Choquistic Regression: Generalizing Logistic Regression using the Choquet Integral

Ali Fallah Tehrani, Weiwei Cheng, and Eyke Hüllermeier

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

Abstract

In this paper, we propose a generalization of logistic regression based on the Choquet integral. The basic idea of our approach, referred to as choquistic regression, is to replace the linear function of predictor variables, which is commonly used in logistic regression to model the log odds of the positive class, by the Choquet integral. Thus, it becomes 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. In experimental studies with real and benchmark data, choquistic regression consistently improves upon standard logistic regression in terms of predictive accuracy.

Keywords: logistic regression, Choquet integral, monotone classification, attribute interaction

1. Introduction

Logistic regression is a well-established statistical method for (probabilistic) classification [1]. In fact, this method is not only interesting from a statistical point of view, but also quite popular in different application fields, such as medicine, psychology, economics and the social sciences, to name a few. This popularity is due to a number of appealing properties of logistic regression, including the following ones:

- Since the model is essentially *linear* in the input attributes, it is easily comprehensible. In particular, the strength of influence of each predictor is directly reflected by the corresponding regression coefficient.
- The influence of each attribute is *monotone* in the sense that an increase of the value of the attribute can only increase (decrease) the probability of the positive class.

Both of the above points, comprehensibility and monotonicity, are important prerequisites for the acceptance of a model by a domain expert. Indeed, in many cases, monotonicity is a very natural requirement. In a medical context, for example, one will expect that tobacco consumption will increase the probability of cancer, and each model violating this constraint will not be considered as faithful and hence be refused by a medical doctor.

The learning of monotone models from data has attracted considerable attention in the machine learning field in recent years [2]. Interestingly, monotonicity is not easily guaranteed for a number of well-known classification methods like, for example, decision trees. Thus, for a decision tree it may easily happen that, depending on the values of the remaining attributes, increasing the value of an attribute (e.g., tobacco consumption) may change the class prediction from positive to negative in one case (e.g., one patient) and from negative to positive in another case (e.g., another patient).

Coming back to logistic regression, the linearity of the model is of course a strong restriction from a learning point of view. Quite often, the response variable (output) depends on the predictor variables (inputs) in a *nonlinear* way. In this paper, we therefore propose an extension of logistic regression that allows for modeling nonlinear relationships between input and output variables while preserving the aforementioned advantages of the approach, namely comprehensibility and monotonicity. Roughly speaking, the basic idea of our approach is to replace the linear function in the logistic regression model by the Choquet integral.

The rest of this paper is organized as follows. In the next section, we give a brief overview of related work. In Section 3, we recall the basic definition of the Choquet integral and some related notions. Logistic regression in its standard form is briefly covered in Section 4, prior to the introduction of our generalized approach in Section 5. Finally, experimental results are presented in Section 6.

2. Related Work

As mentioned earlier, logistic regression is a well-established method in statistics and machine learning, and there is a wealth of literature on this method. Worth mentioning here is that the possibility of interactions between predictor variables has of course also been noticed in the statistical literature [3]. A standard way to handle such interaction effects is to add interaction terms to the linear function of predictor variables, for example in the form of products of pairs of predictors. Thus, however, the aforementioned advantages of logistic

regression are partly lost.

Although the Choquet integral has been widely applied as an aggregation operator in multiple criteria decision making [4–6], it has been used much less in the field of machine learning so far. There are, however, a few notable exceptions. First, the problem of extracting a Choquet integral (or, more precisely, the non-additive measure on which it is defined) in a data-driven way has been addressed in the literature. Essentially, this is a parameter identification problem, which is commonly formalized as a constraint optimization problem, for example using the sum of squared errors as an objective function [7,8]. To this end, [9] proposed an approach based on the use of quadratic forms, while an alternative heuristic, gradient-based method called HLMS (Heuristic Least Mean Squares) was introduced in [10].

Moreover, the Choquet integral has been used for learning classification models. Recently, for example, it has been used for ordinal classification [11, 12]. In [13], the problem of learning an optimal classification function is cast in the setting of margin-maximization. Besides, the Choquet integral has been used as an aggregation operator in the context of ensemble learning, i.e., for combining the predictions of different classifiers [14].

As already mentioned, the problem of monotone classification has received increasing attention in the machine learning community in recent years (despite having been introduced in the literature much earlier [2]). Meanwhile, several machine learning algorithms have been modified so as to guarantee monotonicity in attributes, including nearest neighbor classification [15], decision tree learning [16] and rule induction [17]. Instead of modifying models and algorithms, one can also modify the data. To this end, data pre-processing methods such as re-labeling techniques have been developed. Such methods seek to repair inconsistencies in the training data, so that (standard) classifiers learned on that data will automatically be monotone [18].

