Artificial Intelligence in Medicine (2007) **40**, 45—55

ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Predicting carcinoid heart disease with the noisy-threshold classifier

Marcel A.J. van Gerven [a,*], Rasa Jurgelenaite [a], Babs G. Taal [b], Tom Heskes [a], Peter J.F. Lucas [a]

[a] *Institute for Computing and Information Sciences, Radboud University Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands*
[b] *Netherlands Cancer Institute/Antoni van Leeuwenhoek Hospital, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands*

Received 23 June 2006; received in revised form 20 September 2006; accepted 26 September 2006

KEYWORDS
Carcinoid heart disease;
Bayesian classification;
Causal independence;
Noisy-threshold model

Summary

Objective: To predict the development of carcinoid heart disease (CHD), which is a life-threatening complication of certain neuroendocrine tumors. To this end, a novel type of Bayesian classifier, known as the noisy-threshold classifier, is applied.
Materials and methods: Fifty-four cases of patients that suffered from a low-grade midgut carcinoid tumor, of which 22 patients developed CHD, were obtained from the Netherlands Cancer Institute (NKI). Eleven attributes that are known at admission have been used to classify whether the patient develops CHD. Classification accuracy and area under the receiver operating characteristics (ROC) curve of the noisy-threshold classifier are compared with those of the naive-Bayes classifier, logistic regression, the decision-tree learning algorithm C4.5, and a decision rule, as formulated by an expert physician.
Results: The noisy-threshold classifier showed the best classification accuracy of 72% correctly classified cases, although differences were significant only for logistic regression and C4.5. An area under the ROC curve of 0.66 was attained for the noisy-threshold classifier, and equaled that of the physician's decision-rule.
Conclusions: The noisy-threshold classifier performed favorably to other state-of-the-art classification algorithms, and equally well as a decision-rule that was formulated by the physician. Furthermore, the semantics of the noisy-threshold classifier make it a useful machine learning technique in domains where multiple causes influence a common effect.

* Corresponding author. Tel.: +31 24 365 34 56; fax: +31 24 365 33 56.
E-mail address: marcelge@cs.ru.nl (M.A.J. van Gerven).

0933-3657/$ — see front matter
doi:10.1016/j.artmed.2006.09.003

1. Introduction

Bayesian networks have become a widely accepted formalism for reasoning under uncertainty by providing a concise representation of a joint probability distribution over a set of random variables [1]. This distribution is factorized according to an associated acyclic directed graph (ADG) that represents the independence structure between random variables. However, the construction of a Bayesian network that fully captures this independence structure for a realistic domain, has proven to be a difficult task. It requires either manual specification of the ADG by means of available expert knowledge, or large amounts of high-quality data when we resort to structure learning.

An alternative to the construction of an ADG that fully captures the independence structure that holds between variables within the domain, is to use a fixed or severely constrained graph topology for classification purposes. In the latter context we call a Bayesian network a Bayesian classifier. The use of Bayesian methods in medicine was first proposed by Ledley and Lusted in their classic 1959 paper [2], and one of the first successful implementations of Bayesian classifiers in medicine was De Dombal's system for the diagnosis of acute abdominal pain [3]. The classifier that was used assumes independence of symptoms given the disease, and is known as the naive-Bayes classifier. Over the years, many different Bayesian classifier architectures have been proposed, and many of them focus on lifting the independence assumptions of the naive-Bayes classifier [4]. However, a standard technique such as logistic regression, which is used extensively in medicine, can also be interpreted in terms of a Bayesian classifier architecture (Fig. 1). Other examples of Bayesian classifier architectures can be found in refs. [5–7].

Although, typically, the actual joint probability distribution, and the joint probability distribution that is represented by the Bayesian classifier, differ considerably, this approach can still yield good results with respect to the classification task [8].

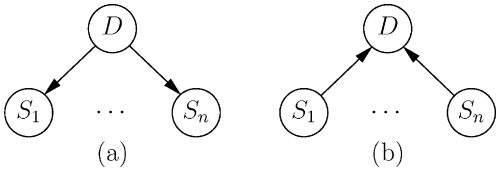

Figure 1 A naive-Bayes classifier (a) is a generative method which models how a disease D leads to symptoms S_i, where symptoms are assumed independent given the disease. Logistic regression (b) is a discriminative method that assumes no such independence and rather assumes that the influences of symptoms combine linearly.

However, a weakness of this approach is that the ad-hoc restrictions that are placed on the underlying graph effectively reduces the Bayesian network to a black box model, making the relation between properties of the domain and classification outcome often difficult to understand. This is an undesirable property; especially in medicine, where ideally one wants to be able to interpret how the classification outcome (such as diagnosed disease or patient prognosis) relates to the available domain knowledge (its causes). The explanation of drawn conclusions is required to increase the acceptance of machine-learning techniques in practice [9,10].

In this paper, we employ a novel Bayesian classifier, introduced in ref. [11], that facilitates this interpretation as it explicitly provides for a semantics in terms of cause and effect relationships [12]. This *noisy-threshold classifier* is based on a generalization of the well-known *noisy-or* model, which has already been used for the purpose of text classification in ref. [13]. In order to demonstrate the merits of the noisy-threshold classifier in a medical context, we apply the technique to the prediction of *carcinoid heart disease*(CHD); a serious condition that arises as a complication of certain neuroendocrine tumors [14]. We demonstrate that the noisy-threshold classifier performs competitively with state-of-the art classification techniques for this medically relevant problem. Furthermore, an expert physician at the Netherlands Cancer Institute (NKI) was consulted, and it is demonstrated how her knowledge concerning CHD relates to the parameters that were estimated for the noisy-threshold classifier.

This paper proceeds as follows. Section 2 introduces the necessary preliminaries and discusses the semantics of the noisy-threshold model, whereas Section 3 describes the medical problem. The use of the noisy-threshold model as a Bayesian classifier is discussed in Section 4. The results on the classification task and the medical interpretation by the expert physician is presented in Section 5. The paper is ended by some concluding remarks in Section 6.

