KYBERNETIKA — VOLUME *51* (2015), NUMBER 3, PAGES 508–524

GENERALIZATIONS OF THE NOISY-OR MODEL

Jiří Vomlel

In this paper, we generalize the noisy-or model. The generalizations are three-fold. First, we allow parents to be multivalued ordinal variables. Second, parents can have both positive and negative influences on their common child. Third, we describe how the suggested generalization can be extended to multivalued child variables. The major advantage of our generalizations is that they require only one parameter per parent. We suggest a model learning method and report results of experiments on the Reuters text classification data. The generalized noisy-or models achieve equal or better performance than the standard noisy-or. An important property of the noisy-or model and of its generalizations suggested in this paper is that it allows more efficient exact inference than logistic regression models do.

Keywords: Bayesian networks, noisy-or model, classification, generalized linear models

Classification: 68T37, 68T30

1. INTRODUCTION

Conditional probability tables (CPTs) that are the basic building blocks of Bayesian networks [9, 14] have, in general, an exponential size with respect to the number of parent variables of the CPT. This has two unpleasant consequences. First, when eliciting model parameters, one needs to estimate an exponential number of parameters. Second, in a case where there is a high number of parent variables, the exact probabilistic inference may become intractable.

On the other hand, real implementations of Bayesian networks (see, e. g., [12]) often have a simple local structure of the CPTs. The noisy-or model [14] is a popular model for describing relations between variables in one CPT of a Bayesian network. Noisy-or is a member of the family of models of independence of causal influence [7], which are also called canonical models [5]. An advantage of these models is that the number of parameters required for their specification is linear with respect to the number of parent variables in CPTs and that they allow applications of efficient inference methods, see, for example, [6, 19]. In [24], Zagorecki and Druzdzel show that practical models, for which the authors do not take noisy-or (or noisy-max) models, even models learned from data, have many CPTs that can be approximated by a noisy-or (noisy-max) model. Additionally, the results presented in [23] suggest that many CPTs from real applications can be parameterized with the aid of a low number of parameters.

DOI: 10.14736/kyb-2015-3-0508

In some applications it is natural to consider multivalued parent variables with values having a natural ordering. In this paper, we propose a generalization of the noisy-or model to multivalued ordinal parent variables. Our first and second proposals differ from the noisy-max model [8] since we keep the child variable binary, no matter what the number of values of the parent variables are. Also, we have only one parameter for each parent. Our generalizations also differ from the generalization of the noisy-or model proposed by Srinivas [20] since in his model the inhibition probabilities cannot depend on the value of the parent variables if the value differs from the value of the child, which we consider to be quite a restrictive requirement for some applications. Our first and second proposals belong to the class of Generalized Linear Models [11] with a non-canonical link function. The link function is the logarithm, while the canonical link function for a binary dependent variable is the logit function.

In Section 4 we discuss methods one can use to learn parameters of the generalized noisy-or model from data. In Section 5 we present results of numerical experiments on the well-known Reuters text classification data. We use this dataset to compare the performance of suggested generalizations of multivalued and binary noisy-or models. We have made the source code and datasets used in our experiments freely available on the Web. In the final part of the paper we describe how the suggested generalization can be further extended to multivalued child variables and perform learning experiments of the suggested model from artificially generated data.

2. MULTIVALUED NOISY-OR

In this Section, we propose a generalization of noisy-or for multivalued parent variables. Let Y be a binary variable taking on values $y \in \{0, 1\}$, and $X_i, i = 1, \ldots, n$ be multivalued discrete variables taking on values[1] $x_i \in \mathcal{X}_i = \{0, 1, \ldots, m_i\}$, $m_i \in \mathbb{N}^+$. The local structure of both the standard (see, e. g., [5]) and the multivalued generalization of the noisy-or can be made explicit with the help of auxiliary variables $X'_i, i = 1, \ldots, n$ taking values also from $\mathcal{X}_i$. The structure is shown in Figure 1.

The CPT $P(Y|X_1, \ldots, X_n)$ is defined using CPTs $P(X'_i|X_i)$ as

$$P(X'_i = 0|X_i = x_i) = (p_i)^{x_i} \quad (1)$$

$$P(X'_i = 1|X_i = x_i) = 1 - (p_i)^{x_i} \ , \quad (2)$$

where (for $i = 1, \ldots, n$) $p_i \in [0, 1]$ is the parameter which defines the probability that the positive value x_i of variable X_i is inhibited. In the formula, we use parentheses to emphasize that x_i is an exponent, not an upper index of p_i. The CPT $P(Y|X'_1, \ldots, X'_n)$ is deterministic and represents the logical OR function. The higher the value x_i of X_i the lower the probability of $X'_i = 0$, which is a desirable property in many applications.

[1]Generally, the values of X_i could be from $\mathbb{R}$. However, in this paper, we consider CPTs in the context of a Bayesian network where parents of one CPT are children in other CPTs. Therefore we assume that all variables are discrete and their values are from $\mathbb{N}^0$.

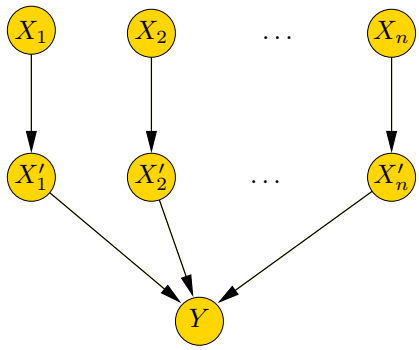

Fig. 1. Noisy-or model with the explicit deterministic (OR) part.

