A generalization of the noisy-or model to multivalued parent variables*

Jiří Vomlel
Institute of Information Theory and Automation of the AS CR
Academy of Sciences of the Czech Republic
http://www.utia.cas.cz/vomlel

Abstract

In this paper we propose a generalization of the noisy-or model to multivalued parent variables. Albeit the proposed generalization is more restrictive than previous proposals, it has several nice properties. In this paper we suggest a method for learning this model and report results of experiments on the Reuters text classification data.

1 Introduction

The conditional probability tables (CPTs) that are the basic building blocks of Bayesian networks [9, 6] have, generally, an exponential size with respect to the number of variables of the CPT. This has two unpleasant consequences. First, during the elicitation of model parameters one needs to estimate an exponential number of parameters. Second, in case of 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. [8]) often have a simple local structure of the CPTs. The noisy-or model [9] is a popular model for describing relations between variables in one CPT of a Bayesian network. Noisy-or is member of a family of models called models of independence of causal influence [4] or canonical models [2]. The advantage of these models is that the number of parameters required for their specification is linear with respect to the number of variables in CPTs and that they allow applications of efficient inference methods, see for example [3, 11].

In this paper we propose a generalization of the noisy-or model to multivalued parent variables. Our proposal differs from the noisy-max model [5] since we keep the child variable binary no matter what the number of states of the parent variables are. Also we have only one parameter for each parent no matter

*This work was supported by the Czech Science Foundation through the project 13-20012S.

the number of states of the parent variables. Our generalization is also different than the generalization of the noisy-or model proposed by Srinivas [12] since in his model the inhibitor probabilities cannot depend on the state of the parent variables if the state differs from the state of the child. We find this to be a quite restrictive requirement for some applications.

We will show that our proposal is closely connected with the Poisson Regression of Generalized Linear Models [7]. Due to this connection we can use methods from Poisson Regression for learning parameters of the generalized noisy-or model from data. In the paper we present results of numerical experiments on the well-known Reuters text classification data. We use this dataset to compare the performance of our multinomial generalization of noisy-or with the standard noisy-or.

2 Multinomial noisy-or

In this section we propose a generalization of noisy-or for multivalued parent variables. Let Y be a binary variable taking states $y \in \{0, 1\}$ and $X_i, i = 1, \ldots, n$ be multivalued discrete variables taking states $x_i \in \{0, 1, \ldots, m_i\}$, $m_i \in \mathbb{N}^+$. The local structure of both the standard (see, e.g., [2]) and the multinomial generalization of the noisy-or can be made explicit as it is shown in Figure 1.

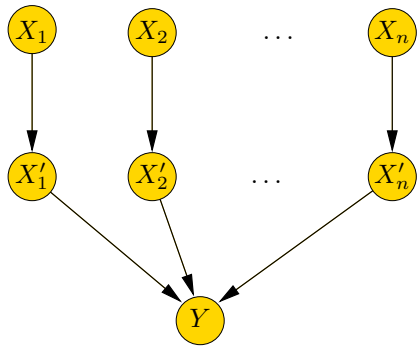

Figure 1: Noisy-or model with the explicit deterministic (OR) part.

The conditional probability table $P(Y|X_1, \ldots, X_n)$ 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 $p_i \in [0, 1]$ is the parameter that defines the probability that the positive value x_i of variable X_i is inhibited. In the formula, we use parenthesis 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.

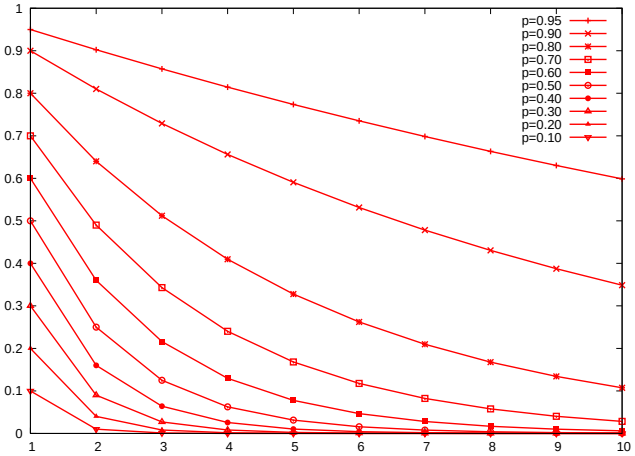

Figure 2: The dependence of $P(X'=0|X=x)$ on p and x.