2. Preliminaries

2.1. Bayesian networks

Bayesian networks provide for a compact factorization of a joint probability distribution over a set of random variables by exploiting the notion of *conditional independence* [1]. Conditional independence can be represented by an acyclic directed graph (ADG) G consisting of vertices $V(G)$ and arcs $A(G)$, and relies on the notion of *d-separation* [1]. Let G be an ADG and P a joint probability distribution over a

set of random variables $\mathbf{X} = \{X_1, \ldots, X_n\}$. We assume that there is a one-to-one correspondence between the vertices $V(G)$ and random variables $\mathbf{X}$. In general, we will use X_v to refer to the random variable that corresponds to a vertex v, and use $X_{\mathbf{U}}$ to refer to the set of random variables $\{X_v | v \in \mathbf{U}, \mathbf{U} \subseteq V(G)\}$. A *Bayesian network* is defined as a pair (G, P), such that G admits the following recursive factorization of the joint probability distribution:

$$P(\mathbf{X}) = \prod_{v \in V(G)} P(X_v | X_{\pi_G(v)}) \tag{1}$$

with $\pi_G(v) = \{v' | (v', v) \in A(G)\}$. To simplify notation, we will use vertices $V(G)$ and random variables in $\mathbf{X}$ interchangeably, where the interpretation will be clear from context. We use x to denote an arbitrary element in the sample space Ω_X of a random variable X, and $\mathbf{x}$ for an element in the sample space $\Omega_{\mathbf{X}} = \Omega_{X_1} \times \cdots \times \Omega_{X_n}$ for a set $\mathbf{X} = \{X_1, \ldots, X_n\}$ of random variables.

2.2. Semantics of the noisy-threshold model

In this section, we will show how to arrive at the noisy-threshold model, by introducing a number of assumptions that are motivated by the semantics in terms of causes and effects, that is taken to hold for causal independence models. Causal independence is a popular way to specify interactions among cause variables [1,12,15–17]. The global structure of a causal independence model is shown in Fig. 2; it expresses the idea that causes $\mathbf{C} = \{C_1, \ldots, C_n\}$ influence a common effect E through hidden variables $\mathbf{H} = \{H_1, \ldots, H_n\}$ and a deterministic function f, called the *interaction function*. The causal independence assumption does not refer to independence between causes, but rather to the assumption that hidden variables H_i are independent of causes $\mathbf{C} \backslash \{C_i\}$ given C_i. Causal independence is therefore also known as *independence of causal influence* or *exception independence*. In practice, causes in a causal independence model can be dependent; for instance, when the model is embedded within a larger network, or if there are direct dependencies between causes. However, if causes are completely observed then it is not necessary to model the dependence structure between cause variables.

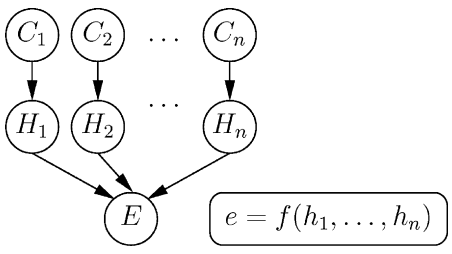

Figure 2 Causal independence model.

In this paper, we assume that causes are either *present* or *absent*. We use x^+ and x^- for $X = \top$ (true) and $X = \bot$ (false), respectively, and interpret $\top$ as 1 and $\bot$ as 0 in an arithmetic context. The individual contribution of a cause C_i to the effect E is realized by the parameter $P(H_i | C_i)$ associated with the hidden variable H_i; if $P(h_i^+ | c_i^+) < 1$ then H_i is said to inhibit the cause C_i. The assumption of *accountability* states that absent causes do not contribute to the effect which implies that $P(h_i^+ | c_i^-) = 0$ [1]. The interaction function f represents in which way the hidden variables H_i, and indirectly also the causes C_i, interact *deterministically* to yield the final effect E. Since variables are binary, f reduces to a Boolean function. It is also useful to introduce a *leak term* whenever it is infeasible to identify all the variables that influence the effect. We model this leak term by postulating a cause C_l, $l \in \{1, \ldots, n\}$, that is always present with which is associated a leak probability $P(h_l^+ | c_l^+)$ [18]. In this manner, we maintain the *closed-world assumption* [19].

It follows from these assumptions that the conditional probability of the effect e^+ given a configuration $\mathbf{c}$ of the causes $\mathbf{C}$ can be obtained from the parameters $P(h_i | c_i)$ as follows [15]:

$$P_f(e^+ | \mathbf{c}) = \sum_{\mathbf{h}:\, f(\mathbf{h})} \prod_{i=1}^{n} P(h_i | c_i), \tag{2}$$

where $P_f(e^+ | \mathbf{h}) = 1 \Leftrightarrow f(\mathbf{h}) = \top$.

As there are 2^{2^n} different n-ary Boolean functions [20,21], the potential number of causal independence models that is admitted by Eq. (2) is huge. However, if we assume that the order of the cause variables does not matter, the Boolean functions become *symmetric* and the number of such functions reduces to 2^{n+1} [21]. An important symmetric Boolean function is the *exact* Boolean function ε_m, which is defined as:

$$\varepsilon_m(h_1, \ldots, h_n) = \top \Leftrightarrow \sum_{j=1}^{n} h_j = m.$$

Any symmetric Boolean function can be decomposed in terms of the exact functions ε_m as follows [21]:

$$f(h_1, \ldots, h_n) = \bigvee_{m=0}^{n} \varepsilon_m(h_1, \ldots, h_n) \wedge \gamma_m \tag{3}$$

where γ_m are Boolean constants dependent on the choice of the symmetric function f. A particularly useful type of symmetric Boolean function is the *threshold* function τ_k, which simply checks whether

there are at least k values $\top$ among the arguments, i.e.:

$$\tau_k(h_1,\ldots,h_n) = \top \Leftrightarrow \sum_{j=1}^{n} h_j \geq k.$$

In terms of causes and effects, the use of the threshold function as the interaction function of a causal independence model expresses the notion that *sufficient* causes should be present in order to induce the effect. Then, the *noisy-threshold model*, as defined in ref. [22], is given by:

$$P_{\tau_k}(e^+|\mathbf{c}) = \sum_{j=k}^{n} \sum_{\mathbf{h}:\varepsilon_j(\mathbf{h})} \prod_{i=1}^{n} P(h_i|c_i). \tag{4}$$

To express a threshold function in terms of Eq. (3), we use $\gamma_0 = \cdots = \gamma_{k-1} = \bot$ and $\gamma_k = \cdots = \gamma_n = \top$. Note that the noisy-or model, with $f(h_1,\ldots,h_n) \Leftrightarrow h_1 \vee \cdots \vee h_n$, corresponds to threshold function τ_1, and the noisy-and model, with $f(h_1,\ldots,h_n) \Leftrightarrow h_1 \wedge \cdots \wedge h_n$, corresponds to threshold function τ_n. Hence, these two commonly used causal independence models are the extremes of a spectrum of causal independence models that are defined by the noisy-threshold function.

