Symmetric Causal Independence Models for Classification

Rasa Jurgelenaite and Tom Heskes
Institute for Computing and Information Sciences
Radboud University Nijmegen
Toernooiveld 1, 6525 ED Nijmegen, The Netherlands

Abstract

Causal independence modelling is a well-known method both for reducing the size of probability tables and for explaining the underlying mechanisms in Bayesian networks. In this paper, we propose an application of an extended class of causal independence models, causal independence models based on the symmetric Boolean function, for classification. We present an EM algorithm to learn the parameters of these models, and study convergence of the algorithm. Experimental results on the Reuters data collection show the competitive classification performance of causal independence models based on the symmetric Boolean function in comparison to noisy OR model and, consequently, with other state-of-the-art classifiers.

1 Introduction

Bayesian networks (Pearl, 1988) are well-established as a sound formalism for representing and reasoning with probabilistic knowledge. However, because the number of conditional probabilities for the node grows exponentially with the number of its parents, it is usually unreliable if not infeasible to specify the conditional probabilities for the node that has a large number number of parents. The task of assessing conditional probability distributions becomes even more complex if the model has to integrate expert knowledge. While learning algorithms can be forced to take into account an expert's view, for the best possible results the experts must be willing to reconsider their ideas in light of the model's 'discovered' structure. This requires a clear understanding of the model by the domain expert. *Causal independence models* (Díez, 1993; Heckerman and Breese, 1994; Srinivas, 1993; Zhang and Poole, 1996) can both limit the number of conditional probabilities to be assessed and provide the ability for models to be understood by domain experts in the field. The main idea of causal independence models is that causes influence a given common effect through intermediate variables and interaction function.

Causal independence assumptions are often used in practical Bayesian network models (Kappen and Neijt, 2002; Shwe et al., 1991). However, most researchers restrict themselves to using only the logical OR and logical AND operators to define the interaction among causes. The resulting probabilistic submodels are called *noisy OR* and *noisy AND*; their underlying assumption is that the presence of either at least one cause or all causes at the same time give rise to the effect. Several authors proposed to expand the space of interaction functions by other symmetric Boolean functions: the idea was already mentioned but not developed further in (Meek and Heckerman, 1997), analysis of the qualitative patterns was presented in (Lucas, 2005), and assessment of conditional probabilities was studied in (Jurgelenaite et al., 2006).

Even though for some real-world problems the intermediate variables are observable (see Visscher et al. (2005)), in many problems these variables are latent. Therefore, conditional probability distributions depend on unknown parameters which must be estimated from data,

using *maximum likelihood* (ML) or *maximum a posteriori* (MAP). One of the most widespread techniques for finding ML or MAP estimates is the *expectation-maximization* (EM) algorithm. Meek and Heckerman (1997) provided a general scheme how to use the EM algorithm to compute the maximum likelihood estimates of the parameters in causal independence models assumed that each local distribution function is collection of multinomial distributions. Vomlel (2006) described the application of the EM algorithm to learn the parameters in the noisy OR model for classification.

The application of an extended class of causal independence models, namely causal independence models with a symmetric Boolean function as an interaction function, to classification is the main topic of this paper. These models will further be referred to as the *symmetric causal independence models*. We present an EM algorithm to learn the parameters in symmetric causal independence models, and study convergence of the algorithm. Experimental results show the competitive classification performance of the symmetric causal independence models in comparison with the noisy OR classifier and, consequently, with other widely-used classifiers.

The remainder of this paper is organised as follows. In the following section, we review Bayesian networks and discuss the semantics of symmetric causal independence models. In Section 3, we state the EM algorithm for finding the parameters in symmetric causal independence models. The maxima of the log-likelihood function for the symmetric causal independence models are examined in Section 4. Finally, Section 5 presents the experimental results, and conclusions are drawn in Section 6.