3. The Discrete Choquet Integral

In this section, we recall the basic definition of the Choquet integral and related notions. The first definition of the Choquet integral for additive measures is due to Vitali [19]. For the general case of a capacity (i.e., a non-additive measure or fuzzy measure), it was later on introduced by Choquet [20]. Yager proposed a generalized version in [21].

Definition 1 (Fuzzy measure) *Let $C = \{c_1, c_2, \ldots, c_m\}$ be a finite set. A discrete fuzzy measure (also called capacity) is a set function $\mu : 2^C \to [0,1]$ which is monotonic ($\mu(A) \leq \mu(B)$ for $A \subseteq B \subseteq C$) and normalized ($\mu(\emptyset) = 0$ and $\mu(C) = 1$). A fuzzy measure μ is called additive if $\mu(A \cup B) = \mu(A) + \mu(B)$ for all $A, B \subset C$ such that $A \cap B = \emptyset$. Obviously, in the case of an additive measure, $\mu(A)$ is simply obtained as follows:*

$$\mu(A) = \sum_{c_i \in A} \mu(\{c_i\}) \; . \tag{1}$$

Definition 2 (Choquet integral) *Let μ be a fuzzy measure on $C = \{c_1, c_2, \ldots, c_m\}$. The discrete Choquet integral of a function $f : C \to \mathbb{R}_+$ with respect to μ is defined as follows:*

$$\mathcal{C}_\mu(f) = \sum_{i=1}^{m} \big(f(c_{(i)}) - f(c_{(i-1)})\big) \cdot \mu(A_{(i)}) \; ,$$

where $(\cdot)$ is a permutation of $\{1, \ldots, m\}$ such that $0 \leq f(c_{(1)}) \leq f(c_{(2)}) \leq \ldots \leq f(c_{(m)})$. Moreover, $A_{(i)}$ is given by the set $\{c_{(i)}, \ldots, c_{(m)}\}$. Finally, $f(c_{(0)}) = 0$ by definition.

Definition 3 (Möbious transform) *The Möbius transform $\boldsymbol{m}_\mu$ of a fuzzy measure μ is defined as follows:*

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

for all $A \subseteq C$.

As a useful property of the Möbius transform, which we shall exploit later on for learning Choquet integrals, we mention that it allows for reconstructing the underlying fuzzy measure:

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

for all $B \subseteq C$. More specifically, we shall make use of the following representation of the Choquet integral:

$$\begin{aligned}
\mathcal{C}_\mu(f) &= \sum_{i=1}^{m} \big(f(c_{(i)}) - f(c_{(i-1)})\big) \cdot \mu(A_{(i)}) \\
&= \sum_{i=1}^{m} f(c_{(i)})(\mu(A_{(i)}) - \mu(A_{(i+1)})) \\
&= \sum_{i=1}^{m} f(c_{(i)}) \sum_{R \subseteq T_{(i)}} \boldsymbol{m}(R) \\
&= \sum_{T \subseteq C} \boldsymbol{m}(T) \times \min_{c_i \in T} f(c_i) \; ,
\end{aligned} \tag{2}$$

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

Definition 4 (k-Additivity) *A fuzzy measure μ is said to be k-order additive or simply k-additive if k is the smallest integer such that $\boldsymbol{m}(A) = 0$ for all $A \subseteq C$ with $|A| > k$.*

Thus, while a Choquet integral is determined by 2^m coefficients in general, the k-additivity of the underlying measure reduces the number of required coefficients to at most

$$\sum_{i=1}^{k} \left(\begin{array}{c} m \\ i \end{array} \right) \; .$$

3.1. Importance of Criteria and Interaction

The (discrete) Choquet integral is often used as an aggregation operator, namely to aggregate the assessments $f(c_i)$ of an object on different criteria c_i into a single evaluation. If the underlying measure μ is additive (i.e., k-additive with $k = 1$), the Choquet integral reduces to a weighted mean

$$\mathcal{C}_\mu(f) = \sum_{i=1}^{m} w_i \cdot f(c_i) \;, \tag{3}$$

with $w_i = \mu(\{c_i\})$ the weight or, say, the importance of the criterion c_i. These weights are non-negative and such that $\sum_{i=1}^{m} w_i = 1$. In this case, there is obviously no interaction between the criteria c_i, i.e., the influence of evaluation $f(c_i)$ on the overall assessment is independent of the other values $f(c_j)$, $j \neq i$.