Remark. Note that the higher is the value x_i of X_i the lower the probability of $X_i'=0$, which is a desirable property in many applications.

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

$$\begin{aligned} P(Y=0|X_1=x_1,\ldots,X_n=x_n) &= \prod_{i=1}^{n} P(X_i'=0|X_i=x_i) \\ &= \prod_{i=1}^{n} (p_i)^{x_i} \qquad (3) \\ P(Y=1|X_1=x_1,\ldots,X_n=x_n) &= 1-\prod_{i=1}^{n} (p_i)^{x_i} \ . \qquad (4) \end{aligned}$$

Remark. Note that if $m_i=1$, i.e. the values x_i of X_i are either 0 or 1, then we get the classical noisy-or model.

In Figure 2 dependence of the inhibitory probability $P(X'=0|X=x)$ on the value x of a variable X is depicted for different values of the parameter p.

It is important to note that contrary to the definition of causal noisy-max [2, Section 4.1.6] we have only one parameter p_i for each parent X_i of Y no matter what is the number of states of X_i. This implies that our model is more restricted. But, on the other hand, the suggested simple parametrization guarantees ordinality, which is in many application a desirable property (as it is also discussed in [2]). Also, since we estimate or elicite (from domain experts) fewer parameters, the estimates are more reliable.

3 Correspondence to Poisson Regression

Next we will show the correspondence of the multionomial noisy-or to the Poisson Regression of Generalized Linear Models [7].

By taking the logarithm of both sides of equation (3) we get

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

Define a new parameter $r_i = \log p_i$. Note that $r_i \in (-\infty, 0]$. Then we get

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

which is the formula of the Poisson regression of the binary variable $1 - Y$. Please, note that the expected value $E(1 - Y|x_1, \ldots, x_n) = P(Y = 0|X_1 = x_1, \ldots, X_n = x_n)$. Therefore

$$\log E((1 - Y)|x_1, \ldots, x_n) \quad = \quad \sum_{i=1}^{n} x_i \cdot r_i \ .$$

This correspondence allows us to apply standard maximum likelihood estimation methods for Poisson regression models to learning multinomial noisy-or. A method typically used to learn the generalized linear models is the iteratively reweighted least squares method [7].

When using a real data that might be modified by a noise or might be generated from a different model it can happen that for some of the $r_i, i = 1, \ldots, n$ parameters we learn positive values. This has a quite natural interpretation in the multinomial noisy-or model. It means that higher values of X_i imply higher inhibitory probability. Therefore we decided to treat positive values of r_i parameters by relabeling the values of X_i from $x_i = 0, 1, \ldots, m_i$ to $m_i - x_i$ in the multinomial noisy-or model. In this way the generalized noisy-or is now capable to treat not only positive (presence of X_i increases probability of $Y = 1$) but also negative influences (presence of X_i decreases probability of $Y = 1$).

4 Experiments

In this section we 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 and were manually classified by personnel from Reuters Ltd. and Carnegie Group, Inc. to several classes according to their topic. In the test we used the split of documents to training and testing sets according to Apté et al. [1]. We performed experiments with preprocessed data for eight largest classes[1].

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

In the experiments we compare the standard noisy-or classifier [13] and our generalized multinomial noisy-or. Both models were learned using the iteratively reweighted least squares method [7] implemented in R – a language and environment for statistical computing [10]. We performed experiments with two versions of both classifiers:

(a) features X_i with both a positive (+) or a negative influence (-) on probability of $Y = 1$ were allowed and treated as it was described in previous section,

(b) features X_i with a negative influence (-) on probability of $Y = 1$ were omitted.

Table 1: Comparisons of the accuracy of the noisy-or and its multinomial generalization. The best achieved accuracy is printed boldface and framed.