2.3. Parameter estimation

The parameters $P(h_i^+|c_i^+)$ of the model can be learned using an *expectation-maximization* (EM) algorithm [23]. EM is a method for finding maximum likelihood estimates of parameters in probabilistic models, where the model depends on (unobserved) hidden variables. Every iteration of an EM algorithm consists of two steps: the expectation step (E-step), which computes the expected value of the hidden variables, and a maximization step (M-step), which computes the maximum likelihood estimates of the parameters given the data.

To learn the parameters in the noisy-threshold classifier, we use the EM algorithm for noisy-threshold models [11]. This EM algorithm is based on the connection between noisy-threshold models and the *Poisson binomial distribution*. Let $\mathbf{p}(\mathbf{c}) = \{p_i | p_i = P(h_i^+|c_i), i = 1,\ldots,n\}$. Then

$$\mathrm{B}(l;\mathbf{p}(\mathbf{c})) = \left\{\prod_{i=1}^{n}(1-p_i)\right\} \sum_{1\leq j_1<\cdots<j_l\leq n} \prod_{z=1}^{l} \frac{p_{j_z}}{1-p_{j_z}}$$

is the *Poisson binomial probability* [24,25], where l denotes the number of successes in n independent trials. The following connection holds between the conditional probabilities in the noisy-threshold model and the Poisson binomial distribution:

$$P_{\tau_k}(e^+|\mathbf{c}) = \sum_{i=k}^{n} \mathrm{B}(i;\mathbf{p}(\mathbf{c})).$$

An analysis of this connection, as well as computationally efficient methods to compute, approximate, or bound this probability distribution, can be found in ref. [22].

Let the data set $\mathcal{D} = \{\mathbf{u}^1,\ldots,\mathbf{u}^N\}$ be a multiset, where instances $\mathbf{u}^j = \{\mathbf{c}^j, e^j\} = \{c_1^j,\ldots,c_n^j, e^j\}$ with $j = 1,\ldots,N$ consist of realizations of causes and the effect. Let $\mathcal{D}^+ \subseteq \mathcal{D}$ denote those instances $\{\mathbf{c}^j, e^j\}$ for which $e^j = \top$, and let $\mathcal{D}^- \subseteq \mathcal{D}$ denote those instances $\{\mathbf{c}^j, e^j\}$ for which $e^j = \bot$. We use $\boldsymbol{\theta} = \{\theta_i | \theta_i = P(h_i^+|c_i^+), i = 1,\ldots,n\}$ to denote the parameters of the noisy-threshold model. Then, on the $(z+1)$th iteration, EM proceeds as follows.

E-step. For every instance, $\mathbf{u}^j = \{\mathbf{c}^j, e^j\}$ with $j = 1,\ldots,N$, set

$$\mathbf{p}^{(z,j)} = \{p_i^{(z,j)} | p_i^{(z,j)} = \theta_i^{(z)} c_i^j, i = 1,\ldots,n\}. \tag{5}$$

Subsequently, we compute the probability

$$P(h_m^+|\mathbf{u}^j,\boldsymbol{\theta}^{(z)}) = \begin{cases} \dfrac{p_m^{(z,j)} \displaystyle\sum_{i=m-1}^{n-1} \mathrm{B}(i;\mathbf{p}_{\setminus m}^{(z,j)})}{\displaystyle\sum_{i=m}^{n} \mathrm{B}(i;\mathbf{p}^{(z,j)})}, & \text{if } \mathbf{u}^j \in \mathcal{D}^+ \\ \dfrac{p_m^{(z,j)} \left(1 - \displaystyle\sum_{i=m-1}^{n-1} \mathrm{B}(i;\mathbf{p}_{\setminus m}^{(z,j)})\right)}{1 - \displaystyle\sum_{i=m}^{n} \mathrm{B}(i;\mathbf{p}^{(z,j)})}, & \text{if } \mathbf{u}^j \in \mathcal{D}^- \end{cases} \tag{6}$$

where $\mathbf{p}_{\setminus m}^{(z,j)} = \{p_i^{(z,j)} | i = 1,\ldots,n, i \neq m\}$ for hidden variables H_m, with $m = 1,\ldots,n$.

M-step. Update the parameter estimates for all $i = 1,\ldots,n$:

$$\theta_i^{(z+1)} = \frac{\sum_{j=1}^{N} P(h_i^+|\mathbf{u}^j,\boldsymbol{\theta}^{(z)})}{\sum_{j=1}^{N} c_i^j}. \tag{7}$$

Generally, the expectation and maximization steps are alternated repeatedly until convergence. However, for small data sets, this may result in overfitting artifacts; an issue to which we return in Section 4.1.

The analysis in this section has shown that causal independence models such as the noisy-threshold model have an interesting semantics in terms of causes and effect, and can be learned using the EM algorithm, given a symmetric Boolean interaction function. The next section describes the

medical problem that is used to illustrate the usefulness of the noisy-threshold model as a classifier.

3. Carcinoid heart disease

Carcinoid tumors belong to the group of neuroendocrine tumors, which are known for the production of vasoactive agents in the presence of metastatic disease; usually hepatic (liver) metastases. Among these agents, serotonin is the most important agent, leading to the characteristic carcinoid syndrome of flushes and diarrhea. The other main characteristic feature of neuroendocrine tumors is the slow progression of most tumors if the histology shows a low-grade pattern [26].

Serotonin overproduction may also cause carcinoid heart disease (CHD), which is characterized by fibrosis of the right sided heart valves as shown in Fig. 3. Fibrosis induces thickening and retraction of the tricuspid valve, leading to tricuspid insufficience and ultimately heart failure, which is the cause of death in as much as one-half of carcinoid patients [14,27]. Since so many carcinoid patients die of CHD, it is important to distinguish patients that are admitted to the clinic into patients that are prone to develop a severe form of carcinoid heart disease, and those that do not develop this severe form. In this way, patients that are at risk can be given more aggressive treatment in order to reduce the probability of the development of CHD. Hence, the classification task for this medical problem will be to classify the patients into these two groups, depending on the attributes that are known at the time of admission to the clinic. We use chd^+ to denote the development of moderate to extreme tricuspid valve insufficience and chd^- to denote the absence, or development of mild tricuspid valve insufficience during patient follow-up.

