset) = 0, \mu(C) = 1 \\ &\mu(A) \leq \mu(B) \text{ for all } A \subseteq B \subseteq C \end{aligned} \tag{1}$$

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

$$\mu(B) = \sum_{A \subseteq B} m_{\mu}(A) \tag{2}$$

for all $B \subseteq C$, where the Möbius transform m_{μ} of the measure μ is defined as follows:

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

A measure μ is said to be k-order additive, or simply k-additive, if k is the smallest integer such that $m(A) = 0$ for all $A \subseteq C$ with $|A| > k$. This property is interesting for several reasons. In particular, 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.

Suppose the "criteria" $c_i \in C$ are simply considered as binary features, which are either present or absent in a set A. Mathematically, $\mu(A)$ can then also be seen as an *integral* of the indicator function of A, namely the function f_A given by $f_A(c) = 1$ if $c \in A$ and $= 0$ otherwise. Now, suppose that $f : C \to \mathbb{R}_+$ is any non-negative function that assigns a *value* to each criterion c_i; for example, $f(c_i)$ might be the degree to which a candidate satisfies criterion c_i. An important question, then, is how to *aggregate* the evaluations of individual criteria, i.e., the values $f(c_i)$, 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 $\mathscr{C}_{\mu}(f)$ 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*

$$\mathscr{C}_{\mu}(f) = \sum_{i=1}^{m} w_i \cdot f(c_i) = \sum_{i=1}^{m} \mu(\{c_i\}) \cdot f(c_i) , \tag{4}$$

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, namely how to define the integral of a function with respect to a non-additive measure (not necessarily restricted to the discrete case), is answered in a satisfactory way by the Choquet integral, which has first been proposed for additive measures by Vitali [23] and later on for non-additive measures by Choquet [4]. In the discrete case, the Choquet integral is formally defined as follows:

$$\mathscr{C}_{\mu}(f) = \sum_{i=1}^{m} \Big(f(c_{(i)}) - f(c_{(i-1)}) \Big) \cdot \mu\left(A_{(i)}\right) ,$$

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

$$\mathscr{C}_\mu(f) = \sum_{T \subseteq C} m(T) \cdot \min_{i \in T} f(c_i) \tag{5}$$

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

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 $y \in \mathscr{Y} = \{0,1\}$ for a given instance represented in terms of a feature vector

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

More specifically, the goal is to learn a classifier $\mathscr{L} : \mathscr{X} \rightarrow \mathscr{Y}$ from a given set of (independent and identically distributed) training data

$$\mathscr{D} = \left\{ (x^{(i)}, y^{(i)}) \right\}_{i=1}^{n} \subset (\mathscr{X} \times \mathscr{Y})^n \tag{6}$$

so as to minimize the risk

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

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 $c_i \in C$. 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 the following, we propose a model that can be seen as an extension of logistic regression.

3.1 Choquistic Regression

The key idea of the method of "choquistic regression" as proposed in [8] 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; thus, the affine function $x \mapsto w_0 + w^\top x$ modeling the log-odds ratio in standard logistic regression is replaced by the Choquet integral. Formally, this leads to the following model:

$$\pi_c \overset{\mathrm{df}}{=} \mathbf{P}(y = 1 \,|\, x) = \frac{1}{1 + \exp\Big(-\gamma\big(\mathscr{C}_\mu(f_x) - \beta\big)\Big)} \ , \tag{8}$$

where $\mathscr{C}_\mu(f_x)$ is the Choquet integral (with respect to the measure μ) of the evaluation function $f_x : \{c_1,\ldots,c_m\} \to [0,1]$ that maps each attribute c_i to a value $x_i = f_x(c_i)$; $\beta,\gamma \in \mathbb{R}_+$ are constants. The value of x_i is normalized in order to turn each predictor variable into a criterion, i.e., a "the higher the better" attribute, and to assure commensurability between the criteria [18].

