

Evidence and Scenario Sensitivities in Naive Bayesian Classifiers

Silja Renooij and Linda C. van der Gaag

Department of Information and Computing Sciences, Utrecht University

P.O. Box 80.089, 3508 TB Utrecht, The Netherlands

{silja,linda}@cs.uu.nl

Abstract

Empirical evidence shows that naive Bayesian classifiers perform quite well compared to more sophisticated network classifiers, even in view of inaccuracies in their parameters. In this paper, we study the effects of such parameter inaccuracies by investigating the sensitivity functions of a naive Bayesian classifier. We demonstrate that, as a consequence of the classifier's independence properties, these sensitivity functions are highly constrained. We investigate whether the various patterns of sensitivity that follow from these functions support the observed robustness of naive Bayesian classifiers. In addition to the standard sensitivity given the available evidence, we also study the effect of parameter inaccuracies in view of scenarios of additional evidence. We show that the standard sensitivity functions suffice to describe such scenario sensitivities.

1 Introduction

Bayesian networks are often employed for classification tasks where an input instance is to be assigned to one of a number of output classes. The actual classifier then is a function which assigns a single class to each input, based on the posterior probability distribution computed from the Bayesian network for the output variable. Such classifiers are often built upon a naive Bayesian network, consisting of a single class variable and a number of observable feature variables, each of which is modelled as being independent of every other feature variable given the class variable. The numerical parameters for such a naive network are generally estimated from data and inevitably are inaccurate.

Experiments have shown time and again that classifiers built on naive Bayesian networks are quite stable: they are competitive with other learners, regardless of the size and quality of the data set from which they are learned (Domingos & Pazzani, 1997; Rish, 2001; Liu et al., 2005). Apparently, inaccuracies in the parameters of the underlying network do not significantly affect the performance of such a naive classifier.

The observed stability of naive Bayesian clas-

sifiers may be attributed to (a combination of) properties of the data and of the classifier function. The commonly used winner-takes-all rule, which assigns an instance to a class which is most probable according to the underlying Bayesian network, for example, seems to contribute to the naive classifier's success (Domingos & Pazzani, 1997). The observed stability may also be attributed to the naive independence properties of the classifier. It has been shown, for example, that naive Bayesian classifiers perform well for both completely independent and functionally dependent feature variables (Domingos & Pazzani, 1997; Rish, 2001).

In this paper, we employ techniques from sensitivity analysis to study the effects of parameter inaccuracies on a naive Bayesian network's posterior probability distributions and thereby contribute yet another explanation of the observed stability of naive Bayesian classifiers. In general, the sensitivity of a posterior probability of interest to parameter variation complies with a simple mathematical function (Castillo *et al.*, 1997; Coupé & Van der Gaag, 2002), called a sensitivity function. We will demonstrate that the independence assumptions underlying a naive Bayesian network constrain these sensi-

tivity functions to such a large extent that they can in fact be established exactly from very limited information. In addition, we study the various sensitivity properties that follow from the constrained functions. We would like to note that sensitivity analysis has been applied before in the context of naive Bayesian classifiers, as a means of providing bounds on the amount of parameter variation that is allowed without changing, for *any* of the possible instances, the class an instance is assigned to (Chan & Darwiche, 2003). We extend on this result with further insights.

For classification problems, it is often assumed that evidence is available for every single feature variable. In numerous application domains, however, this assumption may not be realistic. The question then arises how much impact additional evidence could have on the output of interest and how sensitive this impact is to inaccuracies in the network’s parameters. We introduce the novel notion of scenario sensitivity to capture the latter type of sensitivity and show that the effects of parameter variation in view of scenarios of additional evidence can be established efficiently.

The paper is organised as follows. In Section 2, we present some preliminaries on sensitivity functions and their associated sensitivity properties. In Section 3, we establish the functional forms of the sensitivity functions for naive Bayesian networks and address the ensuing sensitivity properties. In Section 4 we introduce the notion of scenario sensitivity and show that it can be established from standard sensitivity functions. The paper ends with our conclusions and directions for future research in Section 5.