The conditional probability table $P(Y|X_1,\ldots,X_n)$ is then defined as

$$P(Y=0|X_1=x_1,\ldots,X_n=x_n) \quad = \quad \prod_{i=1}^{n} P(X'_i=0|X_i=x_i) \quad = \quad \prod_{i=1}^{n}(p_i)^{x_i} \qquad (3)$$

$$P(Y=1|X_1=x_1,\ldots,X_n=x_n) \quad = \quad 1-\prod_{i=1}^{n}(p_i)^{x_i} \ . \qquad (4)$$

Note that if $m_i = 1$, i. e., the values x_i of X_i are either 0 or 1, then we get the standard noisy-or model [5, 14].

Remark. The suggested generalization influences only the definition of noise $P(X'_i|X_i)$. This implies that the idea can also be similarly applied to models with deterministic parts representing different functions – e. g., the maximum.

In Figure 2 the dependence of the inhibitory probability $P(X'=0|X=x)$ on the value x of a variable X is depicted. In the Figure we can see the shape of the curves representing the dependence for ten different values of the model parameter p.

It is important to note that, contrary to the definition of causal noisy-max [5, Section 4.1.6], we have only one parameter p_i for each parent X_i of Y no matter what the number of values of X_i is. This implies that our model is more restricted. On the other hand, however, the suggested simple parameterization guarantees ordinality, which is a desirable property in many applications (as is also discussed in [5]). Also, since domain experts elicit or learning algorithms estimate fewer parameters, those estimates might be more reliable.

In practical application of noisy-or models, we often lift the requirement that if all parent variables are of the value 0 then the probability of $Y=0$ must be one. One can achieve this by the inclusion of an auxiliary parent variable X_0 whose value is always assumed to be 1. This auxiliary variable is called a leaky cause [5] and its inhibition probability $p_0 = p_L < 1$ is called the leaky probability. This allows the probability

$$P(Y=0|X_1=0,\ldots,X_n=0) \quad = \quad p_L \quad < \quad 1 \ .$$

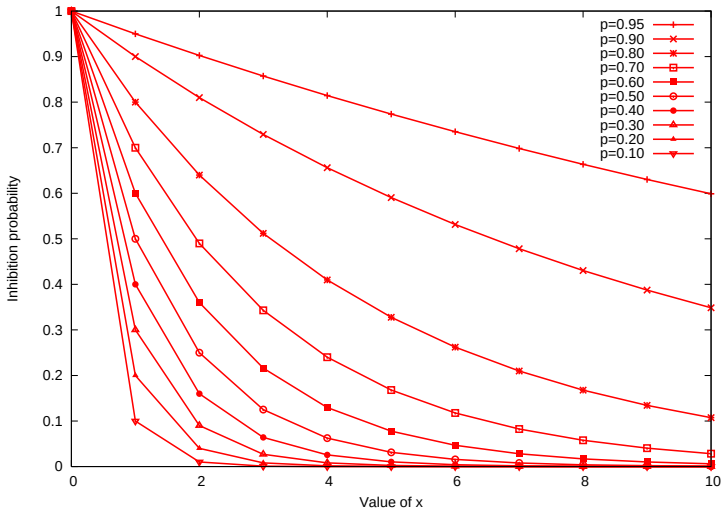

Fig. 2. The dependence of $P(X'_i = 0|X_i = x_i)$ on parameter $p_i = p$ and the variable value $x_i = x$.

In this way we can model unobserved or unknown causes of $Y = 1$. When we include the leaky cause in generalized noisy-or models, Formula (3) is modified to

$$P(Y = 0|X_1 = x_1, \ldots, X_n = x_n) \quad = \quad p_L \prod_{i=1}^{n} (p_i)^{x_i} \quad . \tag{5}$$

Next, we will discuss possible interpretations of multivalued noisy-or. The expression on the left hand side of formula (3) also corresponds to a model where each multivalued variable X_i is replaced by x_i copies of the corresponding binary variable. It means that the probability of the inhibition p_i of multivalued variable X_i corresponds to probability of the inhibition of binary variables repeated in the model as many times as the value of x_i is – with the same parameter value p_i. This seems appropriate, for example, in the classification of text documents discussed in Section 5. The multivalued noisy-or model corresponds to treating each word in a classified text as a separate feature (repeated as many times as it is actually present in the text) with a natural requirement that equivalent words must have the same inhibition probability.

It is possible to give the multivalued noisy-or another interpretation. For $i = 1, \ldots, n$ we can define a new parameter $q_i = p_i^{m_i}$ and replace variable X_i by X'_i so that it takes values $x'_i = \frac{x_i}{m_i}$. The value of q_i is the inhibition probability of a standard noisy-or but with variables X'_i taking fractional values $\frac{x_i}{m_i}$. These values might be interpreted as

degrees of the presence of X_i'. Formula (3) is then modified to:

$$P(Y=0|X_1'=x_1',\ldots,X_n'=x_n') \quad = \quad \prod_{i=1}^{n}(q_i)^{x_i'} \quad \text{, where } x_i' \in \{0, \tfrac{1}{m_i}, \ldots, \tfrac{m_i-1}{m_i}, 1\}.$$

Next, we will describe the relation of the generalized noisy-or model to Generalized Linear Models [11]. Assume the generalized noisy-or model defined by Equation (5) and $p_i > 0, i = 0, 1, \ldots n$. By taking the logarithm on both sides of Equation (5) we get

$$\log P(Y=0|X_1=x_1,\ldots,X_n=x_n) \quad = \quad \log p_L + \sum_{i=1}^{n} x_i \log p_i \quad .$$