The model (8) has several degrees of freedom, namely the fuzzy measure μ (Möbius transform $m = m_\mu$), the threshold β and the scaling parameter γ. The goal of learning is to identify these degrees of freedom on the basis of the training data $\mathscr{D}$. Like in the case of standard logistic regression, it is possible to harness the maximum likelihood (ML) principle for this purpose. The log-likelihood of the parameters can be written as

$$\begin{aligned} l(m,\gamma,\beta) &= \log \mathbf{P}(\mathscr{D}\,|\,m,\beta,\gamma) \\ &= \log\left(\prod_{i=1}^{n} \mathbf{P}(y^{(i)}\,|\,x^{(i)};m,\beta,\gamma)\right) \\ &= \sum_{i=1}^{n} y^{(i)} \log \pi_c^{(i)} + \left(1-y^{(i)}\right)\log\left(1-\pi_c^{(i)}\right). \end{aligned} \tag{9}$$

This is a convex function with respect to m, γ, and β. The problem, now, is to maximize (9) while making sure that μ is a proper fuzzy measure. Formally, this leads to the following constrained optimization problem:

$$\max_{m,\gamma,\beta} \Big\{ -\gamma \sum_{i=1}^{n} (1-y^{(i)})\,(\mathscr{C}_m(x^{(i)})-\beta) \\ - \sum_{i=1}^{n} \log\Big(1+\exp(-\gamma(\mathscr{C}_m(x^{(i)})-\beta))\Big)\Big\}$$

such that

$$0 \le \beta \le 1$$

$$0 < \gamma$$

$$\sum_{T\subseteq C} m(T) = 1 \tag{10}$$

$$\sum_{B\subseteq A\setminus\{c_i\}} m(B\cup\{c_i\}) \ge 0 \quad \forall A\subseteq C,\, c_i\in A \tag{11}$$

4 Efficient Learning of 2-Additive Measures

Solving the above optimization problem is a non-trivial task and may become computationally expensive, mainly due to the constraints on the fuzzy measure μ.

In fact, since (11) needs to be satisfied for all subsets $A \subseteq C$, the number of these monotonicity constraints is given by $m2^{m-1}$ and thus grows exponentially with the number of attributes.

In the following, we restrict ourselves to the specific case of 2-additive fuzzy measures. This restriction is interesting for several reasons. In particular, one may of course hope for a gain in terms of computational efficiency, and indeed, this is the aspect that we shall focus on in the remainder of the paper. Besides, however, let us mention that a restriction of this kind is also interesting from a learning point of view: By allowing one to capture pairwise interactions between attributes, the 2-additive case is a proper generalization of the linear model, while at the same time, it is still reasonable in terms of the number of degrees of freedom. In fact, while the number of parameters to be estimated is exponential (in the number of attributes) in general, it is only quadratic in the 2-additive case. Practically, we could observe that the high flexibility of the general model is rarely needed; on the contrary, it often leads to problems of over-fitting the data, thereby compromising generalization performance.

Coming back to the computational aspect, the number of parameters to be estimated is indeed reduced, since $m(A) = 0$ for all $A \subseteq C$ such that $|A| > 2$. On the other hand, it is important to observe that the number of constraints does *not* reduce: Although the number of summands in each of the constraints (11) becomes smaller (since many of them are now 0), the number of constraints themselves remains the same.

In the following, we shall therefore look for ways to exploit the simplified structure of the 2-additive case in order to reduce the number of constraints. More specifically, we shall propose two alternative formulations of the constraint optimization problem to be solved for ML estimation.

4.1 Alternative Formulation I

To simplify notation, let $C = \{1,\ldots,m\}$ (instead of $C = \{c_1,\ldots,c_m\}$) and let $\mathscr{M}$ denote the class of nonnegative monotone set functions on C, i.e., the class of functions $v : 2^C \rightarrow [0,\infty)$ such that $v(A) \leq v(B)$ for all $A \subseteq B \subseteq C$; for the time being, we neglect the normalization condition (10), as it is less important for our purpose (it constitutes a single constraint that must be added to the optimization problem in order to turn a monotone measure into a fuzzy measure). More specifically, we are interested in the subclass $\mathscr{M}_2 \subset \mathscr{M}$ of 2-additive measures v, i.e., whose Möbius transform satisfies $m_v(A) = 0$ for all $A \subseteq C$ such that $|A| > 2$.

The following characterization is well-known (see, e.g., Proposition 1 in [16]): $v \in \mathscr{M}_2$ if and only if the following constraints $C_{i,X}$ are satisfied for all $i \in C$ and $X \subseteq C_i = C \setminus \{i\}$:

$$C_{i,X} : \quad m_i + \sum_{j \in X} m_{i,j} \geq 0 \ , \tag{12}$$

where $m_i = m_v(\{i\})$ and $m_{i,j} = m_v(\{i,j\})$. Note that the number of constraints (12) is still exponential in m. Yet, we can show that they can be expressed equivalently in terms of a smaller number of constraints (albeit at the expense of introducing additional variables).