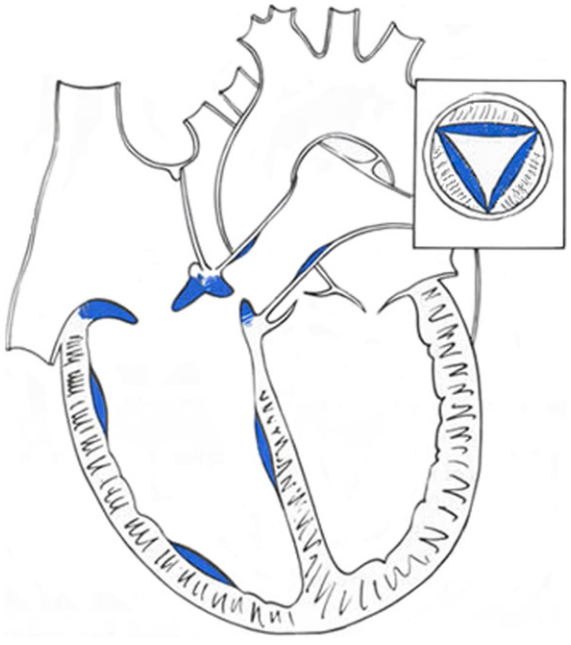

Figure 3 CHD is characterized by heart valve fibrosis as shown in the overlay.

In principle, the physician can make use of the attributes that are measured at admission (Table 1), in order to predict the development of CHD. However, in practice, in order to determine the probability of developing moderate to severe tricuspid valve insufficience, the physician makes use of the following decision rule:

$$P(\text{chd}^+|\mathbf{c}) = \begin{cases} 0.50, \text{if } \text{HIA}^+ \wedge \text{Dia}^+ \wedge \text{HMe}^+ \\ 0.25, \text{if } \text{HIA}^+ \wedge (\text{Dia}^- \wedge \text{HMe}^+ \vee \text{Dia}^+ \wedge \text{HMe}^-) \\ 0.10, \text{if } \text{HIA}^+ \wedge \text{Dia}^- \wedge \text{HMe}^- \vee \text{HIA}^- \wedge \text{Dia}^+ \wedge \text{HMe}^+ \\ 0.03, \text{otherwise}. \end{cases}$$

The aim of this paper is to show that a noisy-threshold model can be used as a Bayesian classifier, where performance is compared both with the physician's classification performance, as well as with standard classification techniques such as the naive-Bayes classifier, logistic regression and decision-trees. The patient attributes are used as cause variables in the definition of a noisy-threshold model, and it is assumed that independence of causal influence, accountability, symmetry and sufficiency hold. As required, variables are binary, and positive states of variables are perceived to be less favorable than negative states, such that they could be responsible for carcinoid heart disease. To train and test Bayesian classifiers for this medical problem, we have used a clinical database consisting of 54 patients that suffered from a neuroendocrine tumor, and for which the grade of tricuspid valve insufficience was known. Twenty-two patients developed moderate or worse tricuspid valve insufficience during follow-up.

Table 1 Patient attributes that are measured at admission

Name	Definition	Name	Definition
HIA	5-HIAA levels	GIL	General illness
CGA	Chromogranin A levels	BOB	Bowel obstruction
DIA	Diarrhea	IBL	Internal bleeding
WHE	Wheezing	FEV	Fever
FLU	Flushing	HME	Hepatic metastases
APA	Abdominal pain		

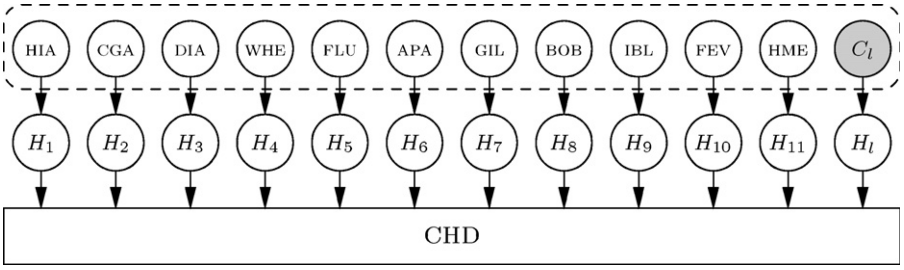

Figure 4 A noisy-threshold model for carcinoid heart disease, where the dashed region represents the total tumor burden for the patient. Note the use of the leak cause C_l in order to model possible hidden causes.

We have not yet touched upon the most important assumption of causal independence models. That is, can the variables be regarded as causes of carcinoid heart disease? For some attributes this is questionable. Diarrhea, for instance, is a symptom of other processes and is therefore not likely to be a cause of carcinoid heart disease. However, we *can* interpret the attributes as risk factors that act as components of the total *tumor burden*, as depicted in Fig. 4. Since the causes are assumed to be completely observed, we refrain from adding additional dependencies between cause variables.

4. The noisy-threshold classifier

4.1. Classifier construction

Construction of a noisy-threshold classifier (NTC) proceeds as follows. We first determine the cause variables $\mathbf{C}$ and effect variable E that are used in the classifier. In the context of a classifier, the cause variables stand for the attributes and the effect variable stands for the class-variable. Secondly, we need to determine the positive states of the variables. In the CHD domain, the positive states are simply defined as the presence of attributes that affect the presence of the class-variable CHD. Once the cause and effect variables have been defined, we need to find both the optimal values for the parameters $P(h_i^+|c_i^+)$ using the EM algorithm of Section 2.3, as well as the correct threshold function τ_k.