Measuring the importance of a criterion c_i becomes obviously more involved if μ is non-additive. Besides, one may then also be interested in a measure of *interaction* between the criteria, either pairwise or even of a higher order. In the literature, measures of that kind have been proposed, both for the importance of single as well as the interaction between several criteria [22].

Given a fuzzy measure μ on C, the *Shaply value* (or importance index) of c_i is defined as follows:

$$\varphi(c_i) = \sum_{A \subseteq C \setminus \{c_i\}} \frac{1}{m \binom{m-1}{|A|}} \left(\mu(A \cup \{c_i\}) - \mu(A)\right).$$

The Shaply value of μ is the vector $\varphi(\mu) = (\varphi(1), \ldots, \varphi(m))$. One can show that $0 \leq \varphi(c_i) \leq 1$ and $\sum_{i=1}^{m} \varphi(c_i) = 1$. Thus, $\varphi(c_i)$ is a measure of the *relative* importance of c_i. Obviously, $\varphi(c_i) = \mu(\{c_i\})$ if μ is additive.

The *interaction index* between criteria c_i and c_j, as proposed by Murofushi and Soneda [23], is defined as follows:

$$I_{i,j} = \sum_{A \subseteq C \setminus \{c_i, c_j\}} \vartheta_A \cdot \Big(\mu(A \cup \{c_i, c_j\}) - \mu(A \cup \{c_i\})) - \mu(A \cup \{c_j\}) + \mu(A)\Big)$$

with

$$\vartheta_A = \frac{1}{(m-1)\binom{m-2}{|A|}} \;.$$

This index ranges between -1 and 1 and indicates a positive (negative) interaction between criteria c_i and c_j if $I_{i,j} > 0$ ($I_{i,j} < 0$).

The interaction index can also be expressed in terms of the Möbius transform:

$$I_{i,j} = \sum_{K \subseteq C \setminus \{c_i, c_j\}, |K| = k} \frac{1}{k+1} \boldsymbol{m}\left(\{c_i, c_j\} \cup K\right) \;.$$

Furthermore, as proposed by Grabisch [24], the definition of interaction can be extended to more than two features, i.e., to feature subsets $T \subseteq C$:

$$I_T = \sum_{k=0}^{m-|T|} \frac{1}{k+1} \sum_{K \subseteq C \setminus T, |K| = k} \boldsymbol{m}(T \cup K) \;.$$

4. Background on Logistic Regression

4.1. Classification

We consider the problem of classification, that is, predicting the value of an output (response) variable $y \in \mathcal{Y}$ given the values of a set of input attributes (predictors) $x_i \in \mathcal{X}_i$, $i = 1, \ldots, m$. The vector

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

is called an *instance*, and $\mathcal{X}$ the *instance space*. We restrict ourselves to binary classification, which means that $\mathcal{Y} = \{0, 1\}$ consists of only two classes, typically called the negative (0) and the positive (1) class. The goal is to learn a classifier $\mathcal{L} : \mathcal{X} \to \mathcal{Y}$ from a given set of training data

$$\mathcal{D} = \left\{(\boldsymbol{x}^{(i)}, y^{(i)})\right\}_{i=1}^{n} \subset (\mathcal{X} \times \mathcal{Y})^n.$$

The data $\mathcal{D}$ is supposed to be an *i.i.d.* (independent and identically distributed) sample generated by an underlying (though unknown) probability measure $\mathbf{P}_{XY}$ on $\mathcal{X} \times \mathcal{Y}$. A common goal, then, is to induce a classifier with minimal risk, where the risk $R(\mathcal{L})$ of a classifier $\mathcal{L}$ is defined as the expected loss:

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

where $\ell(\cdot)$ is a loss function penalizing incorrect predictions. In binary classification, the most commonly used loss is the simple 0/1 loss given by $\ell(\hat{y}, y) = 0$ if $\hat{y} = y$ and $= 1$ if $\hat{y} \neq y$.