2 Symmetric Boolean Functions for Modelling Causal Independence

2.1 Bayesian Networks

A *Bayesian network* $\mathcal{B} = (G, \Pr)$ represents a factorised joint probability distribution on a set of random variables $\mathbf{V}$. It consists of two parts: (1) a qualitative part, represented as an acyclic directed graph (ADG) $G = (\mathbf{V}(G), \mathbf{A}(G))$,

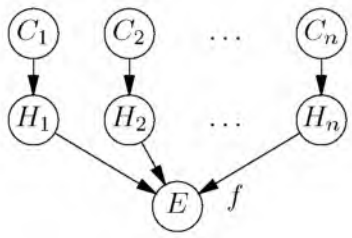

Figure 1: Causal independence model

where there is a 1–1 correspondence between the vertices $\mathbf{V}(G)$ and the random variables in $\mathbf{V}$, and arcs $\mathbf{A}(G)$ represent the conditional (in)dependencies between the variables; (2) a quantitative part Pr consisting of local probability distributions $\Pr(V \mid \pi(V))$, for each variable $V \in \mathbf{V}$ given the parents $\pi(V)$ of the corresponding vertex (interpreted as variables). The joint probability distribution Pr is factorised according to the structure of the graph, as follows:

$$\Pr(\mathbf{V}) = \prod_{V \in \mathbf{V}} \Pr(V \mid \pi(V)).$$

Each variable $V \in \mathbf{V}$ has a finite set of mutually exclusive states. In this paper, we assume all variables to be binary; as an abbreviation, we will often use v^+ to denote $V = \top$ (true) and v^- to denote $V = \bot$ (false). We interpret $\top$ as 1 and $\bot$ as 0 in an arithmetic context. An expression such as

$$\sum_{\psi(H_1,\ldots,H_n)=\top} g(H_1, \ldots, H_n)$$

stands for summing $g(H_1, \ldots, H_n)$ over all possible values of the variables H_k for which the constraint $\psi(H_1, \ldots, H_n) = \top$ holds.

2.2 Semantics of Symmetric Causal Independence Models

Causal independence is a popular way to specify interactions among cause variables. The global structure of a causal independence model is shown in Figure 1; it expresses the idea that causes $C_1, \ldots, C_n$ influence a given common effect E through hidden variables $H_1, \ldots, H_n$ and a deterministic function f, called the *interaction function*. The impact of each cause C_i on the common effect E is independent of each

other cause $C_j, j \neq i$. The hidden variable H_i is considered to be a contribution of the cause variable C_i to the common effect E. The function f represents in which way the hidden effects H_i, and indirectly also the causes C_i, interact to yield the final effect E. Hence, the function f is defined in such a way that when a relationship, as modelled by the function f, between $H_i, i = 1, \ldots, n$, and $E = \top$ is satisfied, then it holds that $f(H_1, \ldots, H_n) = \top$. It is assumed that $\Pr(e^+ \mid H_1, \ldots, H_n) = 1$ if $f(H_1, \ldots, H_n) = \top$, and $\Pr(e^+ \mid H_1, \ldots, H_n) = 0$ if $f(H_1, \ldots, H_n) = \bot$.

A causal independence model is defined in terms of the causal parameters $\Pr(H_i \mid C_i)$, for $i = 1, \ldots, n$ and the function $f(H_1, \ldots, H_n)$. Most papers on causal independence models assume that absent causes do not contribute to the effect (Heckerman and Breese, 1994; Pearl, 1988). In terms of probability theory this implies that it holds that $\Pr(h_i^+ \mid c_i^-) = 0$; as a consequence, it holds that $\Pr(h_i^- \mid c_i^-) = 1$. In this paper we make the same assumption.

In situations in which the model does not capture all possible causes, it is useful to introduce a *leaky cause* which summarizes the unidentified causes contributing to the effect and is assumed to be always present (Henrion, 1989). We model this leak term by adding an additional input $C_{n+1} = 1$ to the data; in an arithmetic context the leaky cause is treated in the same way as identified causes.

The conditional probability of the occurrence of the effect E given the causes $C_1, \ldots, C_n$, i.e., $\Pr(e^+ \mid C_1, \ldots, C_n)$, can be obtained from the causal parameters $\Pr(H_l \mid C_l)$ as follows (Zhang and Poole, 1996):

$$\begin{aligned} &\Pr(e^+ \mid C_1, \ldots, C_n) \\ &= \sum_{f(H_1,\ldots,H_n)=\top} \prod_{i=1}^{n} \Pr(H_i \mid C_i). \end{aligned} \quad (1)$$

In this paper, we assume that the function f in Equation (1) is a Boolean function. However, there are 2^{2^n} different n-ary Boolean functions (Enderton, 1972; Wegener, 1987); thus, the potential number of causal interaction models is huge. However, if we assume that the order of the cause variables does not matter, the Boolean functions become *symmetric* (Wegener, 1987) and the number reduces to 2^{n+1}.