To this end, we define the following measures with respect to a fixed database $\mathcal{D}$ and model M. Let the *true positives* (tp) stand for the number of instances $\mathbf{u}^j \in \mathcal{D}^+$ for which $P(e^+|\mathbf{c}^j) \geq 0.5$ and let the *false negatives* (fn) stand for the number of instances $\mathbf{u}^j \in \mathcal{D}^+$ for which $P(e^+|\mathbf{c}^j) < 0.5$. Likewise, we define the *true negatives* (tn) as the number of instances $\mathbf{u}^j \in \mathcal{D}^-$ for which $P(e^+|\mathbf{c}^j) < 0.5$ and the *false positives* (fp) as the number of instances $\mathbf{u}^j \in \mathcal{D}^-$ for which $P(e^+|\mathbf{c}^j) \geq 0.5$. In order to learn the parameters of the noisy-threshold model, we used a training set $\mathcal{D}_{\text{train}}$ and a validation set $\mathcal{D}_{\text{validate}}$. The validation set is used to counterbalance the overfitting that may occur when learning model parameters. The aim of the learning phase is to maximize both the *classification accuracy*

$$\eta(\mathcal{D}) = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{tn} + \text{fn} + \text{fp}}$$

as a measure of the number of correctly classified cases, and the F_1 *measure*

$$F_1(\mathcal{D}) = \frac{2\pi\rho}{\pi + \rho}$$

as a measure that takes into account the tradeoff between *precision* $\pi = \text{tp}/(\text{tp} + \text{fp})$ and *recall* $\rho = \text{tp}/(\text{tp} + \text{fn})$, which is also known as *sensitivity*. We use these two measures since the accuracy is the obvious measure but may convey the wrong intuition when the classes are not equal in size [28]. Finding the optimal noisy-threshold classifier then proceeds as follows:

(1) Divide the data set $\mathcal{D}$ into the disjoint sets $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{validate}}$ and $\mathcal{D}_{\text{test}}$.
(2) For all noisy-threshold models $P_{\tau_1}, \ldots, P_{\tau_n}$ with $n = |\mathbf{C}|$, use the training data $\mathcal{D}_{\text{train}}$ and the EM-algorithm of Section 2.3 to learn the parameters $P(h_i^+|c_i^+)$.
(3) Select the noisy-threshold model and the number of iterations of the EM-algorithm that maximizes $w_1 \cdot \eta(\mathcal{D}_{\text{validate}}) + w_2 \cdot F_1(\mathcal{D}_{\text{validate}})$ with equal weights $w_1 = w_2$, as the optimal noisy-threshold classifier.

With regard to the clinical data set $\mathcal{D}$, we have used a leave-one-out cross-validation scheme to implement the above algorithm. $\mathcal{D}$ contains too many missing values to simply remove the instances that contain missing data. We have used *mean substitution* [29] as an imputation scheme, and note that *multiple imputation* [30] produced similar results. Let N_i be the number of data samples without missing

data for the variable C_i for all $i = 1, \ldots, n$. If C_i is missing in the sample j then we replace c_i^j in Eqs. (5) and (7) by the estimate

$$\mu_i = \frac{1}{N_i} \sum_{k=1}^{N_i} c_i^k$$

of the prior $P(c_i^+)$.

4.2. Classifier evaluation

In order to evaluate the performance of the noisy-threshold classifier, we compared its classification accuracy with the accuracy of a number of other well-known algorithms. For the comparison we have used the naive-Bayes classifier (NBC), logistic regression (LG) and a decision-tree learning algorithm (C4.5) as implemented by the WEKA machine learning tool [31].[1] Furthermore, we compare the performance of the optimal noisy-threshold classifier with that of the noisy-or classifier P_{τ_1} as a special case [13]. For the naive-Bayes classifier, the posterior probability of developing carcinoid heart disease is given by

$$P(e^+|\mathbf{c}) = P(e^+) \prod_{i=1}^{n} P(c_i|e^+)$$

and for logistic regression, the posterior probability of developing carcinoid heart disease is given by

$$P(e^+|\mathbf{c}) = \frac{1}{1 + e^{-(a_0 + a_1 c_1 + \cdots + a_n c_n)}}$$

where the parameters are estimated from data.

Classification proceeds by selecting the class value that has highest posterior probability. For the decision-tree learning algorithm, classification proceeds by traversing the tree and selecting the class value that is associated with the leaf node. Hence, no posterior probability $P(e^+|\mathbf{c})$ is computed.

As pointed out in ref. [32], when comparing two classification algorithms, the approach preferred to a standard t-test, is to use a binomial test, which uses the number of cases n for which the two classifiers produce a different output, and the number of cases s where the output of the examined classifier was correct, while the output of the reference classifier was wrong. Under the null hypothesis that the two classifiers perform equally well, we compute:

$$q = \sum_{i=s}^{n} \frac{n!}{i!(n-i)!} (0.5)^n$$

for a one-tailed test, and $p = 2q$ for a two-tailed test.

Since the classification accuracy assumes equal costs between false positives and false negatives, we use the *receiver operating characteristics* (ROC) curve to compare the performance of some of the classifiers in terms of the trade off between *sensitivity* $\rho = \mathrm{tp}/(\mathrm{tp} + \mathrm{fn})$ and *specificity* $\sigma = \mathrm{tn}/(\mathrm{tn} + \mathrm{fp})$ for every possible cutoff [33], where ρ is shown on the y-axis, and $1 - \sigma$ is shown on the x-axis. This performance can be quantified by computing the area under the ROC curve (AUC), which has been shown to equal the outcome of the Mann–Whitney U statistic [34]:

$$\mathrm{AUC} = \frac{\sum_{\mathbf{c}^i \in \mathcal{D}^+} \sum_{\mathbf{c}^j \in \mathcal{D}^-} u(\mathbf{c}^i, \mathbf{c}^j)}{|\mathcal{D}^+||\mathcal{D}^-|}$$

where

$$u(\mathbf{c}^i, \mathbf{c}^j) = \begin{cases} 1, & \text{if } P(e^+|\mathbf{c}^i) > P(e^+|\mathbf{c}^j) \\ \frac{1}{2}, & \text{if } P(e^+|\mathbf{c}^i) = P(e^+|\mathbf{c}^j) \\ 0, & \text{if } P(e^+|\mathbf{c}^i) < P(e^+|\mathbf{c}^j) \end{cases}$$

We can interpret this statistic as follows. We assume that there is a ranking between instances in $\mathcal{D}$ such that any deviation from the perfect ranking that ranks all positive examples higher than all negative examples leads to a decrease in the AUC [35]. If $P(e^+|\mathbf{c}^i) > P(e^+|\mathbf{c}^j)$ then we produce a correct ranking, if $P(e^+|\mathbf{c}^i) = P(e^+|\mathbf{c}^j)$ then we break ties at random and produce a correct ranking one-half of the time, and if $P(e^+|\mathbf{c}^i) < P(e^+|\mathbf{c}^j)$ then we produce an incorrect ranking.

5. Results

5.1. Classification performance

Table 2 lists the classification accuracy for noisy-threshold classifiers P_{τ_1} to $P_{\tau_{12}}$. The noisy-threshold classifier P_{τ_6} is selected, based on the validation set $\mathcal{D}_{\text{validate}}$, and shows the best classification accuracy of 0.72 on the test set $\mathcal{D}_{\text{test}}$. Note that this exceeds considerably the classification accuracy of 0.54 for the noisy-or classifier P_{τ_1}.