4.2. Logistic Regression

Logistic regression modifies linear regression for the purpose of predicting (probabilities of) discrete classes instead of real-valued responses. To this end, the probability of the positive class (and hence of the negative class) is modeled as a linear function of the input attributes. More specifically, since a linear function does not necessarily produce values in the unit interval, the response is defined as a generalized linear model, namely in terms of the logarithm of the probability ratio:

$$\log\left(\frac{\mathbf{P}(y = 1 \,|\, \boldsymbol{x})}{\mathbf{P}(y = 0 \,|\, \boldsymbol{x})}\right) = w_0 + \boldsymbol{w}^\top \boldsymbol{x} \;, \tag{4}$$

where $\boldsymbol{w} = (w_1, w_2, \ldots, w_m)^\top \in \mathbb{R}^m$ is a vector of regression coefficients and $w_0 \in \mathbb{R}$ a constant bias (the intercept). A positive regression coefficient $w_i > 0$ means that an increase of the predictor

variable x_i will increase the probability of a positive response, while a negative coefficient implies a decrease of this probability. Besides, the larger the absolute value $|w_i|$ of the regression coefficient, the stronger the influence of x_i.

Since $\mathbf{P}(y = 0 \,|\, \boldsymbol{x}) = 1 - \mathbf{P}(y = 1 \,|\, \boldsymbol{x})$, a simple calculation yields the posterior probability

$$\pi_l \stackrel{\text{df}}{=} \mathbf{P}(y = 1 \,|\, \boldsymbol{x}) = \frac{1}{1 + \exp(-w_0 - \boldsymbol{w}^\top \boldsymbol{x})} \,. \quad (5)$$

The logistic function $z \mapsto \frac{1}{1+\exp(-z)}$, which has a sigmoidal shape, is a specific type of *link function*.

5. Generalized Logistic Regression using the Choquet Integral

5.1. The Choquistic Model

In order to model non-linear dependencies between predictor variables and response, and to take interactions between predictors into account, we propose to extend the logistic regression model by replacing the linear function $\boldsymbol{x} \mapsto w_0 + \boldsymbol{w}^\top \boldsymbol{x}$ in (4) by the Choquet integral. More specifically, we propose the following model

$$\begin{aligned} \pi_c &\stackrel{\text{df}}{=} \mathbf{P}(y = 1 \,|\, \boldsymbol{x}) \\ &= \frac{1}{1 + \exp(-\gamma(\mathcal{C}_\mu(f_{\boldsymbol{x}}) - \beta))} \,, \end{aligned} \quad (6)$$

where $\mathcal{C}_\mu(f_{\boldsymbol{x}})$ is the Choquet integral (with respect to the measure μ) of the evaluation function $f_{\boldsymbol{x}} : \{c_1, \ldots, c_m\} \to [0, 1]$ that maps each attribute c_i to a value $x_i = f_{\boldsymbol{x}}(c_i)$; $\beta, \gamma \in \mathbb{R}$ are constants.

The value of c_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 [25]. A simple transformation, that we shall also employ in our experimental studies, is given by the mapping $z_i = (x_i - m_i)/(M_i - m_i)$, where m_i and M_i are lower and upper bounds for x_i (perhaps estimated from the data); if the influence of x_i is actually negative (i.e., $w_i < 0$), then the mapping $z_i = (M_i - x_i)/(M_i - m_i)$ is used instead.

In order to see that our model (6) is a proper generalization of standard logistic regression, recall that the Choquet integral reduces to a weighted mean (3) in the special case of an additive measure μ. Moreover, consider any linear function $\boldsymbol{x} \mapsto g(\boldsymbol{x}) = w_0 + \boldsymbol{w}^\top \boldsymbol{x}$ with $\boldsymbol{w} = (w_1, \ldots, w_m)^\top$.

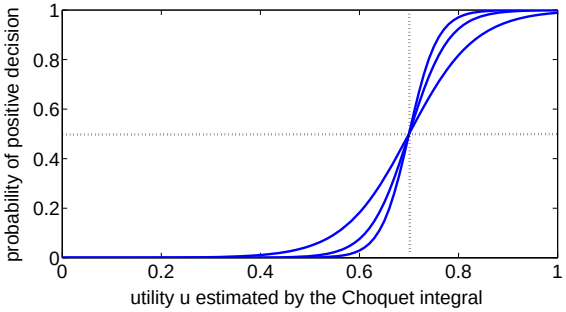

Figure 1: Probability of a positive decision, $\mathbf{P}(y = 1 \,|\, \boldsymbol{x})$, as a function of the estimated degree of utility, $u = U(\boldsymbol{x})$, for a threshold $\beta = 0.7$ and different values of γ.