An important symmetric Boolean function is the *exact* Boolean function ϵ_l, which has function value true, i.e. $\epsilon_l(H_1, \ldots, H_n) = \top$, if $\sum_{i=1}^{n} \nu(H_i) = l$ with $\nu(H_i)$ equal to 1, if H_i is equal to true and 0 otherwise. A symmetric Boolean function can be decomposed in terms of the exact functions ϵ_l as (Wegener, 1987):

$$f(H_1, \ldots, H_n) = \bigvee_{i=0}^{n} \epsilon_i(H_1, \ldots, H_n) \wedge \gamma_i \quad (2)$$

where γ_i are Boolean constants depending only on the function f. For example, for the Boolean function defined in terms of the OR operator we have $\gamma_0 = \bot$ and $\gamma_1 = \ldots = \gamma_n = \top$.

Another useful symmetric Boolean function is the *threshold* function τ_k, which simply checks whether there are at least k trues among the arguments, i.e. $\tau_k(I_1, \ldots, I_n) = \top$, if $\sum_{j=1}^{n} \nu(I_j) \geq k$ with $\nu(I_j)$ equal to 1, if I_j is equal to true and 0 otherwise. To express it in the Boolean constants we have: $\gamma_0 = \cdots = \gamma_{k-1} = \bot$ and $\gamma_k = \cdots = \gamma_n = \top$. Causal independence model based on the Boolean threshold function further will be referred to as the *noisy threshold models*.

2.3 The Poisson Binomial Distribution

Using the property of Equation (2) of the symmetric Boolean functions, the conditional probability of the occurrence of the effect E given the causes $C_1, \ldots, C_n$ can be decomposed in terms of probabilities that exactly l hidden variables $H_1, \ldots, H_n$ are true as follows:

$$\begin{aligned} &\Pr(e^+ \mid C_1, \ldots, C_n) \\ &= \sum_{\substack{0 \leq l \leq n \\ \gamma_l}} \sum_{\epsilon_l(H_1,\ldots,H_n)} \prod_{i=1}^{n} \Pr(H_i \mid C_i). \end{aligned}$$

Let l denote the number of successes in n independent trials, where p_i is a probability of success in the ith trial, $i = 1, \ldots, n$; let $\mathbf{p} = (p_1, \ldots, p_n)$, then $\mathrm{B}(l; \mathbf{p})$ denotes the *Poisson binomial distribution* (Le Cam, 1960; Dar-

roch, 1964):

$$\mathrm{B}(l;\mathbf{p}) = \prod_{i=1}^{n}(1-p_i)\sum_{1\le j_1<\ldots<j_l\le n}\prod_{z=1}^{l}\frac{p_{j_z}}{1-p_{j_z}}.$$

Let us define a vector of probabilistic parameters $\mathbf{p}(C_1,\ldots,C_n) = (p_1,\ldots,p_n)$ with $p_i = \Pr(h_i^+ \mid C_i)$. Then the connection between the Poisson binomial distribution and the class of symmetric causal independence models is as follows.

Proposition 1. *It holds that:*

$$\Pr(e^+ \mid C_1,\ldots,C_n) = \sum_{i=0}^{n}\mathrm{B}(i;\mathbf{p}(C_1,\ldots,C_n))\gamma_i.$$

3 EM Algorithm

Let $\mathbf{D} = \{\mathbf{x}^1,\ldots,\mathbf{x}^N\}$ be a data set of independent and identically distributed settings of the observed variables in a symmetric causal independence model, where

$$\mathbf{x}^j = (\mathbf{c}^j, e^j) = (c_1^j,\ldots,c_n^j,e^j).$$

We assume that no additional information about the model is available. Therefore, to learn the parameters of the model we maximize the conditional log-likelihood

$$CLL(\boldsymbol{\theta}) = \sum_{j=1}^{N}\ln\Pr(e^j \mid \mathbf{c}^j,\boldsymbol{\theta}).$$

The expectation-maximization (EM) algorithm (Dempster et al., 1977) is a general method to find the maximum likelihood estimate of the parameters in probabilistic models, where the data is incomplete or the model has hidden variables.