Table 2 Classification accuracy on $\mathcal{D}_{\text{test}}$ for noisy-threshold classifiers $P_{\tau_1}, \ldots, P_{\tau_{12}}$

NTC	$\eta(\mathcal{D}_{\text{test}})$	NTC	$\eta(\mathcal{D}_{\text{test}})$	NTC	$\eta(\mathcal{D}_{\text{test}})$
P_{τ_1}	0.54	P_{τ_5}	0.69	P_{τ_9}	0.59
P_{τ_2}	0.65	P_{τ_6}	**0.72**	$P_{\tau_{10}}$	0.59
P_{τ_3}	0.65	P_{τ_7}	0.65	$P_{\tau_{11}}$	0.59
P_{τ_4}	0.70	P_{τ_8}	0.57	$P_{\tau_{12}}$	0.59

[1] We use WEKA's default parameter settings; the default imputation method is to interpret a missing value for X as a separate value $x \in \Omega_X$.

Table 3 Classification accuracy and p-values for classification of $\mathcal{D}_{\text{test}}$

Classifier	$\eta(\mathcal{D}_{\text{test}})$	p
Physician	0.69	7.0×10^{-1}
NBC	0.63	2.3×10^{-1}
LG	0.67	6.3×10^{-1}
C4.5	0.44	6.2×10^{-5}
Noisy-or	0.54	6.4×10^{-3}

In order to test how well the NTC performs compared with the physician, and with the other classification algorithms that were discussed in Section 4.2, we have determined the classification accuracy. Table 3 describes the classification accuracy on $\mathcal{D}_{\text{test}}$ for the physician, NBC, LG, C4.5 and noisy-or, and p-values for the null-hypothesis that the classifier accuracy is comparable to that of the NTC P_{τ_6}.

Note that the expert physician's classification accuracy is reasonably high, outperforming all but the noisy-threshold classifier. The noisy-threshold classifier P_{τ_6} shows the best classification accuracy, although the difference is significant only for C4.5 and the noisy-or classifier at a confidence level of $p = 0.05$. For the physician's decision rule, the naive-Bayes classifier, and logistic regression, we cannot reject the null hypothesis that the algorithms may in fact be equally accurate for this data set.

It is well-known that classifiers that show large bias tend to outperform classifiers that show high variance for small data sets, since this reduces the risk of overfitting. For this reason, the naive-Bayes classifier tends to perform well on many data sets [36]. However, although not always reflected in its classification accuracy [8], the assumption of independence between attributes given the class-variable, is a strong assumption which does not hold in general. In contrast, the noisy-threshold classifier's assumptions are motivated by a cause—effect semantics as described in Section 2, and hold for domains where the presence of a sufficient number of causes is sufficient to induce the effect.

Fig. 5 presents the ROC curves for the physician, the noisy-threshold classifier P_{τ_6}, the naive-Bayes classifier and logistic regression, where the area under the curve equals 0.66, 0.66, 0.60 and 0.59, respectively. Although the performance in terms of AUC is mediocre, both the physician's decision rule, and the noisy-threshold classifier show a considerably better performance than the other standard classification techniques. The ROC curve does demonstrate a potential danger of using the noisy-threshold classifier, especially when the causal assumptions are not satisfied. Whereas the naive-Bayes classifier is able to gradually increase the true positive rate at the expense of increasing the true negative rate, the noisy-threshold classifier fails to accomplish this for all true positive rates. This is a consequence of the model assumption that absent causes cannot contribute to the effect; the probability $P_{\tau_k}(e^+|\mathbf{c}^i)$ of assigning an instance to the positive class equals zero whenever the number of present causes is less than the chosen threshold k.

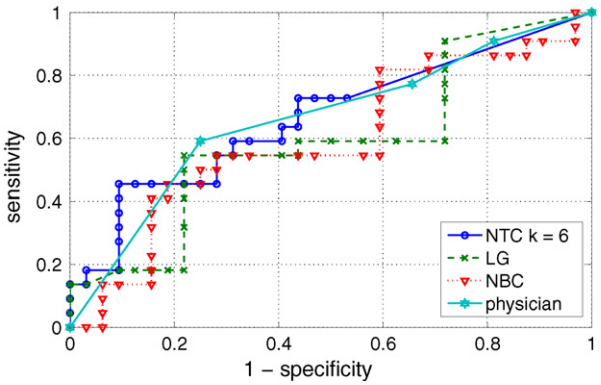

Figure 5 ROC curve for the noisy-threshold classifier, logistic regression, and the naive-Bayes classifier, where the straight line segment in the NTC curve is a consequence of the model assumption that absent causes do not contribute to the effect.

5.2. Medical interpretation

In this section, we look at the noisy-threshold classifier for CHD from a medical point of view. Prior to presenting the resulting classifier, we have asked the physician to indicate how important the individual attributes were felt to be with respect to predicting the development of carcinoid heart disease.

According to the physician, progressive carcinoid disease is often accompanied by the carcinoid syndrome, which is characterized by diarrhea (DIA) caused by increased bowel motility due to serotonin overproduction, by periodical flushing attacks (FLU) due to the synergistic interaction between various vasoactive agents, and sometimes by wheezing (WHE). As discussed in Section 3, serotonin overproduction is thought to play a key role in the etiology of CHD and it can be measured indirectly by means of the urinary 5-HIAA level (HIA) since this is a metabolite of serotonin. Hence, the variables related to the carcinoid syndrome are indicative of serotonin overproduction and ultimately CHD. It is therefore assumed that the variables HIA, DIA, FLU and to a lesser extent WHE have a high predictive value. Serotonin overproduction is itself caused by

the carcinoid tumor in the presence of particular metastases; hormones released by carcinoid tumors are often destroyed by the liver before they reach the general circulation to cause symptoms. Therefore, only hepatic metastases (HME), or metastases that can release hormones directly into the general circulation, can produce the carcinoid syndrome. According to the physician, the presence of hepatic metastases (HME) during hospitalization is indicative of CHD development, since this is a requirement for serotonin overproduction. The plasma chromogranin A (CGA) level is used as a general marker of neuroendocrine activity and tumor extensiveness [37]. Although not regarded as important as the previously discussed attributes, the physician expected CGA to have a high predictive value since extensive tumors with high neuroendocrine activity are more likely to cause CHD. In contrast, the variables IBL, FEV, APA and BOB were not thought to predict CHD very well. Local progression of hypervascular primary tumors into the lumen of the small bowel is often the cause of internal bleeding (IBL), but is not thought to be related to metastatic disease. Fever (FEV) can be caused by hepatic metastases, as captured by the variable HME, but it is also a non-specific symptom that is not necessarily caused by carcinoid disease in the first place. Abdominal pain (APA) and bowel obstruction (BOB) are often caused by complications due to the primary tumor and were assumed to be unrelated to the development of CHD. According to the physician, general illness (GIL) could be indicative of the development of carcinoid heart disease; a poor condition is often due to extensive metastases and therefore a high probability of serotonin overproduction. In general, the physician expected that at least some of the risk factors should occur together in order to cause CHD.