Define new parameters $\beta_i = \log p_i, i = 1, \ldots, n$, and $\beta_0 = p_L$. Note that, since $0 < p_i \leq 1, i = 0, 1, \ldots, n$, it holds for $i = 0, 1, \ldots, n$ that

$$-\infty \quad < \quad \beta_i \quad \leq \quad 0 \quad . \tag{6}$$

Then we can write

$$\log P(Y=0|X_1=x_1,\ldots,X_n=x_n) \quad = \quad \beta_0 + \sum_{i=1}^{n} x_i\beta_i \quad = \quad \boldsymbol{\beta}^T\boldsymbol{x} \quad , \tag{7}$$

where $\boldsymbol{x}$ denotes vector $(1, x_1, \ldots, x_n)$, and $\boldsymbol{\beta}$ denotes vector $(\beta_0, \beta_1, \ldots, \beta_n)$. Note that, since the expected value

$$\begin{aligned} E(1-Y|x_1,\ldots,x_n) \quad &= \quad 1 \cdot P(Y=0|x_1,\ldots,x_n) + 0 \cdot P(Y=1|x_1,\ldots,x_n) \quad (8) \\ &= \quad P(Y=0|x_1,\ldots,x_n), \quad (9) \end{aligned}$$

it holds that

$$\log E((1-Y)|x_1,\ldots,x_n) \quad = \quad \boldsymbol{\beta}^T\boldsymbol{x} \quad . \tag{10}$$

3. PARENTS WITH POSITIVE AND NEGATIVE INFLUENCE

In this section we will go one step further with generalizations of the noisy-or model and allow positive values of $\beta_i, i = 1, \ldots, n$. We can give each positive value β_i a quite natural interpretation in generalized noisy-or models – it can mean that the higher values of corresponding X_i imply a higher probability of inhibition of positive influence on Y.

We will allow parents $X_i, i = 1, \ldots, n$ of Y to have a negative influence on probability of the value $Y = 1$ with increasing values of x_i. In the generalized noisy-or model we can treat the parents with positive values of β_i by relabeling their values $x_i \in \{0, 1, \ldots, m_i\}$ to $(m_i - x_i) \in \{m_i, \ldots, 1, 0\}$. In this way the generalized noisy-or is now capable of treating not only positive influences (the presence of X_i increases the probability of $Y = 1$) but also negative influences (the presence of X_i decreases the probability of $Y = 1$).

When learning the generalized noisy-or with positive and negative influences of parents, we need to restrict values of $\beta_i, i = 1, \ldots, n$ by a single constraint:

$$\beta_0 + \sum_{i \in \{1,\ldots,n\}:\beta_i>0} \beta_i m_i \quad \leq \quad 0 \quad , \tag{11}$$

which generalizes the condition (19) from [13, p. 81] to multivalued parents. The positive values of β_i cannot be interpreted as inhibition probabilities $p_i = \exp(\beta_i)$, but they may get an inhibition interpretation with the aid of the transformation discussed above:

$$\begin{aligned} P(Y=0|\boldsymbol{x}) &= \exp\left(\beta_0 + \sum_{i\in\{1,\ldots,n\}} \beta_i x_i\right) \\ &= \exp\left(\beta'_0 + \sum_{i\in\{1,\ldots,n\}:\beta_i>0} \beta'_i(m_i - x_i) + \sum_{i\in\{1,\ldots,n\}:\beta_i\leq 0} \beta_i x_i\right) \end{aligned} \quad (12)$$

where $\beta'_i = -\beta_i$, and $\beta'_0 = \beta_0 + \sum_{i\in\{1,\ldots,n\}:\beta_i>0} \beta_i m_i$, which is non-positive due to the constraint (11). It is now possible to get a probabilistic interpretation of the parents' influence in the model with the new non-positive parameters β'_0 and β'_i.

In the experiments reported in Section 5 we learned these models by, first, restricting $\beta_i \leq 0$ for $i = 0, 1, \ldots, n$ and then transforming the parents X_i with learned $\beta_i \sim 0$ by the above transformation. Finally, we learned the model again requiring $\beta_i \leq 0$ for $i = 0, 1, \ldots, n$. This is easier than optimization with the constraint 11.

4. LEARNING PARAMETERS OF THE GENERALIZED NOISY-OR

We will distinguish between two basic versions of generalized noisy-or models:

(b) a generalized noisy-or with binary parent variables, which we will refer to as *generalized binary noisy-or*,

(m) a generalized noisy-or with multivalued parent variables, which we will refer to as *generalized multivalued noisy-or*.

For both versions the learning algorithms will be the same; they will differ only in data used for learning. In the version where all variables are binary, the data will be transformed so that all values are either 0 or 1. In this paper we define the transformed value[2] to be 0 if and only if the original value is 0, otherwise it is 1.

We consider two versions of the generalized noisy-or models:

(+) include only parents $X_i, i = 1, \ldots, n$ of Y that have a positive influence on the probability of $Y = 1$ (i. e., negative β_i),

(±) include all parents from $X_i, i = 1, \ldots, n$ of Y. Those parents that have negative influence on the probability of $Y = 1$ (i. e., positive β_i) have their values renumbered reversely (as discussed in Section 3) and then are included in the generalized noisy-or models with a positive influence.

[2]Our motivation is the classification of text documents. In the binary version we just consider whether a word is present or not in the document – we use a threshold of 0.5. Generally, any threshold from the interval $(0, m)$ can be applied.

For both versions of the generalized noisy-or model we use a constraint learning method with the constraint $\boldsymbol{\beta} \leq \mathbf{0}$. In this way we guarantee that, for all possible values of $\boldsymbol{x}$, we get probability values (i. e., values from $[0, 1]$). In the experiments reported in Section 5 we used a quasi-Newton method with box constraints [3] implemented in R [15]. When searching for the maximum of the conditional log-likelihood we used the formula (22) for the gradient derived in Appendix A. To avoid infinite numbers we added a small positive value (10^{-10}) to the numerator and the denominator in Formula (22). The algorithm was started from ten different initial values of $\boldsymbol{\beta}$, generated randomly from interval $[-1, +1]$.

To summarize, if we combine the options discussed above we get four different models:

- Generalized binary noisy-or with parents of positive influence[3] (noisy-or+),
- Generalized binary noisy-or with parents of positive and also negative influence (noisy-or±),
- Generalized multivalued noisy-or with parents of positive influence (m-noisy-or+),
- Generalized multivalued noisy-or with parents of positive and also negative influence (m-noisy-or±).