This function can also be written in the form

$$\begin{aligned} g(\boldsymbol{x}) &= w_0 + \sum_{i=1}^{m} (w_i p_i + |w_i|(M_i - m_i) z_i) \\ &= w_0 + \sum_{i=1}^{m} w_i p_i + \sum_{i=1}^{m} |w_i|(M_i - m_i) z_i \\ &= w_0' + \left(\sum_{i=1}^{m} u_i \right)^{-1} \sum_{i=1}^{m} u_i' z_i \\ &= \gamma \left(\sum_{i=1}^{m} u_i' z_i - \beta \right), \end{aligned}$$

where $p_i = m_i$ if $w_i \geq 0$ and $p_i = M_i$ if $w_i < 0$, $u_i = |w_i|(M_i - m_i)$, $\gamma = \left(\sum_{i=1}^{m} u_i\right)^{-1}$, $u_i' = u_i/\gamma$, $w_0' = w_0 + \sum_{i=1}^{m} w_i p_i$, $\beta = -w_0'/\gamma$. By definition, the u_i' are non-negative and sum up to 1, which means that $\sum_{i=1}^{m} u_i' z_i$ is a weighted mean of the z_i that can be represented by a Choquet integral.

The model (6) can be seen as a two-step process: The first step consists of an assessment of the input $\boldsymbol{x}$ in terms of a utility degree

$$u = U(\boldsymbol{x}) = \mathcal{C}_\mu(f_{\boldsymbol{x}}) \in [0, 1].$$

Then, in a second step, a discrete choice (yes/no decision) is made on the basis of this utility. Roughly speaking, this is done through a "probabilistic thresholding" at the utility threshold β. If $U(\boldsymbol{x}) > \beta$, then the decision tends to be positive, whereas if $U(\boldsymbol{x}) < \beta$, it tends to be negative. The precision of this decision is determined by the parameter γ (see Fig. 1): For large γ, the decision function converges toward the step function $u \mapsto \mathbb{I}(u > \beta)$, jumping from 0 to 1 at β. For small γ, this function is smooth, and there is a certain probability to violate the threshold rule $u \mapsto \mathbb{I}(u > \beta)$. This might be due to the fact that, despite being important for decision making, some properties of the instances to be classified are not captured by the utility function. In that case, the utility $U(\boldsymbol{x})$, estimated on the basis of the given attributes, is not a perfect predictor for the decision eventually made. Thus, the parameter γ can also be seen as an indicator of the quality of the classification model.

5.2. Parameter Estimation

The model (6) has several degrees of freedom: The fuzzy measure μ (Möbius transform $\boldsymbol{m} = \boldsymbol{m}_\mu$) determines the (latent) utility function, while the utility threshold β and the scaling parameter γ determine the discrete choice model. The goal of learning is to identify these degrees of freedom on the basis of the training data $\mathcal{D}$. 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

$$\begin{aligned} l(\boldsymbol{m}, \gamma, \beta) &= \log \mathbf{P}(\mathcal{D} \mid \boldsymbol{m}, \beta, \gamma) \\ &= \log \left(\prod_{i=1}^{n} \mathbf{P}(y^{(i)} \mid \boldsymbol{x}^{(i)}; \boldsymbol{m}, \beta, \gamma) \right) \qquad (7) \\ &= \sum_{i=1}^{n} y^{(i)} \log \pi_c^{(i)} + \left(1 - y^{(i)}\right) \log \left(1 - \pi_c^{(i)}\right). \end{aligned}$$

One easily verifies that (7) is convex with respect to $\boldsymbol{m}, \gamma$, and β. In principle, maximization of the log-likelihood can hence be accomplished by means of standard gradient-based optimization methods. However, since we have to assure that μ is a proper fuzzy measure and, hence, that $\boldsymbol{m}$ guarantees the corresponding monotonicity and boundary conditions, we actually need to solve a *constrained* optimization problem (let $C = \{c_1, \ldots, c_m\}$ denote the set of predictor variables):

$$\begin{aligned} \max_{\boldsymbol{m}, \gamma, \beta} \Big\{ & -\gamma \sum_{i=1}^{n} (1 - y^{(i)}) \left(\mathcal{C}_{\boldsymbol{m}}(\boldsymbol{x}^{(i)}) - \beta\right) \\ & - \sum_{i=1}^{n} \log \Big(1 + \exp(-\gamma \left(\mathcal{C}_{\boldsymbol{m}}(\boldsymbol{x}^{(i)}) - \beta\right)) \Big) \Big\} \end{aligned}$$

such that

$$\begin{aligned} & 0 \leq \beta \leq 1 \\ & 0 < \gamma \\ & \sum_{T \subseteq C} \boldsymbol{m}(T) = 1 \\ & \sum_{B \subseteq A \setminus \{c_i\}} \boldsymbol{m}(B \cup \{c_i\}) \geq 0 \quad \forall A \subseteq C, \forall c_i \in C. \end{aligned}$$

A solution to this problem can be produced by standard solvers. Concretely, we used the `fmincon` function implemented in the optimization toolbox of Matlab. This method is based on a sequential quadratic programming approach.