Fig. 6 depicts the actual estimates of prior probabilities $P(c_i^+)$ and conditional probabilities $P(h_i^+|c_i^+)$, for the noisy-threshold classifier that was used for predicting CHD. The predictive value of the variables HIA, DIA, FLU and WHE is reflected in the reasonably high associated probabilities $P(h_i^+|c_i^+)$ with $i \in \{1,3,4,5\}$, which range from 0.67 to 0.91, where wheezing is indeed seen to be of less predictive value than the other attributes. The presence of hepatic metastases (HME) is also an important predictor of CHD, as is indicated by the high probability $P(h_{11}^+|c_{11}^+) = 0.92$. Notice that most patients that are admitted already present with such metastases, which is reflected by the high prior probability $P(c_{11}^+) = 0.78$. Contrary to the physician's expectations, CGA was not a very good predictor of CHD, with $P(h_i^+|c_i^+) = 0.53$. In hindsight, this may be explained by the fact that CGA overproduction does not necessarily reflect serotonin overproduction, and if it does, it may be redundant information given that we have observed HIA, which is a metabolite of serotonin. Internal bleeding (IBL) and fever (FEV), with $P(h_i^+|c_i^+) = 0.12$ and $P(h_i^+|c_i^+) = 0.13$, respectively, did not contribute much to the effect. Unexpectedly, both abdominal pain (APA) and bowel obstruction (BOB) had relatively high probability values $P(h_i^+|c_i^+)$ of 0.80 and 0.84, respectively. After some deliberation, the physician gave the following possible explanation. Since abdominal pain and bowel obstruction are often caused by complications due to the primary tumor, both APA and BOB indicate a midgut tumor with possible mesenterial fibrosis. A midgut localization is a prerequisite for serotonin overproduction, and mesenterial fibrosis is thought to be related to tricuspid valve fibrosis [38], and therefore, the presence of these variables could have been indicative of the development of CHD. General illness (GIL) had a high probability value of $P(h_i^+|c_i^+) = 0.93$. Five out of seven patients that suffered from general illness indeed developed CHD. The threshold function τ_6 corresponds to the physician's assessment that the presence of just one risk factor is generally insufficient to cause CHD,

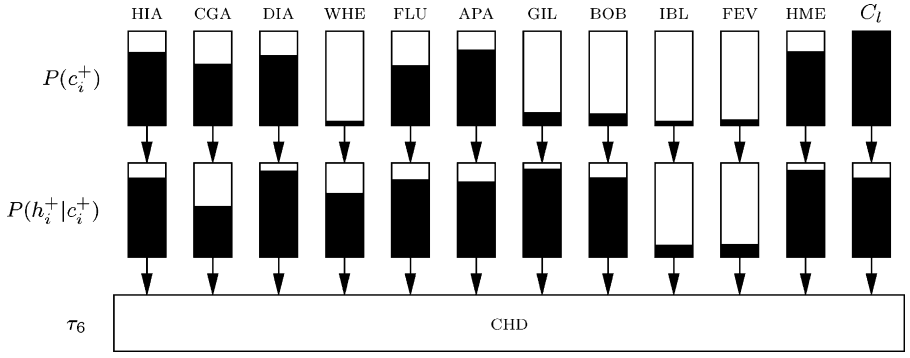

Figure 6 Estimates of priors $P(c_i^+)$, and conditional probabilities $P(h_i^+|c_i^+)$, for the noisy-threshold classifier with threshold function τ_6.

whereas the presence of all risk factors is much too strict a requirement as a cause for CHD; demonstrating that the noisy-threshold model as a generalization of both the noisy-or and noisy-and model can be the proper choice for realistic domains.

6. Conclusions

The noisy-threshold classifier is a novel type of classifier that has a well-defined semantics in terms of causes and effect. Due to the independence assumptions that are made by the classifier, parameters can be reliably estimated without needing to resort to huge amounts of data. This is an important feature since many domains are characterized by limited amounts of data, as discussed in ref. [39]. Learning Bayesian classifiers from data is to be contrasted with the construction of a full Bayesian network that captures available domain knowledge, which, although possible, can be very resource intensive for realistic domains.

We have demonstrated that the noisy-threshold classifier performs comparably with the decision rule that is used by an expert physician, and competitively with state-of-the-art classifiers, on an important classification task in oncology. Furthermore, it significantly outperforms the noisy-or classifier, as a special case of the noisy-threshold classifier, for this data set. The semantics of the noisy-threshold classifier enables an interpretation in terms of available domain knowledge, as is illustrated by the physician's interpretation of classifier parameters. Nevertheless, one should be cautious when defining the positive states of the cause variables since negative states cannot contribute to the effect, as reflected by the straight line segment of the ROC curve. The competitive classification performance and well-defined semantics make the noisy-threshold classifier a promising new machine learning technique, as was demonstrated here in the context of medical prognosis. Currently, the technique is being applied in the context of the analysis of gene expression data.

Various extensions to the noisy-threshold classifier are possible, that increase its applicability. One extension would be to incorporate graded or continuous variables that allow a more natural representation of risk factors such as abdominal pain or fever. Another extension would be the incorporation of time for the noisy-threshold model, analogous to the generalization of noisy-or models to temporal noisy-or models as was realized in ref. [40], and applied to modeling the spread of nasopharyngeal cancer [41]. Furthermore, lifting the assumption of independence of causal influence by allowing multiple causes to influence the same hidden variables may lead to more realistic models. We leave these extensions as topics for further research.

Acknowledgements

This research was sponsored by the Netherlands Organization for Scientific Research (NWO) under grant numbers 612.066.201 and FN4556. We would like to thank the anonymous reviewers for their valuable comments.