Class	# test documents	binomial (+ and -)	binomial (only +)	multinomial (+ and -)	multinomial (only +)
earn	1083	**95.61**	95.02	94.29	94.66
acq	696	**94.20**	91.78	92.01	91.87
crude	121	**97.58**	**97.58**	96.12	96.12
money-fx	87	**96.67**	**96.67**	96.30	96.44
interest	81	96.67	96.67	**97.03**	**97.03**
trade	75	97.44	97.44	**98.13**	**98.13**
ship	36	98.77	98.77	**99.13**	**99.13**
grain	10	**99.91**	**99.91**	99.86	99.86
total	2189				

The results of experiments are summarized in Table 1. The accuracy is reported using the percentage scale, it is the relative proportion of correctly classified documents either as belonging to the given class or not. From Table 1 we can see that standard noisy-or performs better for larger models, while multinomial noisy-or is better at smaller models. The model for the class *grain* is very small, it has one feature only and also the difference between the models' accuracy is very small – it is 0.046, which corresponds to one test case only. In Table 2 we provide the number of selected features for models from Table 1.

We decided to include into the models all features that were not rejected as irrelevant at the significance level 0.1. In the experiments, we observed that the classification accuracy could be slightly improved if the significance was increased to 0.3, this would also slightly improve the AIC criteria[2] However, since the gain was not large we decide to prefer simpler models. Also, it has

[2]The AIC criteria takes into account both the log-likelihood and the number of parameters of the learned model. The lower the AIC the better the model.

Table 2: Comparisons of the number of selected features for the noisy-or and its multinomial generalization.

Class	# test documents	binomial (+ and -)	binomial (only +)	multinomial (+ and -)	multinomial (only +)
earn	1083	17	14	13	12
acq	696	28	20	23	20
crude	121	4	4	3	3
money-fx	87	4	4	4	3
interest	81	3	3	2	2
trade	75	5	6	4	4
ship	36	2	2	3	3
grain	10	1	1	1	1
total	2189				

a very limited influence on the two tested models' preference, which is of our major interest in this paper. However, it might be topic for a future research to apply exhaustive feature selection methods that would find optimal models for the families of our interest.

5 An example

In this section we use the class *ship* to illustrate the benefits of treating the features as multinomial. In the first example we present the standard noisy-or model and in the second the multinomial noisy-or model. Both models were learned by the iteratively reweighted least squares method [7] and contain only significant features for the significance level 0.1. The accuracy of the noisy-or model was 98.77%, while the multinomial noisy-or model achieved accuracy 99.13%. Even if more features are included in the standard noisy-or model the accuracy remains lower than the accuracy of the multinomial noisy-or model.

Example 1 (The noisy-or model for the ship class)**.** In Figure 3 the structure of the noisy-or model for the ship class is presented (in the examples we do not make the deterministic part explicit). The variables are all binary, taking values 0 or 1. The leaky cause has a fixed value 1. The conditional probability $P(class.ship = 0|chip = s, vessel = v)$ is defined as

$$P(class.ship = 0|ship = s, vessel = v) \quad = \quad (p_1)^s \cdot (p_2)^v \cdot p_0 \ ,$$

where $s \in \{0, 1\}$ is the state of feature *ship* and $v \in \{0, 1\}$ is the state of feature *vessel*. The values of parameters p_1, p_2 were estimated to be

$$\begin{aligned} p_1 &= \exp(r_1) = \exp(-0.773407) \doteq 0.461438 \\ p_2 &= \exp(r_2) = \exp(-1.980023) \doteq 0.138066 \end{aligned}$$

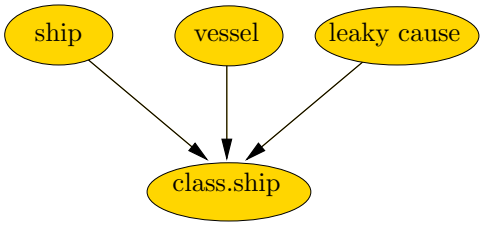

Figure 3: Noisy-or model for the ship class.

and the leaky parameter $p_0 = \exp(r_0)$ was estimated to be

$$p_0 \quad = \quad \exp(r_0) \;=\; \exp(-0.005252) \doteq\; 0.994762 \;\; .$$

This model has accuracy 98.77%.