2 Preliminaries

To investigate the effects of inaccuracies in its parameters, a Bayesian network can be subjected to a sensitivity analysis. In such an analysis, the effects of varying a single parameter on an output probability of interest are studied. These effects are described by a simple mathematical function, called a *sensitivity function*. If, upon varying a single parameter, the parame-

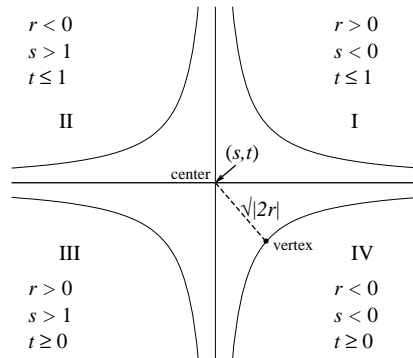


Figure 1: The possible hyperbolas and their constants, where the constraints on r , s and t are specific for sensitivity functions.

ters pertaining to the same conditional distribution are co-varied proportionally, then the sensitivity function is a quotient of two linear functions in the parameter x under study (Castillo *et al.*, 1997; Coupé & Van der Gaag, 2002); more formally, the function takes the form

$$f(x) = \frac{a \cdot x + b}{c \cdot x + d}$$

where the constants a, b, c, d are built from the assessments for the non-varied parameters. The constants of the function can be feasibly computed from the network (Coupé & Van der Gaag, 2002; Kjærulff & Van der Gaag, 2000).

A sensitivity function is either a *linear* function or a fragment of a *rectangular hyperbola*. A rectangular hyperbola takes the general form

$$f(x) = \frac{r}{x - s} + t$$

where, for a function with a, b, c, d as before, we have that $s = -\frac{d}{c}$, $t = \frac{a}{c}$, and $r = \frac{b}{c} + s \cdot t$. The hyperbola has two branches and the two asymptotes $x = s$ and $f(x) = t$; Figure 1 illustrates the locations of the possible branches relative to the asymptotes. Since both x and $f(x)$ are probabilities, the two-dimensional space of their feasible values is defined by $0 \leq x, f(x) \leq 1$; this space is termed the *unit window*. Since any sensitivity function should be well-behaved in the unit window, a hyperbolic sensitivity function is a fragment of just one of the four possible branches in Figure 1.

From a sensitivity function, various properties can be computed. Here we briefly re-

view the properties of *sensitivity value* (Laskey, 1995), *vertex-proximity* and *admissible deviation* (Van der Gaag & Renooij, 2001). The *sensitivity value* for a parameter is the absolute value $\left| \frac{\partial f}{\partial x}(x_0) \right|$ of the first derivative of the sensitivity function at the assessment x_0 for the parameter; it captures the effect of an infinitesimally small shift in the parameter on the output probability. The impact of a larger shift is strongly dependent upon the location of the *vertex* of the sensitivity function, that is, the point where $\left| \frac{\partial f}{\partial x}(x) \right| = 1$. A vertex that lies within the unit window basically marks the transition from parameter values with a high sensitivity value to parameter values with a low sensitivity value, or vice versa. If an output probability is used for establishing the most likely value of the output variable, the effect of parameter variation on the most likely output value is of interest. The *admissible deviation* for a parameter now is a pair (α, β) , where α is the amount of variation allowed to values smaller than its assessment without changing the most likely output value, and β is the amount of variation allowed to larger values; the symbols \leftarrow and \rightarrow are used to indicate that variation is allowed up to the boundaries of the unit window.

3 Sensitivities in Naive Classifiers

Upon being subjected to a sensitivity analysis, the independence assumptions of a naive Bayesian network strongly constrain the general form of the resulting sensitivity functions. In fact, given just limited information, the exact functions can be readily determined for each class value and each parameter. In this section we derive these functions and discuss their ensuing sensitivity properties.

3.1 Functional forms

Before establishing the sensitivity functions for naive Bayesian networks, we introduce some notational conventions. We consider a naive Bayesian network with nodes $V(G) = \{C\} \cup E$, $E = \{E_1, \dots, E_n\}$, $n \geq 2$, and arcs $A(G) = \{(C, E_i) \mid i = 1, \dots, n\}$; C is called the *class variable* of the network and E_i are termed its

feature variables. We now study the sensitivity of the posterior probability $\Pr(c \mid e)$ of a class value c given an input instance e in terms of a parameter $x = p(e'_v \mid c')$, where e'_v denotes some value of the feature variable E_v and c' is a value of the class variable. The original assessment specified in the network for the parameter x is denoted by x_0 . The notation p_c is used for the posterior probability of interest prior to parameter variation. We use $f_{\Pr(c|e)}(x)$ to denote the sensitivity function that expresses the probability of interest in terms of the parameter x .