5. EXPERIMENTS

In this section, we will describe experiments we performed with the well-known Reuters-21578 collection (Distribution 1.0) of text documents. The text documents from this dataset appeared on the Reuters newswire in 1987. Personnel from Reuters Ltd. and Carnegie Group, Inc. classified the documents manually into several classes according to their topic. In the test, we further divided documents into training and testing sets according to Apté et al. [2]. We performed experiments with preprocessed data in the eight largest classes[4]. To reduce the feature space we only kept relevant features in our data. Namely, for each class we excluded from the data all features with a correlation to the class of less than 0.3. To allow interested readers to replicate our experiments easily, we have made our R code and the datasets used in experiments available on the Web [5].

In the experiments we compare all versions of the generalized noisy-or classifiers and the generalized multivalued noisy-or classifier with two versions of the logistic classifier, as they are all defined in Section 4.

We decided to include in the models all features that were not rejected as irrelevant at the significance level 0.1. We also performed the experiments with the significance level increased to 0.3. In this way, we can increase the number of features, but we prefer simpler models since there was no significant increase in the accuracy for most classes. However, it may be a topic for future research to apply exhaustive feature selection methods that would find optimal models for the families of our interest.

[3]This is the standard noisy-or. However, to stress the relation to other noisy-or models discussed in this paper, we will use the abbreviation noisy-or+.

[4]The preprocessed dataset is available at `http://web.ist.utl.pt/acardoso/datasets/`.

[5]The code and data are available at `http://www.utia.cas.cz/vomlel/generalized-noisy-or/`

In Table 1 we present the accuracy, sensitivity, specificity, and number of selected features of individual classifiers for binarized and multivalued data. All tested classifiers could be tuned up so that they sacrifice specificity to sensitivity and vice versa by a modification of the threshold. But we did not experiment with the threshold, we simply kept it fixed to 1/2.

noisy-or+				
	acc.	sens.	spec.	feat.
earn	92.46	96.49	88.52	9
acq	89.49	83.76	92.16	6
crude	97.58	61.98	99.66	4
money-fx	96.98	59.77	98.53	7
interest	96.67	17.28	99.72	3
trade	98.40	65.33	99.57	8
ship	98.77	50.00	99.58	3
grain	99.91	90.00	99.95	1

m-noisy-or+				
	acc.	sens.	spec.	feat.
earn	93.60	95.48	91.77	7
acq	85.20	59.20	97.32	4
crude	98.54	82.64	99.47	4
money-fx	97.03	47.13	99.10	5
interest	97.85	49.38	99.72	2
trade	98.36	72.00	99.29	5
ship	99.13	66.67	99.67	4
grain	99.91	90.00	99.95	1

noisy-or±				
	acc.	sens.	spec.	feat.
earn	92.46	96.49	88.52	10
acq	90.04	83.62	93.03	9
crude	97.58	61.98	99.66	4
money-fx	96.98	59.77	98.53	7
interest	96.67	17.28	99.72	3
trade	98.40	65.33	99.57	8
ship	98.77	50.00	99.58	3
grain	99.91	90.00	99.95	1

m-noisy-or±				
	acc.	sens.	spec.	feat.
earn	93.60	95.48	91.77	8
acq	85.20	59.20	97.32	6
crude	98.54	82.64	99.47	4
money-fx	97.03	47.13	99.10	5
interest	97.85	49.38	99.72	2
trade	98.36	72.00	99.29	5
ship	99.13	66.67	99.67	4
grain	99.91	90.00	99.95	1

b-logistic				
	acc.	sens.	spec.	feat.
earn	94.34	94.92	93.76	7
acq	89.90	74.57	97.05	7
crude	97.99	71.07	99.56	7
money-fx	96.67	33.33	99.29	10
interest	97.35	35.80	99.72	3
trade	98.81	80.00	99.48	12
ship	99.22	63.89	99.81	4
grain	99.86	70.00	100.00	3

m-logistic				
	acc.	sens.	spec.	feat.
earn	94.47	96.49	92.50	12
acq	91.64	81.61	96.32	9
crude	98.58	80.99	99.61	8
money-fx	97.08	45.98	99.19	11
interest	96.80	22.22	99.67	4
trade	98.86	85.33	99.34	14
ship	99.09	50.00	99.91	4
grain	99.77	50.00	100.00	3

Tab. 1. Results on binarized and multivalued data.

We summarize the results of experiments in Tables 2 and 3. We report the accuracy using the percentage scale, which is the relative proportion of correctly classified documents either as belonging to the given class or not. We print the best achieved accuracy across both versions of data, bold and framed.

From Tables 2 and 3 we can see that multivalued noisy-or is more often better than binary noisy-or. There is almost no difference between the performance of noisy-or+

	nr. doc	noisy-or+	noisy-or±	b-logistic
earn	1083	92.46	92.46	**94.47**
acq	696	89.49	90.04	**91.64**
crude	121	97.58	97.58	**98.58**
money-fx	87	96.98	96.98	**97.08**
interest	81	96.67	96.67	96.80
trade	75	98.40	98.40	**98.86**
ship	36	98.77	98.77	99.09
grain	10	**99.91**	**99.91**	99.77

Tab. 2. Comparisons of the accuracy of the generalized noisy-or classifiers and the logistic classifier on binarized data.

	nr. doc	m-noisy-or+	m-noisy-or±	m-logistic
earn	1083	93.60	93.60	94.34
acq	696	85.20	85.20	89.90
crude	121	98.54	98.54	97.99
money-fx	87	97.03	97.03	96.67
interest	81	**97.85**	**97.85**	97.35
trade	75	98.36	98.36	98.81
ship	36	99.13	99.13	**99.22**
grain	10	**99.91**	**99.91**	99.86

Tab. 3. Comparisons of the accuracy of the generalized noisy-or classifiers and the logistic classifier on multivalued data.

and noisy-or±. In the case of binary data, noisy-or± is better than noisy-or+ for one class only; in the case of multivalued data both classifiers perform equally well. An explanation might be that in the Reuters data there are not many words whose presence would be significant for concluding that a document does not belong to a certain class. The difference between noisy-or+ and noisy-or± might turn out to be significant for a different dataset. It is important that in our experiments m-noisy-or± is never worse than m-noisy-or+, which is a desirable property for any good generalization of noisy-or.