References

[1] Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference, 2nd edition, San Francisco, CA: Morgan Kaufmann; 1988.
[2] Ledley R, Lusted L. Reasoning foundation of medical diagnosis: symbolic logic, probability, and value theory aid our understanding of how physicians reason. Science 1959;130:9—21.
[3] de Dombal FT, Leaper D, Staniland J, Horrocks J, McCann A. Computer aided diagnosis of acute abdominal pain. Br Med J 1972;2:9—13.
[4] Spiegelhalter D, Knill-Jones R. Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology. J R Stat Soc 1984;147:35—77.
[5] Sahami M. Learning limited dependence Bayesian classifiers. In: Second international conference on knowledge discovery in databases. Portland, OR: AAAI Press; 1996. p. 335—8.
[6] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Machine Learn 1997;29:131—63.
[7] Cheng J, Greiner R. Comparing Bayesian network classifiers. In: Proceedings of the fifteenth conference on uncertainty in artificial intelligence. Stockholm: Morgan Kaufmann; 1999. p. 101—7.
[8] Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learn 1997;29:103—30.
[9] Teach R, Shortliffe E. An analysis of physician attitudes regarding computer-based clinical consultation systems. Comput Biomed Res 1981;14:542—58.
[10] Lacave C, Díez F. A review of explanation methods for Bayesian networks. Knowledge Eng Rev 2002;17(2):107—27.
[11] Jurgelenaite R, Heskes T. EM algorithm for symmetric causal independence models. In: Proceedings of the seventeenth European conference on machine learning. Heidelberg, Germany: Springer-Verlag; 2006. p. 234—45.
[12] Heckerman D, Breese J. A new look at causal independence. In: Proceedings of the tenth conference on uncertainty in artificial intelligence. San Francisco, CA: Morgan Kaufmann; 1994. p. 286—92.
[13] Vomlel J. Exploiting functional dependence in Bayesian network inference. In: Proceedings of the eighteenth conference on uncertainty in artificial intelligence. San Francisco, CA: Morgan Kaufmann; 2002. p. 528—35.
[14] Zuetenhorst J, Bonfrer J, Korse C, Bakker R, van Tinteren H, Taal BG. Carcinoid heart disease: the role of urinary 5-HIAA excretion and plasma levels of TGF- β and FGF. Cancer 2003;97:1609—15.

[15] Zhang N, Poole D. Exploiting causal independence in Bayesian network inference. J Artif Intell Res 1996;5:301—28.
[16] Díez F. Parameter adjustment in Bayes networks. The generalized noisy OR-gate. In: Proceedings of the ninth conference on uncertainty in artificial intelligence. San Francisco, CA: Morgan Kaufmann; 1993. p. 99—105.
[17] Lucas P. Bayesian network modelling by qualitative patterns. Artif Intell 2005;163:233—63.
[18] Pradham M, Provan G, Middleton B, Henrion M. Knowledge engineering for large belief networks. In: de Mantaras RL, Poole D, editors. Proceedings of the tenth conference on uncertainty in artificial intelligence. San Fransisco, CA: Morgan Kaufmann; 1994. p. 484—90.
[19] Reiter R. On closed-world data bases. In: Gallaire H, Minker J, editors. Logic and databases. New York, NY: Plenum Press; 1978. p. 55—76.
[20] Enderton H. A mathematical introduction to logic. New York, NY: Academic Press, Inc.; 1972.
[21] Wegener I. The complexity of boolean functions. New York, NY: John Wiley & Sons; 1987.
[22] Jurgelenaite R, Heskes T, Lucas P. Noisy threshold functions for modelling causal independence in Bayesian networks. Tech. Re ICIS-R06014. Nijmegen, The Netherlands: Radboud University; 2006.
[23] Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc 1977;39: 1—38.
[24] Edwards A. The meaning of binomial distribution. Nature 1960;186:1074—6.
[25] Cam LL. An approximation theorem for the Poisson binomial distribution. Pacific J Math 1960;10:1181—97.
[26] Zuetenhorst J, Taal B. Metastatic carcinoid tumors: a clinical review. Oncologist 2005;10(2):123—31.
[27] Zuetenhorst J, Taal B. Carcinoid heart disease. New Engl J Med 2003;348:2359—61.
[28] van Rijsbergen C. Information retrieval, 2nd edition, London, UK: Butterworths; 1979.
[29] Kline R. Principles and practice of structural equation modeling. New York, NY: Guilford; 1998.
[30] Rubin D. Multiple imputation for nonresponse in surveys. New York, NY: Wiley; 1987.
[31] Witten I, Frank E. Data mining: practical machine learning tools and techniques, 2nd edition, San Francisco, CA: Morgan Kaufmann; 2005.
[32] Salzberg S. On comparing classifiers: pitfalls to avoid and a recommended approach. Data Mining and Knowledge Discovery 1997;1:317—27.
[33] Egan J. Signal detection theory and ROC analysis. New York, NY: Academic Press; 1975.
[34] Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J Math Psychol 1975;12:387—415.
[35] Cortes C, Mohri M. AUC optimization vs. error rate minimization. In: Thrun S, Saul L, Schölkopf B, editors. Advances in neural information processing systems, vol. 16. Cambridge, MA: MIT Press; 2004.
[36] Kohavi R, Wolpert DH. Bias plus variance decomposition for zero-one loss functions. In: Saitta L, editor. Machine learning: proceedings of the thirteenth international conference. San Mateo, CA: Morgan Kaufmann; 1996. p. 275—83.
[37] Nobels F, Kwekkeboom D, Bouillon R, Lamberts S. Chromogranin A: its clinical values as marker of endocrine tumours. Eur J Clin Invest 1998;28:431—40.
[38] Modlin I, Shapiro M, Kidd M. Carcinoid tumors and fibrosis: an association with no explanation. Am J Gastroenterol 2004;99:2466—78.
[39] van Gerven M, Lucas P. Using background knowledge to construct Bayesian classifiers for data-poor domains. In: Bramer M, Coenen F, Allen T, editors. Proceedings of AI-2004, the twenty-fourth SGAI international conference on innovative techniques and applications of artificial intelligence. London, UK: Springer-Verlag; 2004. p. 269—82.
[40] Galán S, Díez F. Networks of probabilistic events in discrete time. Int J Approx Reason 2002;30:181—202.
[41] Galán S, Aguado F, Díez F, Mira J. Nasonet, modeling the spread of nasopharyngeal cancer with networks of probabilistic events in discrete time. Artif Intell Med 2002;25(3): 247—64.