Proposition 1. *Condition (12) is equivalent to the following condition: For all $i \in C$, there exist $\alpha_{i,j}$, $j \in C_i$, such that*

$$\begin{aligned} &\alpha_{i,j} \geq 0 \\ &\sum_{j \in C_i} \alpha_{i,j} \leq 1 \\ &m_i \geq 0 \\ &m_{i,j} \geq -\alpha_{i,j} \cdot m_i \end{aligned} \tag{13}$$

Proof: Let $v \in \mathscr{M}_2$ and suppose (12) to hold. For $i \in C$, (12) with $X = \emptyset$ implies $m_i \geq 0$. Now, define $C_i^- = \{j \in C_i \,|\, m_{i,j} < 0\}$, $C_i^+ = \{j \in C_i \,|\, m_{i,j} \geq 0\}$, and let

$$\alpha_{i,j} = \begin{cases} 0 & \text{if } j \in C_i^+ \\ \frac{|m_{i,j}|}{m_i} & \text{if } j \in C_i^- \end{cases}$$

Since (12) holds with $X = C_i^-$, we have

$$\sum_{j \in C_i^-} |m_{i,j}| \leq m_i \ ,$$

and therefore

$$\sum_{j \in C_i} \alpha_{i,j} = \sum_{j \in C_i^-} \alpha_{i,j} = \sum_{j \in C_i^-} \frac{|m_{i,j}|}{m_i} = \frac{1}{m_i} \sum_{j \in C_i^-} |m_{i,j}| \leq 1.$$

Moreover, $m_{i,j} \geq -\alpha_{i,j} \cdot m_i$ holds by definition, both for $j \in C_i^+$ and $j \in C_i^-$. Thus, condition (13) holds, and hence (12) implies (13).

Now, suppose that (13) holds. Then, $m_i \geq 0$ and for any $\emptyset \neq X \subseteq C_i$,

$$\begin{aligned} m_i + \sum_{j \in X} m_{i,j} &\geq m_i + \sum_{j \in X} -\alpha_{i,j} \cdot m_i \\ &= m_i - m_i \sum_{j \in X} \alpha_{i,j} \\ &\geq m_i(1 - \sum_{j \in X} \alpha_{i,j}) \geq 0 \end{aligned}$$

Thus, condition (12) holds, and hence (13) implies (12). Q.E.D.

As a consequence of the above result, the constraints (11) can be replaced by the equivalent constraints (13). Thus, the number of constraints can indeed be reduced from exponential to quadratic, namely to $2m^2$ inequalities. On the other hand,

(13) also comes with a disadvantage: While the constraints (11) are all linear, some of the constraints (13) are *nonlinear* (albeit convex); indeed, recall that the $\alpha_{i,j}$ are introduced as new variables that need to be determined simultaneously with the m_i and $m_{i,j}$.

4.2 Alternative Formulation II

Our second reformulation of the problem is based on a theoretical result showing that the class $\mathscr{M}_2$ or, more specifically, the class of normalized measures in $\mathscr{M}_2$ (i.e., those v whose Möbius function additionally satisfies (10), forms a convex polytope. The extreme points of this polytope are exactly those $\{0,1\}$-valued measures whose Möbius transforms are of the form

$$m_A(X) = \begin{cases} 1 & \text{if } X = A \\ 0 & \text{otherwise} \end{cases}, \qquad A \in \mathscr{E}$$

or of the form

$$m'_B(X) = \begin{cases} 1 & \text{if } \emptyset \neq X \subsetneq B \\ -1 & \text{if } X = B \\ 0 & \text{otherwise} \end{cases}, \qquad A \in \mathscr{E}',$$

where $\mathscr{E} = \{A \subseteq C \,|\, 1 \leq |A| \leq 2\}$ and $\mathscr{E}' = \{B \subseteq C \,|\, |B| = 2\}$ [17]. In other words, each feasible solution m can be written as a convex combination of these m^2 extreme points:

$$m = \sum_{A \in \mathscr{E}} \alpha_A \cdot m_A + \sum_{B \in \mathscr{E}'} \alpha'_B \cdot m'_B \tag{14}$$

Consequently, the constraints (10–11) can be replaced by (14) in conjunction with the following constraints:

$$\begin{aligned} &\alpha_A \geq 0 \\ &\alpha'_B \geq 0 \\ &\sum_{A \in \mathscr{E}} \alpha_A + \sum_{B \in \mathscr{E}'} \alpha_B = 1 \end{aligned}$$

Like in our first reformulation, the number of constraints is thus significantly reduced, this time even without introducing nonlinearities, albeit again at the cost of a quadratic number of additional variables. More concretely, we end up with m^2 additional variables while reducing the number of constraints to $m^2 + 1$.