From the tested models the best performing model is the logistic regression model. Its accuracy is the best for five classes in the binary case and for one class in the multivariate case. m-noisy-or± has the best performance for two classes.

6. ONE STEP FURTHER: A MULTIVALUED GRADED CHILD VARIABLE

A natural generalization of the noisy-or model to multivalued parents and a multivalued child is the noisy-max. Assume that Y takes on values $y \in \{0, 1, \ldots, m\}$, where $m = \max\{m_1, \ldots, m_n\}$. Similarly as the noisy-or, the noisy-max can also be defined with the help of auxiliary variables X'_i and inhibition probabilities p_{y,x_i}. The latter can be different for different values x_i of X_i and y of Y. The structure is the same as for noisy-or

defined in Figure 1. It is convenient to express conditional probabilities in the form of cumulative distribution function for $y = 0, 1, \ldots, m$ and $x_i = 0, 1, \ldots, m_i$

$$P(Y \leq y|\boldsymbol{x}) \quad = \quad \prod_{i=1}^{n} P(X_i' \leq y|X_i = x_i) \quad = \quad \prod_{i=1}^{n} p_{y,x_i} \ , \tag{13}$$

which requires $m \cdot (n + \sum_{i=1}^{n} m_i)$ model parameters since $p_{m,x_i}, i = 1, \ldots, n$ is defined to be one. See [5] for a discussion of different versions of noisy-max. When all variables are ordinal, which is the case considered in this paper, then it is reasonable to assume that, for $y = 0, 1, \ldots, m$,

$$x_i < x_i' \quad \text{implies} \quad p_{y,x_i} \geq p_{y,x_i'} \ . \tag{14}$$

Following the generalization presented in Section 2 of this paper, we can further extend the generalized noisy-or also for a multivalued child variable so that the number of model parameters is still n (or $n + 1$ if we also consider the leaky cause). We define

$$P(X_i' \leq y|X_i = x_i) \quad = \quad (p_i)^{R(x_i - y)} \ , \tag{15}$$

where R is the ramp function:

$$R(x) \quad = \quad \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } x > 0. \end{cases} \tag{16}$$

The meaning is that X_i with values $x_i \leq y$ does not contribute to $Y \geq y$. Please note that this definition satisfies (14). As a consequence, Formulas (5) and (4) are generalized for a multivalued variable Y as follows. For $y = 0, 1, \ldots, m$ it holds

$$P(Y \leq y|\boldsymbol{x}) \quad = \quad (p_L)^{R(m-y)} \prod_{i=1}^{n} (p_i)^{R(x_i - y)}. \tag{17}$$

Let $(y, \boldsymbol{x})$ denote a data vector with $\boldsymbol{x} = (x_0, x_1, \ldots, x_n)$ and x_0 be fixed at value m. Assume that both x_i and y take values from the set $\{0, 1, \ldots, m\}$, and probabilities $p_0, p_1, \ldots, p_n$ are non-zero. Denote the vector of logarithms of the model's probability parameters by $\boldsymbol{\beta} = (\log(p_0), \log(p_1), \ldots, \log(p_n))$ with $p_0 = p_L$. Then we can write Formula (17) as

$$P(Y \leq y|\boldsymbol{x}) \quad = \quad \exp\left(\boldsymbol{\beta}^T R(\boldsymbol{x} - \boldsymbol{y})\right) \ , \tag{18}$$

where $\boldsymbol{y}$ is the vector of $n + 1$ copies of the value y.

Remark. Note that if $y = m$ then $R(m - y) = 0$ and for $i = 1, \ldots, n$ it holds that $R(x_i - y) = 0$. This implies that for all $\boldsymbol{x} = (x_1, \ldots, x_n)$ it holds that $P(Y \leq m|\boldsymbol{x}) = 1$.

If we set $m = 1$ then we get the generalized noisy-or from the previous sections of this paper. If $m = 1$ and $m_i = 1, i = 1, \ldots, n$ then we get the standard noisy-or. Our generalization from this section is a special case of the graded noisy-max model proposed by Díez in [4, Definition 2]. The main difference is that our model requires only $n + 1$ parameters (including the leaky cause p_L). We will refer to our model as the *Simple Graded Noisy-max*.

Learning of Simple Graded Noisy-max

We learn the Simple Graded Noisy-max model using a constraint learning method with the constraint $\boldsymbol{\beta} \leq \mathbf{0}$. This constraint guarantees that we get probability values for all possible values of $\boldsymbol{x}$ in formula (18) . To maximize the conditional log-likelihood, we use a quasi-Newton method with box constraints [3] implemented in R [15]. The method uses Formula (25) for the gradient derived in Appendix B. To avoid infinite numbers we added a small positive value (10^{-10}) to the numerator and the denominator in Formula (25). The algorithm was initialized at ten different values of $\boldsymbol{\beta}$ generated randomly from interval $[-1, +1]$.

Since we did not have real data for a learning experiment we created data artificially. We used a Simple Graded Noisy-max model with three parents, each taking values from $\{0, 1, 2\}$. The child variable was also ternary and took values from $\{0, 1, 2\}$. During the data generation process we repeated the following two steps. First, we randomly generated a parent configuration $\boldsymbol{x}$ (with uniform probability), and then we randomly generated value y with probabilities defined by Formula 18 for the given $\boldsymbol{x}$. We performed experiments with different numbers of generated data vectors. We used the learning algorithm described above to learn the parameters $\boldsymbol{\beta}$ of the Simple Graded Noisy-max.