6. Experimental Results

Experimentally, we compared our generalized variant (6) to the standard version (4) of logistic regression on several benchmark data sets. As mentioned earlier, standard logistic regression (LR) can be seen as a special case of choquistic regression (CR), which is obtained by restricting the fuzzy measure μ to an additive measure. What we expect, therefore, is an improved predictive accuracy thanks to the increased flexibility of choquistic regression, namely its ability to capture nonlinear dependencies between input attributes. It should be noted, however, that such an improvement, despite being plausible, is not self-evident. In fact, if the true underlying dependency is indeed a linear one, at least approximately, then standard logistic regression will be the model of choice, whereas choquistic regression may tend to overfit the training data and hence generalize worse.

6.1. Data

Even though the topic is receiving more and more interest in the machine learning community, benchmark data for monotone classification is not as abundant as for conventional classification. In our experiments, we used six benchmark data sets from the UCI repository[1] and the WEKA machine learning toolbox[2] plus real data that has been extracted from an industrial polyester dyeing process [26]. In what follows, we give a brief description of each of these data sets.

Employee Selection (ESL) This data set contains profiles of applicants for certain industrial jobs. The values of the four input attributes were determined by psychologists based upon psychometric test results and interviews with the candidates. The output is an overall score on an ordinal scale between 1 and 9, corresponding to the degree of suitability of each candidate to this type of job. We binarized the output value by distinguishing between suitable (score 5-9) and unsuitable (score 1-4) candidates.

Employee Rejection/Acceptance (ERA) This data set originates from an academic decision-making experiment. The input attributes are features of a candidate such as past experience, verbal skills, etc., and the output is the subjective judgment of a decision-maker, measured on an ordinal scale from 1 to 9, to which degree he or she tends to accept the applicant for the job. We binarized the output value by distinguishing between acceptance (score 4-9) and rejection (score 1-3).

Lecturers Evaluation (LEV) This data set contains examples of anonymous lecturer evaluations, taken at the end of MBA courses. Students were asked to score their lecturers according to four attributes such as oral skills and contribution to their professional/general knowledge. The output was a total evaluation of each lecturer's performance,

[1] `http://archive.ics.uci.edu/ml/`

[2] `http://www.cs.waikato.ac.nz/ml/weka/`

data set	#inst.	#attr.
CYD 1–7	120	3
DBS	120	8
CPU	209	6
ESL	488	4
ERA	1000	4
LEV	1000	4
CEV	1728	6

Table 1: Data sets and their properties.

measured on an ordinal scale from 0 to 4. We binarized the output value by distinguishing between good (score 3-4) and bad evaluation (score 0-2).

Den Bosch (DBS) This data set contains 8 attributes describing houses in the city of Den Bosch: area, number of bedrooms, type of house, volume, storeys, type of garden, garage, and price. The output is a binary variable indicating whether the price of the house is low or high (depending whether or not it exceeds a threshold).

CPU This is a standard benchmark data set from the UCI repository. It contains nine attributes, three of which were removed since they are obviously of no predictive value (vendor name, model name, ERP).

Car Evaluation (CEV) This data set contains 6 attributes describing a car, namely, buying price, price of the maintenance, number of doors, capacity in terms of persons to carry, the size of luggage boot, estimated safety of the car. The output is the overall evaluation of the car: unacceptable, acceptable, good, very good. We binarized this evaluation into (un)acceptable versus (very) good.

Color Yield (CYD) Finally, we took data from an industrial polyester dyeing process that was also analyzed in [26]. Here, the output variable is the color yield, which has been measured as a function of three important factors: disperse dyes concentration, temperature and time of dyeing. Corresponding experiments have been made for seven different colors, giving rise to seven data sets. Each of these data sets was binarized by thresholding the color yield at its median value.

An overview of the data sets together with their main properties is given in Table 1.