5 Experiments

The collection of data for experimental evaluation is a bit hindered by the fact that choquistic regression is a method for learning *monotone models*, i.e., models in which the probability of a positive output is an increasing function of each

Table 1 Data sets and their properties

	data set	#instances	#attributes	source
1	Employee Selection (ESL)	488	4	WEKA
2	Employee Rejection/Acceptance (ERA)	1000	4	WEKA
3	Lecturers Evaluation (LEV)	1000	4	WEKA
4	CPU	209	6	UCI
5	Mammographic (MMG)	961	5	UCI
6	Car Evaluation (CEV)	1728	6	UCI
7	Auto MPG	392	7	UCI
8	Den Bosch (DBS)	120	8	[5]
9	Breast Cancer (BCC)	286	7	UCI
10	Social Workers Decisions (SWD)	1000	10	[6]

Table 2 Classification accuracy in terms of 0/1 loss (mean ± standard deviation derived from 10 repeats of 5-fold cross-validation)

data set	CR-orig	CR-AI	CR-AII	LR
ESL	.0655 ± .0225	.0668 ± .0227	.0639 ± .0208	.0678 ± .0255
ERA	.2908 ± .0312	.2880 ± .0292	.2907 ± .0312	.2873 ± .0275
LEV	.1478 ± .0202	.1491 ± .0222	.1530 ± .0213	.1686 ± .0240
CPU	.0241 ± .0223	.0244 ± .0197	.0196 ± .0236	.0672 ± .0346
MMG	.1685 ± .0240	.1697 ± .0232	.1661 ± .0232	.1712 ± .0268
CEV	.0743 ± .0127	.0835 ± .0120	.0726 ± .0135	.1382 ± .0170
MPG	.0663 ± .0244	.0644 ± .0281	.0636 ± .0254	.0627 ± .0277
DBS	.1413 ± .0715	.1330 ± .0648	.1130 ± .0645	.1472 ± .0573
BCC	.3041 ± .0581	.2840 ± .0556	.3065 ± .0524	.3079 ± .0586
SWD	.2186 ± .0187	.2169 ± .0276	.2143 ± .0225	.2202 ± .0244

input attribute. Data sets for which monotonicity of this kind is a reasonable assumption are less frequent than standard classification data. Nevertheless, we managed to collect 10 such data sets; Table 1 provides a summary of their main properties. Those with a numerical or ordered categorical output were binarized by thresholding at the median. Moreover, all input attributes were normalized.

Experimentally, we compared three versions of choquistic regression, the original formulation from Section 3.1 (CR-orig), the first reformulation from Section 4.1 (CR-AI), and the second reformulation from Section 4.2 (CR-AII). To make the implementations as comparable as possible, we applied the same solver to the different optimization problems, namely the `fmincon` function implemented in the optimization toolbox of Matlab. This function provides a method for constrained nonlinear optimization based on sequential quadratic programming.

In terms of classification accuracy, the different implementations of choquistic regression should perform exactly the same, at least theoretically, because they seek to maximize the same likelihood function under different but equivalent constraints.

Table 3 Runtime complexity of the alternative implementations on different data sets (name, number of attributes, number of instances) measured in terms of CPU time (mean ± standard deviation in seconds) for different sample sizes (in % of the complete data set)