In Figure 3 we present the results of our experiments. The left-hand side plot describes the dependence of the total sum of the Kullback-Leibler divergence of the conditional probability tables of the true model and the model learned from the generated data on the size of the training dataset. We compare two models: the Simple Graded Noisy-max (the full line) and the full conditional probability table (CPT) computed from relative frequencies of data vectors in generated data (the dashed line). Note the logarithmic scale of the vertical axis. On the right-hand side we present the dependence of Euclidean distance of parameter vectors $\boldsymbol{\beta}$ of the true and the learned models on the size of the training dataset.

From the plots we can see that training datasets of a relatively small size (~ 1000) are sufficient for learning the parameters of the Simple Graded Noisy-max. The parameters of the full CPT are much harder to estimate properly. In the experiment, a general CPT required 54 parameters to be learned while to specify the Simple Graded Noisy-max we needed to learn only 4 parameters. This explains the difference in the learning speed.

7. COMPARISON WITH OTHER MODELS FOR GRADED VARIABLES

Probably the most popular model for modeling ordered response variables is the cumulative logit model. McCullagh calls this model the Proportional Odds Model [10]. Cumulative logit models are known as Graded Response Models [17] in the Item Response Theory. They were applied in Bayesian networks by Russel Almond et al. [1]. The conditional probability of a Graded Response Model is defined as

$$P(Y \leq y|\boldsymbol{x}) \quad = \quad \frac{\exp(\alpha_y + \boldsymbol{\beta}^T \boldsymbol{x})}{1 + \exp(\alpha_y + \boldsymbol{\beta}^T \boldsymbol{x})} \ . \tag{19}$$

Note that the Graded Response Model is specified by $m+n$ parameters while the Simple Graded Noisy-max requires only $n + 1$ parameters.

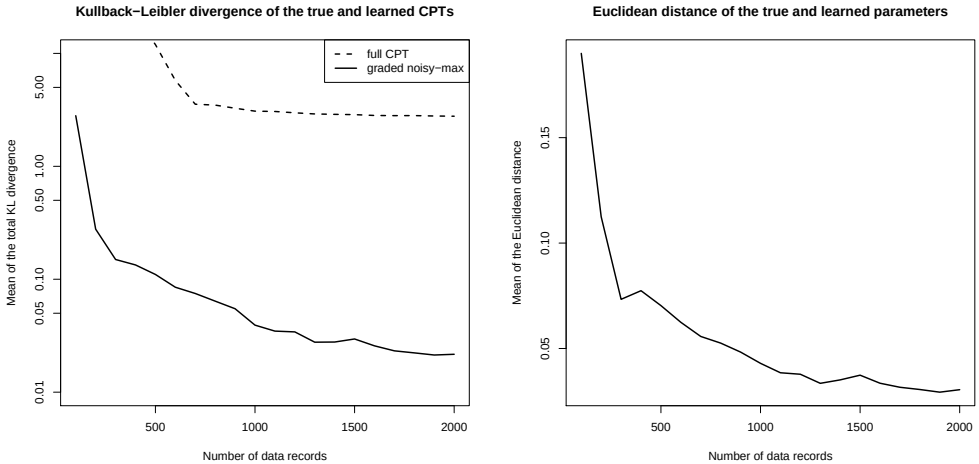

Fig. 3. Dependence of the quality of the learned simple graded noisy-max models on the size of the training dataset.

In Figures 4 and 5 we compare the conditional probability distributions of a Simple Graded Noisy-max and a Graded Response Models – both with one parent X. Variable Y takes on four values and each curve corresponds to one of these values. In Figure 4 the conditional probabilities of a Simple Graded Noisy-max are presented while in Figure 5 we can see the conditional probabilities of a Graded Response Model. The horizontal axis corresponds to the values of x and the vertical axis to the probabilities of $P(Y = y|x)$.

Remark. One might try to further generalize the generalized linear model with the log link function to a multivalued graded child variable by defining

$$P(Y \leq y|\boldsymbol{x}) \quad = \quad \exp(\alpha_y + \boldsymbol{\beta}^T \boldsymbol{x}) \tag{20}$$

and by requiring non-positive α_y and $\boldsymbol{\beta}$. But this model does not seem useful since it is not possible for more than two values y of Y to be the most probable at least for certain values of $\boldsymbol{\beta}^T \boldsymbol{x}$. See Figure 6, in which we can observe that only two values y of Y attain maximum probability at a certain interval of values x.

8. CONCLUSIONS

In this paper we propose generalizations of the popular noisy-or model to multivalued parent variables and allow parent variables with both positive and negative influence on the child variable. To learn generalized noisy-or models we use a quasi-Newton method with box constraints, while for the logistic regression we use iteratively reweighted least-squares method. In the experiments with the Reuters text collection, generalized noisy-or models perform equally well or better than standard noisy-or models. Generalized