Example 2 (The multinomial noisy-or model for the ship class)**.** In Figure 4 the structure of the multinomial noisy-or model for the ship class is presented. The variable *ship* takes values from the set $\{0, 1, \ldots, 9\}$, variables *vessel* and

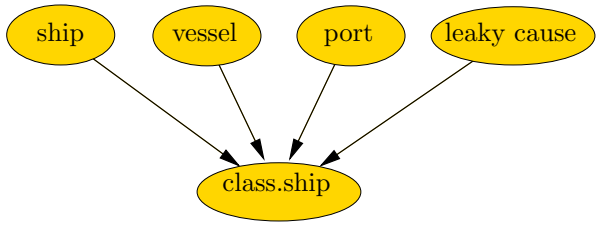

Figure 4: Noisy-or model for the ship class.

port take values from the set $\{0, 1, \ldots, 5\}$. The leaky cause has fixed state 1. The conditional probability $P(Class.ship = 0|Ship = s, Vessel = v, Port)$ is defined as

$$P(Class.ship = 0|Ship = s, Vessel = v) \quad = \quad (p_1)^s \cdot (p_2)^v \cdot (p_3)^p \cdot p_0 \;\; ,$$

where $s \in \{0, 1, \ldots, 9\}$ is the state of feature *ship*, $v \in \{0, 1, \ldots, 5\}$ is the state of feature *vessel*, and $p \in \{0, 1, \ldots, 5\}$ is the state of feature *port*. The values of parameters p_1, p_2, p_3 were estimated to be

$$\begin{aligned} p_1 \quad &= \quad \exp(r_1) \;=\; \exp(-0.467276) \;\doteq\; 0.626707 \\ p_2 \quad &= \quad \exp(r_2) \;=\; \exp(-1.361929) \;\doteq\; 0.256166 \\ p_3 \quad &= \quad \exp(r_3) \;=\; \exp(-0.500009) \;\doteq\; 0.606525 \end{aligned}$$

and the leaky parameter $p_0 = \exp(r_0)$ was estimated to be

$$p_0 \quad = \quad \exp(r_0) \;=\; \exp(-0.001273) \doteq\; 0.998728 \;\; .$$

This model has accuracy 99.13%, which is higher than the accuracy of noisy-or from Example 1.

6 Conclusions

In this paper we proposed a generalization of the popular noisy-or model to multivalued explanatory variables. We showed the correspondence of this model to the Poisson family of generalized linear models and applied iteratively reweighted least squares method to learning of these models. In the experiments with the Reuters text collection the standard noisy-or performed better for larger models, while the multinomial noisy-or was better for smaller models.

Acknowledgments

I am grateful to Remco Bouckaert from The University of Auckland, New Zealand for his suggestion to consider generalizations of noisy-or classifier [13] to multinomial variables.

References

[1] Ch. Apté, F. Damerau, and S. M. Weiss. Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems*, 12(3):233–251, 1994.

[2] F. J. Díez and M. J. Druzdzel. Canonical probabilistic models for knowledge engineering. Technical Report CISIAD-06-01, UNED, Madrid, Spain, 2006.

[3] F. J. Díez and S. F. Galán. An efficient factorization for the noisy MAX. *International Journal of Intelligent Systems*, 18:165–177, 2003.

[4] D. Heckerman and J. Breese. A new look at causal independence. In *Proceedings of Tenth Conference on Uncertainty in Artificial Intelligence, Seattle, WA*, pages 286–292. Morgan Kaufmann, 1994.

[5] Max Henrion. Practical issues in constructing a Bayes' Belief Network. In *Proceedings of the Third Conference Annual Conference on Uncertainty in Artificial Intelligence*, pages 132–139. AUAI Press, 1987.

[6] F. V. Jensen and T. D. Nielsen. *Bayesian Networks and Decision Graphs, 2nd ed.* Springer, 2007.

[7] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 1989.

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

[9] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufman, San Mateo, CA, 1988.

[10] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

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

[12] Sampath Srinivas. A generalization of the noisy-or model. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, pages 208–215. Morgan Kaufmann, 1993.

[13] J. Vomlel. Noisy-or classifier. *International Journal of Intelligent Systems*, 21:381–398, 2006.