data	CR	20%	40%	60%	80%	100%
ESL	orig	0.26 ± 0.05	0.31 ± 0.02	0.38 ± 0.02	0.45 ± 0.13	0.63 ± 0.05
4	AI	0.41 ± 0.13	0.50 ± 0.07	0.68 ± 0.13	0.80 ± 0.17	1.05 ± 0.18
488	AII	0.31 ± 0.09	0.39 ± 0.07	0.50 ± 0.06	0.61 ± 0.04	0.70 ± 0.04
ERA	orig	0.23 ± 0.03	0.36 ± 0.01	0.50 ± 0.02	0.63 ± 0.01	0.78 ± 0.02
4	AI	0.53 ± 0.10	0.90 ± 0.08	1.06 ± 0.16	1.20 ± 0.20	1.35 ± 0.18
1000	AII	0.31 ± 0.05	0.52 ± 0.07	0.70 ± 0.09	1.12 ± 0.14	1.32 ± 0.16
LEV	orig	0.34 ± 0.04	0.55 ± 0.05	0.71 ± 0.04	0.88 ± 0.07	1.03 ± 0.07
4	AI	0.96 ± 0.23	1.41 ± 0.21	1.84 ± 0.24	2.25 ± 0.18	2.50 ± 0.19
1000	AII	0.49 ± 0.07	0.76 ± 0.05	1.04 ± 0.10	1.68 ± 0.15	1.90 ± 0.14
CPU	orig	0.77 ± 0.18	1.95 ± 3.39	3.37 ± 5.42	6.9 ± 8.97	14.23 ± 11.33
6	AI	1.85 ± 0.22	2.56 ± 0.52	2.79 ± 0.71	3.42 ± 0.18	6.11 ± 2.71
209	AII	0.50 ± 0.31	1.28 ± 0.24	1.33 ± 0.29	1.68 ± 0.56	2.06 ± 0.66
MMG	orig	0.39 ± 0.15	0.56 ± 0.06	0.79 ± 0.12	0.95 ± 0.09	1.07 ± 0.11
6	AI	1.19 ± 0.24	1.77 ± 0.47	2.06 ± 0.61	2.71 ± 1.60	3.24 ± 1.96
961	AII	0.52 ± 0.13	0.83 ± 0.11	1.13 ± 0.10	1.54 ± 0.18	1.78 ± 0.19
CEV	orig	2.45 ± 0.24	3.84 ± 0.38	5.09 ± 0.41	5.79 ± 0.51	6.74 ± 0.41
6	AI	5.36 ± 0.55	7.53 ± 1.00	9.89 ± 0.96	11.93 ± 2.83	13.72 ± 2.56
1728	AII	2.11 ± 0.33	3.68 ± 0.31	5.23 ± 0.52	6.88 ± 0.59	7.88 ± 0.58
MPG	orig	1.83 ± 0.71	2.15 ± 0.62	2.69 ± 0.59	3.18 ± 0.54	3.45 ± 0.65
7	AI	2.58 ± 0.32	2.54 ± 0.66	3.46 ± 0.89	3.84 ± 0.75	4.15 ± 0.92
392	AII	0.61 ± 0.21	0.72 ± 0.12	0.95 ± 0.24	1.02 ± 0.19	1.3 ± 0.13
DBS	orig	5.68 ± 1.11	5.36 ± 1.23	5.61 ± 1.02	5.59 ± 0.72	5.47 ± 1.05
8	AI	2.51 ± 1.81	2.88 ± 1.29	3.03 ± 1.42	3.17 ± 0.96	4.08 ± 1.10
120	AII	0.71 ± 0.19	0.78 ± 0.34	0.76 ± 0.18	0.82 ± 0.12	0.91 ± 0.13
BCC	orig	1.22 ± 0.56	1.10 ± 0.27	1.19 ± 0.23	1.47 ± 0.38	1.47 ± 0.25
9	AI	2.29 ± 1.09	2.04 ± 1.52	2.16 ± 0.95	2.88 ± 2.5	2.97 ± 2.30
286	AII	0.47 ± 0.24	0.47 ± 0.06	0.55 ± 0.55	0.66 ± 0.11	0.78 ± 0.07
SWD	orig	292.4 ± 31.1	382.8 ± 42.24	371.3 ± 12.67	394.0 ± 36.62	427.5 ± 36.62
10	AI	17.9 ± 13.4	27.82 ± 12.13	32.11 ± 10.10	32.35 ± 10.05	33.14 ± 10.77
1000	AII	4.7 ± 0.71	8.80 ± 1.34	13.01 ± 1.44	18.24 ± 2.21	22.66 ± 1.73

Practically, of course, different formulations of the optimization problem will yield slightly different solutions, although these differences should be small. This expectation is confirmed by the result of a 5-fold cross validation, which is summarized in Table 2; this table also shows results for standard logistic regression (LR) as a baseline.