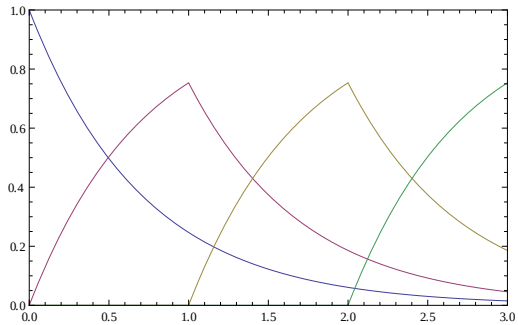

Fig. 4. Conditional probability distributions of a Simple Graded Noisy-max

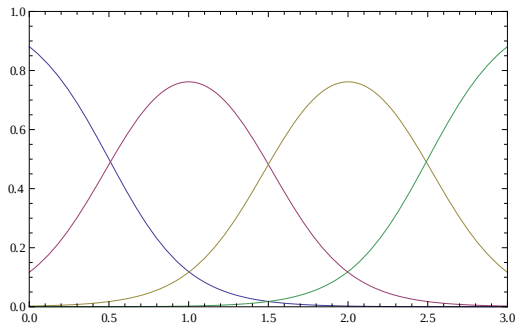

Fig. 5. Conditional probability distributions of a Graded Response Model

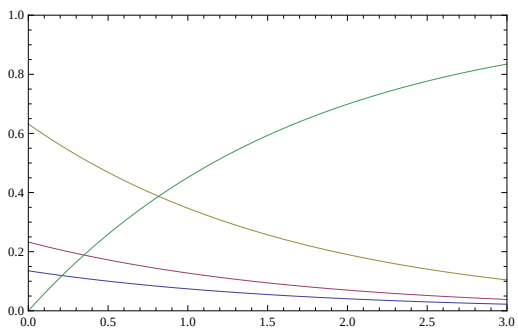

Fig. 6. Conditional probability distributions of a Graded Log Model

noisy-or models represent handy generalizations of noisy-or for real applications with multivalued variables and/or with parent variables being a mixture of variables having either positive or negative influence on their child variables.

The generalizations of noisy-or can be used as local models for conditional probability tables (CPTs) of a Bayesian network similarly as logistic regression models are used in BNs called sigmoid belief networks [13, 16]. The noisy-or and its generalizations have an advantage over logistic regression models. The exact probabilistic inference with densely connected sigmoid belief networks is intractable and approximate inference methods have to be used [18]. On the other hand, the exact probabilistic inference with BNs with (generalized) noisy-or and noisy-max can still be tractable, since this CPT can be nicely decomposed using CP tensor decomposition [6, 19] with the rank equal to the number of states of the child variable.

APPENDIX

A. THE CONDITIONAL LOG-LIKELIHOOD AND ITS GRADIENT FOR THE BINARY RESPONSE

Let $(y, \boldsymbol{x})$ denote a data vector from a training dataset $\mathcal{D}$, where $\boldsymbol{x} = (x_0, x_1, \ldots, x_n)$, and x_0 is fixed at value 1. Assume that x_i take values from the set $\{0, 1, \ldots, m\}$ and y from the set $\{0, 1\}$. Under the generalized noisy-or model the probability of y given $\boldsymbol{x}$ is defined as:

$$\begin{aligned} \log P(y=0|\boldsymbol{x}) &= \boldsymbol{\beta}^T \boldsymbol{x} \\ \log P(y=1|\boldsymbol{x}) &= \log(1 - \exp(\boldsymbol{\beta}^T \boldsymbol{x})) \ . \end{aligned}$$

The conditional log-likelihood of data given this model is

$$\ell(\boldsymbol{\beta}) = \sum_{(y,\boldsymbol{x})\in\mathcal{D}} (1-y)\boldsymbol{\beta}^T \boldsymbol{x} \ + \ y \log(1 - \exp(\boldsymbol{\beta}^T \boldsymbol{x})) \ . \tag{21}$$

The gradient of the conditional log-likelihood is the vector of partial derivatives with respect to $\boldsymbol{\beta}$:

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \sum_{(y,\boldsymbol{x})\in\mathcal{D}} (1-y)\boldsymbol{x} \ - \ y\boldsymbol{x} \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{x})}{1 - \exp(\boldsymbol{\beta}^T \boldsymbol{x})} \\ &= \sum_{(y,\boldsymbol{x})\in\mathcal{D}} \frac{(1-y)\boldsymbol{x} - (1-y)\boldsymbol{x}\exp(\boldsymbol{x}^T\boldsymbol{\beta}) - y\boldsymbol{x}\exp(\boldsymbol{\beta}^T \boldsymbol{x})}{1 - \exp(\boldsymbol{\beta}^T \boldsymbol{x})} \\ &= \sum_{(y,\boldsymbol{x})\in\mathcal{D}} \boldsymbol{x} \frac{(1-y) - \exp(\boldsymbol{\beta}^T \boldsymbol{x})}{1 - \exp(\boldsymbol{\beta}^T \boldsymbol{x})} \ . \end{aligned} \tag{22}$$

B. THE CONDITIONAL LOG-LIKELIHOOD AND ITS GRADIENT FOR THE GRADED RESPONSE

Let $(y, \boldsymbol{x})$ be a data vector from $\mathcal{D}$. Under the simple graded noisy-max (defined in Section 6) the probability of observing y given $\boldsymbol{x}$ is

$$P(y|\boldsymbol{x}) \quad = \quad \exp\left(\boldsymbol{\beta}^T R(\boldsymbol{x}-\boldsymbol{y})\right) - H(y)\exp\left(\boldsymbol{\beta}^T R(\boldsymbol{x}-\boldsymbol{y}+\mathbf{1})\right) \ ,$$

where $\boldsymbol{y}$ is the vector of $n+1$ copies of the value y and H is the step function:

$$H(y) \quad = \quad \begin{cases} 0 & \text{if } y \leq 0 \\ 1 & \text{otherwise.} \end{cases} \tag{23}$$

The conditional log-likelihood of data $\mathcal{D}$ given the model is

$$\ell(\boldsymbol{\beta}) \quad = \quad \sum_{(y,\boldsymbol{x})\in\mathcal{D}} \log\left(\exp\left(\boldsymbol{\beta}^T R(\boldsymbol{x}-\boldsymbol{y})\right) - H(y)\exp\left(\boldsymbol{\beta}^T R(\boldsymbol{x}-\boldsymbol{y}+\mathbf{1})\right)\right) \ . \tag{24}$$

The gradient of the conditional log-likelihood is the vector of partial derivatives with respect to $\boldsymbol{\beta}$:

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \tag{25}$$

$$= \sum_{(y,\boldsymbol{x})\in\mathcal{D}} \frac{\exp\left(\boldsymbol{\beta}^T R(\boldsymbol{x}-\boldsymbol{y})\right) R(\boldsymbol{x}-\boldsymbol{y}) - H(y)\exp\left(\boldsymbol{\beta}^T R(\boldsymbol{x}-\boldsymbol{y}+\mathbf{1})\right) R(\boldsymbol{x}-\boldsymbol{y}+\mathbf{1})}{\exp\left(\boldsymbol{\beta}^T R(\boldsymbol{x}-\boldsymbol{y})\right) - H(y)\exp\left(\boldsymbol{\beta}^T R(\boldsymbol{x}-\boldsymbol{y}+\mathbf{1})\right)} \ .$$

ACKNOWLEDGEMENT

I am grateful to Remco Bouckaert from The University of Auckland, New Zealand for his suggestion to consider generalizations of noisy-or classifier [21] to multivalued variables and to anonymous reviewers for their comments and suggestions that helped me to improve the paper. This work was supported by the Czech Science Foundation through Project 13-20012S. A preliminary version of this paper appeared in the proceedings of The 16th Czech-Japan Seminar on Data Analysis and Decision Making under Uncertainty [22].

(Received June 19, 2014)

REFERENCES

[1] R. G. Almond, R. J. Mislevy, L. Steinberg, D. Yan, and D. Williamson: Bayesian Networks in Educational Assessment. Statistics for Social and Behavioral Sciences. Springer, New York 2015. DOI:10.1007/978-1-4939-2125-6

[2] Ch. Apté, F. Damerau, and S. M. Weiss: Automated learning of decision rules for text categorization. ACM Trans. Inform. Syst. *12* (1994), 3, 233–251. DOI:10.1145/183422.183423

[3] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu: A limited memory algorithm for bound constrained optimization. SIAM J. Sci. Comput. *16* (1995), 1190–1208. DOI:10.1137/0916069

[4] F. J. Díez: Parameter adjustment in Bayes networks. The generalized noisy OR gate. In: Proc. Ninth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann 1993, pp. 99–105. DOI:10.1016/b978-1-4832-1451-1.50016-0

[5] F. J. Díez and M. J. Druzdzel: Canonical Probabilistic Models for Knowledge Engineering. Technical Report CISIAD-06-01, UNED, Madrid 2006.

[6] F. J. Díez and S. F. Galán: An efficient factorization for the noisy MAX. Int. J. Intell. Syst. *18* (2003), 165–177. DOI:10.1002/int.10080

[7] D. Heckerman and J. Breese: A new look at causal independence. In: Proc. Tenth Conference on Uncertainty in Artificial Intelligence, Seattle, Morgan Kaufmann 1994, pp. 286–292. DOI:10.1016/b978-1-55860-332-5.50041-9

[8] M. Henrion: Practical issues in constructing a Bayes' Belief Network. In: Proc. Third Conference Annual Conference on Uncertainty in Artificial Intelligence, AUAI Press 1987, pp. 132–139.

[9] F. V. Jensen and T. D. Nielsen: Bayesian Networks and Decision Graphs. Second edition. Springer, 2007. DOI:10.1007/978-0-387-68282-2

[10] P. McCullagh: Regression models for ordinal data. J. Roy. Statist. Soc. Series B (Methodological) *42* (1980), 109–142.

[11] P. McCullagh and J. A. Nelder: Generalized Linear Models. Chapman and Hall, London 1989. DOI:10.1007/978-1-4899-3242-6

[12] R. A. Miller, F. E. Fasarie, and J. D. Myers: Quick medical reference (QMR) for diagnostic assistance. Medical Comput. *3* (1986), 34–48.

[13] R. M. Neal: Connectionist learning of belief networks. Artif. Intell. *56* (1992), 1, 71–113. DOI:10.1016/0004-3702(92)90065-6

[14] J. Pearl: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufman, San Mateo 1988.

[15] R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna 2008.

[16] F. Rijmen: Bayesian networks with a logistic regression model for the conditional probabilities. Int. J. Approx. Reas. *48* (2008), 2, 659–666. In Memory of Philippe Smets. DOI:10.1016/j.ijar.2008.01.001

[17] F. Samejima: Estimation of Latent Ability Using a Response Pattern of Raded Scores (Psychometric Monograph No. 17). Psychometric Society, Richmond 1969.

[18] L. K. Saul, T. Jaakkola, and M. I. Jordan: Mean field theory for sigmoid belief networks. J. Artif. Intell. Res. *4* (1996), 61–76.

[19] P. Savický and J. Vomlel: Exploiting tensor rank-one decomposition in probabilistic inference. Kybernetika *43* (2007), 5, 747–764.

[20] S. Srinivas: A generalization of the noisy-or model. In: Proc. Ninth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann 1993, pp. 208–215. DOI:10.1016/b978-1-4832-1451-1.50030-5

[21] J. Vomlel: Noisy-or classifier. Int. J. Intell. Syst. *21* (2006), 381–398. DOI:10.1002/int.20141

[22] J. Vomlel: A generalization of the noisy-or model to multivalued parent variables. In: Proc. 16th Czech-Japan Seminar on Data Analysis and Decision Making under Uncertainty 2013, pp. 19–27.

[23] J. Vomlel and P. Tichavský: On tensor rank of conditional probability tables in Bayesian networks. A preprint arXiv:1409.6287, 2014.

[24] A. Zagorecki and M. J. Druzdzel: Knowledge engineering for Bayesian networks: How common are noisy-MAX distributions in practice? IEEE Trans. Systems, Man, and Cybernetics: Systems *43* (2013) 186–195. DOI:10.1109/tsmca.2012.2189880

Jiří Vomlel, Institute of Information Theory and Automation — Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 4, 18208 Praha 8 and Faculty of Management, University of Economics, Prague, Jarošovská 1117/II, 37701 Jindřichův Hradec. Czech Republic.

e-mail: vomlel@utia.cas.cz