Let $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_n)$ be the parameters of the symmetric causal independence model where $\theta_i = \Pr(h_i^+ \mid c_i^+)$. Then, after some calculations, the $(z+1)$-th iteration of the EM algorithm for symmetric causal independence models is given by:

Expectation step: For every data sample $\mathbf{x}^j = (\mathbf{c}^j, e^j)$ with $j = 1,\ldots,N$, we form

$$\mathbf{p}^{(z,j)} = (p_1^{(z,j)},\ldots,p_n^{(z,j)}) \text{ where } p_i^{(z,j)} = \theta_i^{(z)}c_i^j.$$

Let us define

$$\mathbf{p}_{\setminus k}^{(z,j)} = (p_1^{(z,j)},\ldots,p_{k-1}^{(z,j)},p_{k+1}^{(z,j)},\ldots,p_n^{(z,j)}).$$

Subsequently, for all hidden variables H_k with $k = 1,\ldots,n$ we compute the probability $\Pr(h_k^+ \mid e^j,\mathbf{c}^j,\boldsymbol{\theta}^{(z)})$ where

$$\Pr(h_k^+ \mid e^j,\mathbf{c}^j,\boldsymbol{\theta}^{(z)}) = \frac{p_k^{(z,j)}\sum_{i=0}^{n-1}\mathrm{B}\left(i;\mathbf{p}_{\setminus k}^{(z,j)}\right)\gamma_{i+1}}{\sum_{i=0}^{n}\mathrm{B}\left(i;\mathbf{p}^{(z,j)}\right)\gamma_i} \text{ if } e^j = 1,$$

and

$$\Pr(h_k^+ \mid e^j,\mathbf{c}^j,\boldsymbol{\theta}^{(z)}) = \frac{p_k^{(z,j)}\left(1-\sum_{i=0}^{n-1}\mathrm{B}\left(i;\mathbf{p}_{\setminus k}^{(z,j)}\right)\gamma_{i+1}\right)}{1-\sum_{i=0}^{n}\mathrm{B}\left(i;\mathbf{p}^{(z,j)}\right)\gamma_i} \text{ if } e^j = 0.$$

Maximization step: Update the parameter estimates for all $k = 1,\ldots,n$:

$$\theta_k = \frac{\sum_{1\le j\le N}c_k^j\Pr(h_k^+ \mid e^j,\mathbf{c}^j,\boldsymbol{\theta}^{(z)})}{\sum_{1\le j\le N}c_k^j}.$$

4 Analysis of the Maxima of the Log-likelihood Function

Generally, there is no guarantee that the EM algorithm will converge to a global maximum of log-likelihood. In this section, we investigate the maxima of the conditional log-likelihood function for symmetric causal independence models.

4.1 Noisy OR and Noisy AND Models

In this section we will show that the conditional log-likelihood for the noisy OR and the noisy AND models has only one maximum. Since the conditional log-likelihood for these models is not necessarily concave we will use a monotonic transformation to prove the absence of the stationary points other than global maxima.

First, we establish a connection between the maxima of the log-likelihood function and the maxima of the corresponding composite function.

Proposition 2. ***(Global optimality condition for concave functions (Boyd and Vandenberghe, 2004))***
Suppose $h(\mathbf{q}) : \mathbf{Q} \to \Re$ *is concave and differentiable on* $\mathbf{Q}$. *Then* $\mathbf{q}^* \in \mathbf{Q}$ *is a global maximum if and only if*

$$\nabla h(\mathbf{q}^*) = \left(\frac{\partial h(\mathbf{q}^*)}{\partial q_1}, \ldots, \frac{\partial h(\mathbf{q}^*)}{\partial q_n} \right)^T = \mathbf{0}.$$

Further we consider the function

$$CLL(\boldsymbol{\theta}) = h(\mathbf{q}(\boldsymbol{\theta})).$$

Let $CLL(\boldsymbol{\theta})$ and $h(\mathbf{q}(\boldsymbol{\theta}))$ be twice differentiable functions, and let $\mathbf{q}(\boldsymbol{\theta})$ be a differentiable, injective function where $\boldsymbol{\theta}(\mathbf{q})$ is its inverse. Then the following relationship between the stationary points of the functions CLL and h holds.