6.2. Results

Classification accuracy was measured in terms of 0/1 loss and determined by randomly splitting the data into two parts, one half for training and one half for testing. To prevent over-fitting we restrict the choquistic model to k-additive measures, where k is select by 10-fold cross validation on the training set. This was repeated 100 times, and the accuracy degrees were averaged. The results of the experiments are summarized in Table 2. As can be seen, CR achieves a consistent improvement and outperforms LR on all data sets.

dats set	LR	CR
ESL	0.0621 ± 0.0096	0.0547 ± 0.0105
ERA	0.2849 ± 0.0140	0.2756 ± 0.0170
LEV	0.1669 ± 0.0134	0.1340 ± 0.0115
DBS	0.1443 ± 0.0371	0.1560 ± 0.0405
CPU	0.0400 ± 0.0093	0.0119 ± 0.0138
CEV	0.1883 ± 0.0066	0.0346 ± 0.0076
CYD-1	0.1254 ± 0.0074	0.0729 ± 0.0066
CYD-2	0.2004 ± 0.0091	0.0717 ± 0.0078
CYD-3	0.1512 ± 0.0238	0.0762 ± 0.0163
CYD-4	0.1289 ± 0.0253	0.0496 ± 0.0201
CYD-5	0.1242 ± 0.0099	0.0204 ± 0.0057
CYD-6	0.1604 ± 0.0085	0.0383 ± 0.0083
CYD-7	0.1958 ± 0.0207	0.0646 ± 0.0089

Table 2: Classification performance in terms of the mean and standard deviation of 0/1 loss.

As mentioned before, as one of its key features, the Choquet integral offers interesting information about the importance of individual attributes as well as the interaction between them. In fact, in many practical applications, this type of information is at least as important as the predictive accuracy of the model. A detailed analysis of this type of information is beyond the scope of this paper. Let us mention, however, an interesting albeit plausible observation we made during our experiments, namely a positive correlation between the degree of attribute interaction and the improvement achieved by CR as compared to LR in terms of accuracy.

As an illustration, Fig. 2 provides a visual representation of the interaction between the three attributes in the color yield data sets, namely for CYD-1 and CYD-7. Degrees of interaction are shown as levels of gray, which means that light and dark fields strongly silhouetted against the color of the diagonal indicate a high degree of interaction. Obviously, the interaction is not very strong in the case of CYD-1, but more pronounced for CYD-7. This is in agreement with the improvement in terms of accuracy, which is much higher in the latter case.

7. Summary and Conclusions

In this paper, we have advocated the use of the discrete Choquet integral in the context of binary classification with monotonicity constraints. More specifically, we have used the Choquet integral for representing a latent utility function in the logistic regression model. Thus, it becomes possible to capture nonlinear dependencies and interactions among predictor variables in a convenient way.

Admittedly, as mentioned above, the same can in

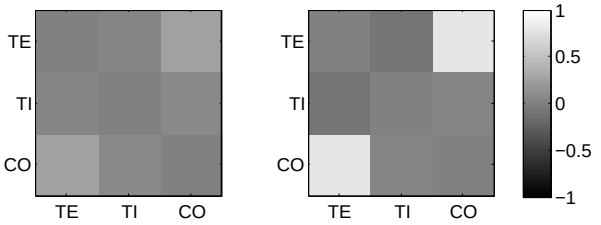

Figure 2: Visualization of the interaction index for data sets CYD-1 (left) and CYD-7 (right). Comparing CR and LR in terms of 0/1 loss, the reduction of prediction error is about twice as much as for CYD-7. For the ease of representation, the values on the diagonal are set to 0.

principle be achieved by any other type of nonlinear function, such as polynomials. Then, however, some important properties of logistic regression may get lost, including the comprehensibility of the model. The Choquet integral is especially appealing from this point of view, as it offers measures of the importance of individual and the interaction among subsets of attributes. Besides, the Choquet integral can ensure a monotone dependency between the output and the individual input attributes, which may become difficult for other nonlinear functions.

An interesting question to be addressed in future work concerns the restriction of the choquistic model to k-additive measures. A restriction of that kind may have two important advantages: First, it may prevent from over-fitting the data in cases where the full flexibility of the Choquet integral is actually not needed. Second, since less parameters need to be identified, the computational complexity will be reduced, too. The key problem is how to select a suitable k in an efficient way (without simply trying all values).

Apart from that, our current approach could be generalized in other directions, first from the dichotomous (binary) case to polytomous (multiclass) classification, and second to more general types of monotonicity. In fact, not all variables are monotone in a strict sense, but instead have ideal values somewhere in the middle of their domain. This type of preference can be modeled conveniently in terms of fuzzy sets, and monotone models can then be learned on membership degrees as predictor variables.

Acknowledgments: This work was supported by the German Research Foundation (DFG). We thank Maryam Nasiri for providing us the color yield data.

References

[1] David Hosmer and Stanley Lemeshow. *Applied Logistic Regression*. Wiley, 2nd edition, 2000.

[2] Arie Ben-David, Leon Sterling, and Yoh-Han Pao. Learning and classification of monotonic ordinal concepts. *Computational Intelligence*, 5(1):45–49, 1989.