What we are of course most interested in is the runtime performance of the different implementations, which we measured in terms of CPU usage.[1] The results, which are summarized in Table 3, convey a quite clear picture: While the original implementation CR-orig is superior or at least competitive for data sets with up to 6 attributes, it is visibly outperformed by the alternative formulations for $m > 6$ attributes, and the difference in runtime rapidly increases with m. This is in agreement with our expectations: An exponential number of constraints is no big obstacle provided the number of attributes is small. In this case, a reduction from exponential to quadratic does not compensate for the additional overhead caused by introducing new variables. Due to the exponential growth of the number of constraints in CR-orig, however, this situation quickly changes in favor of CR-AI and CR-AII with an increasing number of attributes; indeed, as can be seen from the SWD data, the runtime of CR-orig becomes unacceptable as soon as $m > 9$.

This is also confirmed by another experiment we did with this data set: From the total of 10 attributes, we randomly samples $m \in \{5, 6, \ldots, 10\}$, trained a CR model on the data set reduced to these k attributes (using the tree methods CR-orig, CR-AI and CR-AII) and measured the runtime. This was repeated many times and the runtime was averaged. Fig. 1 shows this average runtime as a function of m.

Comparing the two alternatives CR-AI and CR-AII, it seems that the latter is consistently faster, although the growth of the runtime as a function of m is in both cases much more moderate than for CR-orig. Again, this is not unexpected against the background of the results from the previous section.

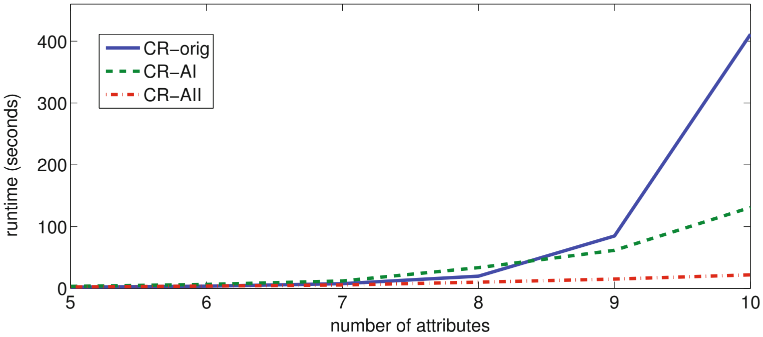

Fig. 1 Average runtime on the SWD data as a function of the number of attributes included

6 Discussion

Our experimental results are in complete agreement with the theoretical complexity (in terms of the number of constraints and the number of variables involved) of the optimization problems. Thus, learning the Choquet integral for classification can indeed be made more efficient by exploiting the special structure of the problem

[1] Experiments were carried out on an Intel Core(TM) i7-2600 CPU with 3.40GHz and 8 GB RAM under Windows 7.

in the case of 2-additive fuzzy measures, essentially reducing the complexity from exponential to quadratic in the number of attributes.

In order to compare the different variants of the problem (CR-orig, CR-AI, CR-AII), we decided to use a rather general optimization method that can handle all of them without the need for specific adaptations. An interesting alternative, of course, is to implement each of the variants individually and as efficiently as possible, seeking for a more specialized solver that allows for exploiting the respective problem structure in an optimal way. In particular, this appears to be important for a more thorough comparison of the two alternatives we proposed, respectively, in Sections 4.1 and 4.2.

Theoretically, CR-AII seems to be advantageous to CR-AI, and indeed, the experimental results are in agreement with this presumption. Nevertheless, the reformulation in Section 4.1 should not be abandoned rashly. First, as just mentioned, it might be possible to improve its efficiency by means of specialized optimization techniques; one may think, for example, of an alternating optimization scheme in which, repeatedly, the $\alpha_{i,j}$ are fixed while the $m_{i,j}$ are optimized and vice versa, thereby circumventing the issue of nonlinearity.

Moreover, CR-AII might be more amenable for a generalization to the case of k-additive measures, $k > 2$. In this regard, the second approach is arguably difficult: Firstly, it is known that for $k > 2$, the extreme points of the convex polytope of k-additive measures are not all $\{0,1\}$-valued. Secondly, and more importantly, the number of these extreme points is expected to grow extremely fast, knowing that the number of extreme points of the polytope of additive measures on m variables grows like the sequence of Dedekind numbers [15].