Lemma 1. *Suppose,* $\boldsymbol{\theta}^*$ *is a stationary point of* $CLL(\boldsymbol{\theta})$. *Then there is a corresponding point* $\mathbf{q}(\boldsymbol{\theta}^*)$, *which is a stationary point of* $h(\mathbf{q}(\boldsymbol{\theta}))$.

Proof. Since the function $\mathbf{q}(\boldsymbol{\theta})$ is differentiable and injective, its Jacobian matrix $\frac{\partial(q_1,\ldots,q_n)}{\partial(\theta_1,\ldots,\theta_n)}$ is positive definite. Therefore, from the chain rule it follows that if $\nabla CLL(\boldsymbol{\theta}^*) = \mathbf{0}$, then $\nabla h(\mathbf{q}(\boldsymbol{\theta}^*)) = \mathbf{0}$. □

Proposition 3. *If* $h(\mathbf{q}(\boldsymbol{\theta}))$ *is concave and* $\boldsymbol{\theta}^*$ *is a stationary point of* $CLL(\boldsymbol{\theta})$, *then* $\boldsymbol{\theta}^*$ *is a global maximum.*

Proof. If $\boldsymbol{\theta}^*$ is a stationary point, then from Lemma 1 it follows that $\mathbf{q}(\boldsymbol{\theta}^*)$ is also stationary. From the global optimality condition for concave functions the stationary point $\mathbf{q}(\boldsymbol{\theta}^*)$ is a maximum of $h(\mathbf{q}(\boldsymbol{\theta}))$, thus from the definition of global maximum we get that for all $\boldsymbol{\theta}$

$$CLL(\boldsymbol{\theta}) = h(\mathbf{q}(\boldsymbol{\theta})) \leq h(\mathbf{q}(\boldsymbol{\theta}^*)) = CLL(\boldsymbol{\theta}^*).$$

□

Given Proposition 3 the absence of the local optima can be proven by introducing such a monotonic transformation $\mathbf{q}(\boldsymbol{\theta})$ that the composite function $h(\mathbf{q}(\boldsymbol{\theta}))$ would be concave. As it is a known result that the Hessian matrix of the log-likelihood function for logistic regression is negative-semidefinite, and hence the problem has no local optima, we will use transformations that allow us to write the log-likelihood for the noisy OR and noisy AND models in a similar form as that of the logistic regression model.

The conditional probability of the effect in a noisy OR model can be written:

$$\Pr(e^+ \mid \mathbf{c}, \boldsymbol{\theta}) = 1 - \prod_{i=1}^{n} \Pr(h_i^- \mid c_i)$$
$$= 1 - \prod_{i=1}^{n} (1 - \theta_i)^{c_i} = 1 - \exp\left(\sum_{i=1}^{n} \ln(1 - \theta_i) c_i \right).$$

Let us choose a monotonic transformation $q_i = -\ln(1 - \theta_i), i = 1, \ldots, n$. Then the conditional probability of the effect in a noisy OR model equals

$$\Pr(e^+ \mid \mathbf{c}, \mathbf{q}) = 1 - e^{-\mathbf{q}^T \mathbf{c}}.$$

Let us define $z^j = \mathbf{q}^T \mathbf{c}^j$ and $f(z^j) = \Pr(e^+ \mid \mathbf{c}^j, \mathbf{q})$, then the function h reads

$$h(\mathbf{q}) = \sum_{j=1}^{N} e^j \ln f(z^j) + (1 - e^j) \ln(1 - f(z^j)). \quad (3)$$

Since $f'(z^j) = 1 - f(z^j)$, the first derivative of h is

$$\frac{\partial h(\mathbf{q})}{\partial \mathbf{q}} = \sum_{j=1}^{N} \frac{f'(z^j)(e^j - f(z^j))}{f(z^j)(1 - f(z^j))} \mathbf{c}^j$$
$$= \sum_{j=1}^{N} \frac{e^j - f(z^j)}{f(z^j)} \mathbf{c}^j.$$

To prove that the function h is concave we need to prove that its Hessian matrix is negative semidefinite. The Hessian matrix of h reads

$$\frac{\partial^2 h(\mathbf{q})}{\partial \mathbf{q} \partial \mathbf{q}^T} = -\sum_{j=1}^{N} \frac{1 - f(z^j)}{f(z^j)^2} e^j \mathbf{c}^j \mathbf{c}^{j\,T}.$$

As the Hessian matrix of h is negative semidefinite, the function h is concave. Therefore, from Proposition 3 it follows that every stationary point of the log-likelihood function for the noisy OR model is a global maximum.