[3] James Jaccard. *Interaction Effects in Logistic Regression*. Saga University Papers Series on Quantitative Applications in the Social Sciences, 07-135. Saga Publications, 2001.

[4] Michel Grabisch, Toshiaki Murofushi, and Michio Sugeno, editors. *Fuzzy Measures and Integrals: Theory and Applications*. Physica, 2000.

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

[6] Vicenç Torra. Learning aggregation operators for preference modeling. In Johannes Fürnkranz and Eyke Hüllermeier, editors, *Preference Learning*, pages 317–333. Springer, 2011.

[7] Vicenç Torra and Yasuo Narukawa. *Modeling Decisions: Information Fusion and Aggregation Operators*. Springer, 2007.

[8] Michel Grabisch. Modelling data by the Choquet integral. In Vicenç Torra, editor, *Information Fusion in Data Mining*, pages 135–148. Springer, 2003.

[9] Takehiro Mori and Toshiaki Murofushi. An analysis of evaluation model using fuzzy measure and the Choquet integral. In *Proceedings of the 5th Fuzzy System Symposium*, pages 207–212. Japan Society for Fuzzy Sets and Systems, 1989.

[10] Michel Grabisch. A new algorithm for identifying fuzzy measures and its application to pattern recognition. In *Proceedings of IEEE International Conference on Fuzzy Systems*, volume 1, pages 145–150. IEEE, 1995.

[11] Silvia Angilella, Salvatore Greco, and Benedetto Matarazzo. Non-additive robust ordinal regression with Choquet integral, bipolar and level dependent Choquet integrals. In Joao Paulo Carvalho, Didier Dubois, Uzay Kaymak, and Joao Miguel 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.

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

[13] Silvia Angilella, Salvatore Greco, and Benedetto Matarazzo. The most representative utility function for non-additive robust ordinal regression. In Eyke Hüllermeier, Rudolf Kruse, and Frank Hoffmann, editors, *Computational Intelligence for Knowledge-Based Systems Design*, volume 6178 of *Lecture Notes in Computer Science*, pages 220–229. Springer, 2010.

[14] Michel Grabisch and Jean-Marie Nicolas. Classification by fuzzy integral: performance and

tests. *Fuzzy Sets and Systems*, 65(2-3):255–271, 1994.

[15] Wouter Duivesteijn and Ad Feelders. Nearest neighbour classification with monotonicity constraints. In Walter Daelemans, Bart Goethals, and Katharina Morik, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 5211 of *Lecture Notes in Computer Science*, pages 301–316. Springer, 2008.

[16] Rob Potharst and Ad Feelders. Classification trees for problems with monotonicity constraints. *ACM SIGKDD Explorations Newsletter*, 4(1):1–10, 2002.

[17] Krzysztof Dembczyński, Wojciech Kotlowski, and Roman Slowinski. Learning rule ensembles for ordinal classification with monotonicity constraints. *Fundamenta Informaticae*, 94(2):163–178, 2009.

[18] Ad Feelders. Monotone relabeling in ordinal classification. In Geoffrey Webb, Bing Liu, Chengqi Zhang, Dimitrios Gunopulos, and Xindong Wu, editors, *Proceedings of the 10th IEEE International Conference on Data Mining*, pages 803–808. IEEE Computer Society, 2010.

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

[20] Gustave Choquet. Theory of capacities. *Annales de l'institut Fourier*, 5:131–295, 1954.

[21] Ronald Yager. Generalized OWA aggregation operators. *Fuzzy Optimization and Decision Making*, 3(1):93–107, 2004.

[22] Michel Grabisch. The representation of importance and interaction of features by fuzzy measures. *Pattern Recognition Letters*, 17(6):567–575, 1996.

[23] Toshiaki Murofushi and S. Soneda. Techniques for reading fuzzy measures (III): interaction index. In *Proceedings of the 9th Fuzzy Systems Symposium*, pages 693–696, 1993.

[24] Michel Grabisch. k-order additive discrete fuzzy measures and their representation. *Fuzzy Sets and Systems*, 92(2):167–189, 1997.

[25] Francois Modave and Michel Grabisch. Preference representation by a Choquet integral: commensurability hypothesis. In Bernadette Bouchon-Meunier and Ronald Yager, editors, *Proceedings of the 7th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 164–171. Editions EDK, 1998.

[26] Maryam Nasiri. Fuzzy regression modeling of colour yield in dyeing polyester with disperse dyes. Master's thesis, Textile Engineering Department, Isfahan University of Technology, 2003.