Acknowledgements. This work was supported by the German Research Foundation (DFG).

References

[1] Angilella, S., Greco, S., Matarazzo, B.: Non-additive robust ordinal regression with Choquet integral, bipolar and level dependent Choquet integrals. In: Proc. IFSA/EUSFLAT 2009, Lisbon, Portugal, pp. 1194–1199 (2009)

[2] Beliakov, G., James, S.: Citation-based journal ranks: the use of fuzzy measures. Fuzzy Sets and Systems 167(1), 101–119 (2011)

[3] Ben-David, A.: Monotonicity maintenance in information-theoretic machine learning algorithms. Machine Learning 19, 29–43 (1995)

[4] Choquet, G.: Theory of capacities. Annales de l'institut Fourier 5, 131–295 (1954)

[5] Daniels, H., Kamp, B.: Applications of MLP networks to bond rating and house pricing. Neural Computation and Applications 8, 226–234 (1999)

[6] David, A.B.: (2010), http://mldata.org/repository/data/viewslug/datasets-arie_ben_david-swd/ (last accessed on June 21, 2012)

[7] Tehrani, A.F., Cheng, W., Dembczy, K., Hüllermeier, E.: Learning Monotone Nonlinear Models Using the Choquet Integral. In: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (eds.) ECML PKDD 2011. LNCS, vol. 6913, pp. 414–429. Springer, Heidelberg (2011)

[8] Fallah Tehrani, A., Cheng, W., Hüllermeier, E.: Choquistic regression: Generalizing logistic regression using the Choquet integral. In: 7th Int. Conf. of the European Society for Fuzzy Logic and Technology, EUSFLAT 2011, Aix-les-Bains, France, pp. 868–875 (2011)

[9] Fallah Tehrani, A., Cheng, W., Hüllermeier, E.: Preference learning using the Choquet integral: The case of multipartite ranking. IEEE Transactions on Fuzzy Systems (forthcoming, 2012)

[10] Feelders, A.: Monotone relabeling in ordinal classification. In: Proc. of the 10th IEEE International Conference on Data Mining, pp. 803–808. IEEE Press, Piscataway (2010)

[11] Grabisch, M.: Fuzzy integral in multicriteria decision making. Fuzzy Sets and Systems 69(3), 279–298 (1995)

[12] Grabisch, M.: Modelling data by the Choquet integral. In: Torra, V. (ed.) Information Fusion in Data Mining, pp. 135–148. Springer, Heidelberg (2003)

[13] Grabisch, M., Nicolas, J.M.: Classification by fuzzy integral: performance and tests. Fuzzy Sets and Systems 65(2-3), 255–271 (1994)

[14] Grabisch, M., Murofushi, T., Sugeno, M.: Fuzzy Measures and Integrals: Theory and Applications. Physica-Verlag, Heidelberg (2000)

[15] Miranda, P., Grabisch, M.: On vertices of the k-additive monotone core. In: Proc. IFSA/EUSFLAT 2009, Lisbon, Portugal, pp. 76–81 (2009)

[16] Miranda, P., Grabisch, M., Gil, P.: Axiomatic structure of k-additive capacities. Mathematical Social Sciences 49, 153–178 (2005)

[17] Miranda, P., Combarro, E.F., Gil, P.: Extreme points of some families of non-additive measures. European J. of Operational Research 174, 1865–1884 (2006)

[18] Modave, F., Grabisch, M.: Preference representation by a Choquet integral: commensurability hypothesis. In: Proc. of the 7th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, pp. 164–171 (1998)

[19] Potharst, R., Feelders, A.: Classification trees for problems with monotonicity constraints. ACM SIGKDD Explorations Newsletter 4(1), 1–10 (2002)

[20] Sugeno, M.: Theory of fuzzy integrals and its application. PhD thesis, Tokyo Institute of Technology, Tokyo, Japan (1974)

[21] Torra, V.: Learning aggregation operators for preference modeling. In: Fürnkranz, J., Hüllermeier, E. (eds.) Preference Learning, pp. 317–333. Springer, Heidelberg (2011)

[22] Torra, V., Narukawa, Y.: Modeling Decisions: Information Fusion and Aggregation Operators. Springer, Heidelberg (2007)

[23] Vitali, G.: Sulla definizione di integrale delle funzioni di una variabile. Annali di Matematica Pura ed Applicata 2(1), 111–121 (1925)