The conditional probability of the effect in a noisy AND model can be written:

$$\begin{aligned}\Pr(e^{j+} \mid \mathbf{c}, \boldsymbol{\theta}) &= \prod_{i=1}^{n} \Pr(h_i^+ \mid c_i) \\ &= \prod_{i=1}^{n} \theta_i^{c_i} = \exp\left(\sum_{i=1}^{n} \ln \theta_i c_i\right).\end{aligned}$$

Let us choose a monotonic transformation $q_i = \ln \theta_i, i = 1, \ldots, n$. Then the conditional probability of the effect in a noisy AND model equals

$$\Pr(e^{j+} \mid \mathbf{c}, \mathbf{q}) = e^{\mathbf{q}^T \mathbf{c}}.$$

Let us define $z^j = \mathbf{q}^T \mathbf{c}^j$ and $f(z^j) = \Pr(e^+ \mid \mathbf{c}^j, \mathbf{q})$. The function h is the same as for the noisy OR model in Equation (3). Combined with $f'(z^j) = f(z^j)$, it yields the first derivative of h

$$\begin{aligned}\frac{\partial h(\mathbf{q})}{\partial \mathbf{q}} &= \sum_{j=1}^{N} \frac{f'(z^j)(e^j - f(z^j))}{f(z^j)(1 - f(z^j))} \mathbf{c}^j \\ &= \sum_{j=1}^{N} \frac{e^j - f(z^j)}{1 - f(z^j)} \mathbf{c}^j\end{aligned}$$

and Hessian matrix

$$\frac{\partial^2 h(\mathbf{q})}{\partial \mathbf{q} \partial \mathbf{q}^T} = -\sum_{j=1}^{N} \frac{f(z^j)}{(1 - f(z^j))^2}(1 - e^j)\mathbf{c}^j \mathbf{c}^{j\,T}.$$

Hence, the function h is concave, and the log-likelihood for the noisy AND model has no other stationary points than the global maxima.

4.2 General Case

The EM algorithm is guaranteed to converge to the local maxima or saddle points. Thus, we can only be sure that the global maximum, i.e. a point $\boldsymbol{\theta}^*$ such that $CLL(\boldsymbol{\theta}^*) \geq CLL(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}$, will be found if the log-likelihood has neither saddle points nor local maxima. However, the log-likelihood function for a causal independence model with any symmetric Boolean function does not always fulfill this requirement as it is shown in the following counterexample.

Example 1. Let us assume a data set $\mathbf{D} = \{(1,1,1,1),(1,0,1,0)\}$ and an interaction function ϵ_1, i.e. $\gamma_1 = 1$ and $\gamma_0 = \gamma_2 = \gamma_3 = 0$. To learn the hidden parameters of the model describing this interaction we have to maximize the conditional log-likelihood function

$$\begin{aligned}CLL(\boldsymbol{\theta}) &= \ln[\theta_1(1-\theta_2)(1-\theta_3) \\ &+(1-\theta_1)\theta_2(1-\theta_3) + (1-\theta_1)(1-\theta_2)\theta_3] \\ &+\ln[1-\theta_1(1-\theta_3) - (1-\theta_1)\theta_3].\end{aligned}$$

Depending on a choice for initial parameter settings $\boldsymbol{\theta}^{(0)}$, the EM algorithm for symmetric causal independence models converges to one of the maxima:

$$\begin{aligned}&CLL(\boldsymbol{\theta})_{\max} \\ &= \begin{cases} 0 & \text{at } \boldsymbol{\theta} = (0,1,0), \\ -1.386 & \text{at } \boldsymbol{\theta} \in \{\left(\theta_1, 0, \frac{1}{2}\right), \left(\frac{1}{2}, 0, \theta_3\right)\}. \end{cases}\end{aligned}$$

Obviously, only the point $\boldsymbol{\theta} = (0,1,0)$ is a global maximum of the log-likelihood function while the other obtained points are local maxima.

The discussed counterexample proves that in general case the EM algorithm for symmetric causal independence models does not necessarily converge to the global maximum.

5 Experimental Results

For our experiments we use Reuters data collection, which allows us to evaluate the classification performance of large symmetric causal independence models where the number of cause variables for some document classes is in the hundreds.

5.1 Evaluation Scheme

Since we do not have an efficient algorithm to perform a search in the space of symmetric Boolean functions, we chose to model the interaction among cause and effect variables by means of Boolean threshold functions, which seem to be the most probable interaction functions for the given domains.

Given the model parameters $\boldsymbol{\theta}$, the testing data $\mathbf{D}_{test}$ and the classification threshold $\frac{1}{2}$, the classifications and misclassifications for both classes are computed. Let *tp (true positives)* stand for the number of data samples $(\mathbf{c}^j, e^{j+}) \in \mathbf{D}_{test}$ for which $\Pr(e^+ \mid \mathbf{c}^j, \boldsymbol{\theta}) \geq \frac{1}{2}$ and *fp (false*

positives) stand for the number of data samples $(\mathbf{c}^j, e^{j+}) \in \mathbf{D}_{test}$ for which $\Pr(e^+ \mid \mathbf{c}^j, \boldsymbol{\theta}) < \frac{1}{2}$. Likewise, *tn (true negatives)* is the number of data samples $(\mathbf{c}^j, e^{j-}) \in \mathbf{D}_{test}$ for which $\Pr(e^+ \mid \mathbf{c}^j, \boldsymbol{\theta}) < \frac{1}{2}$ and *fp (false positives)* is the number of data samples $(\mathbf{c}^j, e^{j-}) \in \mathbf{D}_{test}$ for which $\Pr(e^+ \mid \mathbf{c}^j, \boldsymbol{\theta}) \geq \frac{1}{2}$. To evaluate the classification performance we use *accuracy*, which is a measure of correctly classified cases,

$$\eta = \frac{tp + tn}{tp + tn + fn + fp},$$

and *F-measure*, which combines *precision* $\pi = \frac{tp}{tp+fp}$ and *recall* $\rho = \frac{tp}{tp+fn}$,

$$F = \frac{2\pi\rho}{\pi + \rho}.$$

5.2 Reuters Data Set

We used the Reuters-21578 text categorization collection containing the Reuters new stories preprocessed by Karčiauskas (2002). The training set contained 7769 documents and the testing set contained 3018 documents. For every of the ten document classes the most informative features were selected using the expected information gain as a feature selection criteria, and each document class was classified separately against all other classes. We chose to use the same threshold for the expected information gain as in (Vomlel, 2006), the number of selected features varied from 23 for the corn document class to 307 for the earn document class. While learning the values of the hidden parameters the EM algorithm was stopped after 50 iterations. The accuracy and F-measure for causal independence models with the threshold interaction function $k = 1, \ldots, 4$ are given in tables 1 and 2. Even though the threshold to select the relevant features was tuned for the noisy OR classifier, for 5 document classes the causal independence models with other interaction function than logical OR provided better results.

The accuracy and F-measure of the noisy OR model and a few other classifiers on the Reuters data collection reported in (Vomlel, 2006) show the competitive performance of the noisy OR model.

Table 1: Classification accuracy for symmetric causal independence models with the interaction function $\tau_k, k = 1, \ldots, 4$ for Reuters data set; N_{Class} is number of documents in the corresponding class.

Class	N_{Class}	τ_1	τ_2	τ_3	τ_4
earn	1087	96.3	**97.2**	**97.2**	96.8
acq	719	93.1	**93.2**	**93.2**	93.0
crude	189	**98.1**	**98.1**	97.6	97.7
money-fx	179	95.8	95.8	95.9	**96.0**
grain	149	**99.2**	99.0	98.2	97.9
interest	131	96.5	**96.8**	96.7	96.7
trade	117	96.6	97.0	**97.3**	**97.3**
ship	89	**98.9**	98.8	98.7	98.6
wheat	71	**99.5**	99.2	98.8	98.5
corn	56	**99.7**	99.4	99.1	98.8

6 Discussion

In this paper, we discussed the application of symmetric causal independence models for classification. We developed the EM algorithm to learn the parameters in symmetric causal independence models and studied its convergence. The reported experimental results indicate that it is unnecessary to restrict causal independence models to only two interaction functions, logical OR and logical AND. Competitive classification performance of symmetric causal independence models present them as a potentially useful additional tool to the set of classifiers.

The current study has only examined the problem of learning conditional probabilities of hidden variables. The problem of learning an optimal interaction function has not been addressed. Efficient search in symmetric Boolean function space is a possible direction for future research.

Acknowledgments

This research, carried out in the TimeBayes project, was supported by the Netherlands Organization for Scientific Research (NWO) under project number FN4556. The authors are grateful to Gytis Karčiauskas for the preprocessed Reuters data. We would also like to thank Jiří

Table 2: Classification F-measure for symmetric causal independence models with the interaction function $\tau_k, k = 1, \ldots, 4$ for Reuters data set; N_{Class} is number of documents in the corresponding class.

Class	N_{Class}	τ_1	τ_2	τ_3	τ_4
earn	1087	95.0	**96.1**	**96.1**	95.6
acq	719	**85.3**	84.3	84.5	83.8
crude	189	84.5	**85.7**	80.7	81.0
money-fx	179	60.9	62.1	62.6	**62.7**
grain	149	**92.7**	89.9	80.7	77.2
interest	131	40.2	**55.0**	53.3	54.0
trade	117	51.0	61.2	**63.7**	**63.7**
ship	89	**79.5**	77.7	74.5	71.5
wheat	71	**90.3**	81.8	71.4	66.2
corn	56	**91.8**	83.6	72.5	61.5

Vomlel for sharing his code and insights.

References

S. Boyd and L. Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press.

J. Darroch. 1964. On the distribution of the number of successes in independent trials. *The Annals of Mathematical Statistics*, 35:1317–1321.

A.P. Dempster, N.M. Laird and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–38.

F.J. Díez. 1993. Parameter adjustment in Bayes networks. The generalized noisy OR-gate. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, pages 99–105.

H.B. Enderton. 1972. *A Mathematical Introduction to Logic*. Academic Press, San Diego.

D. Heckerman and J.S. Breese. 1994. A new look at causal independence. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pages 286–292.

M. Henrion. 1989. Some practical issues in constructing belief networks. *Uncertainty in Artificial Intelligence*, 3:161–173.

R. Jurgelenaite, T. Heskes and P.J.F. Lucas. 2006. Noisy threshold functions for modelling causal independence in Bayesian networks. *Technical report ICIS–R06014*, Radboud University Nijmegen.

H.J. Kappen and J.P. Neijt. 2002. Promedas, a probabilistic decision support system for medical diagnosis. *Technical report*, SNN - UMCU.

G. Karčiauskas. 2002. Text Categorization using Hierarchical Bayesian Network Classifiers. *Master thesis*, Aalborg University.

L. Le Cam. 1960. An approximation theorem for the Poisson binomial distribution. *Pacific Journal of Mathematics*, 10:1181–1197.

P.J.F. Lucas. 2005. Bayesian network modelling through qualitative patterns. *Artificial Intelligence*, 163:233–263.

C. Meek and D. Heckerman. 1997. Structure and parameter learning for causal independence and causal interaction models. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, pages 366–375.

J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kauffman Publishers.

M.A. Shwe, B. Middleton, D.E. Heckerman, M. Henrion, E.J. Horvitz, H.P. Lehmann and G.F. Cooper. 1991. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base, I – the probabilistic model and inference algorithms. *Methods of Information in Medicine*, 30:241–255.

S. Srinivas. 1993. A generalization of the noisy-or model. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, pages 208–215.

S. Visscher, P.J.F. Lucas, M. Bonten and K. Schurink. 2005. Improving the therapeutic performance of a medical Bayesian network using noisy threshold models. In *Proceedings of the Sixth International Symposium on Biological and Medical Data Analysis*, pages 161–172.

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

I. Wegener. 1987. *The Complexity of Boolean Functions*. John Wiley & Sons, New York.

N.L. Zhang and D. Poole. 1996. Exploiting causal independence in Bayesian networks inference. *Journal of Artificial Intelligence Research*, 5:301–328.