The following proposition now captures the sensitivity function that describes the output probability of interest of a naive Bayesian network in terms of a single parameter associated with one of the feature variables. The proposition shows that this function is highly constrained as a result of the independences in the network and the parameter being conditioned on a value of the class variable. In fact, for any class value and any parameter, only one of four different functions can result. We would like to note that the proposition holds only for parameters and output probabilities that allow for a hyperbolic sensitivity function.

Proposition 1. *Let E_v be a feature variable with the value e_v and let $x = p(e'_v \mid c')$ be a parameter associated with E_v . Then, the sensitivity function $f_{\Pr(c|e)}(x)$ has one of the following forms:*

$f_{\Pr(c e)}(x)$	$e_v = e'_v$	$e_v \neq e'_v$
$c = c'$	$\frac{x}{x - s_1}$	$\frac{x - 1}{x - s_2}$
$c \neq c'$	$p_c \cdot \frac{x_0 - s_1}{x - s_1}$	$p_c \cdot \frac{x_0 - s_2}{x - s_2}$

where $s_1 = x_0 - \frac{x_0}{p'}$ and $s_2 = x_0 + \frac{(1-x_0)}{p'}$ are the vertical asymptotes of the corresponding functions and p' is the original value of $\Pr(c' \mid e)$.

Proof. We prove the property stated in the proposition for $e_v = e'_v$; the proof for $e_v \neq e'_v$ is analogous. We begin by writing the probability $\Pr(c, e)$ in terms of the network's parameters:

$$\Pr(c, e) = \prod_{E_i \in E \setminus \{E_v\}} p(e_i \mid c) \cdot p(c) \cdot p(e_v \mid c),$$

where e_i is the value of the variable E_i in the input instance e . This probability relates to the parameter $x = p(e'_v | c')$ as

$$\Pr(c', e)(x) = \prod_{E_i \in E \setminus \{E_v\}} p(e_i | c') \cdot p(c') \cdot x$$

for $c = c'$ and as

$$\Pr(c, e)(x) = \prod_{E_i \in E} p(e_i | c) \cdot p(c)$$

for $c \neq c'$. Similarly, $\Pr(e)$ can be written as

$$\Pr(e) = \Pr(c', e) + \sum_{c \neq c'} \Pr(c, e),$$

where $\Pr(c', e)$ again relates to the parameter x as indicated above and the other term is constant with respect to the parameter. For $c = c'$ we now find for the sensitivity function that

$$f_{\Pr(c'|e)}(x) = \frac{\Pr(c', e)(x)}{\Pr(e)(x)} = \frac{a \cdot x}{a \cdot x + d} = \frac{x}{x - s_1}$$

and for $c \neq c'$ that

$$f_{\Pr(c|e)}(x) = \frac{\Pr(c, e)(x)}{\Pr(e)(x)} = \frac{b_c}{a \cdot x + d} = \frac{r_c}{x - s_1}$$

with the constants

$$\begin{aligned} a &= \prod_{E_i \in E \setminus \{E_v\}} p(e_i | c') \cdot p(c') \\ b_c &= \prod_{E_i \in E} p(e_i | c) \cdot p(c) \\ d &= \sum_{c \neq c'} \Pr(c, e) \end{aligned}$$

The constant $s_1 = -\frac{d}{a}$ is the vertical asymptote that is shared by all functions; its value is determined from

$$\begin{aligned} s_1 &= -\frac{\Pr(e) - \Pr(c', e)}{\Pr(c', e)/p(e'_v | c')} \\ &= x_0 - \frac{x_0}{p'} \end{aligned}$$

In addition, the function $f_{\Pr(c'|e)}(x)$ has a horizontal asymptote at $t = \frac{a}{a} = 1$. The functions $f_{\Pr(c|e)}(x)$ with $c \neq c'$, all have $t = \frac{0}{a} = 0$. The constants $r_c = \frac{b_c}{a}$ of the latter functions are determined from $p_c = \frac{r_c}{x_0 - s_1}$. \square

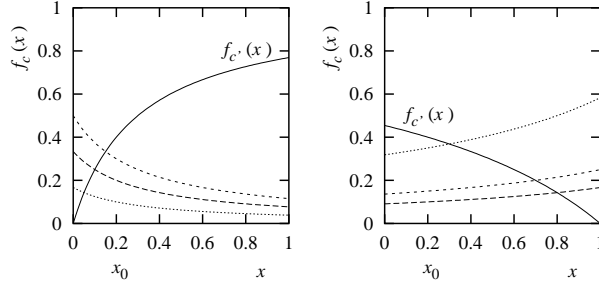


Figure 2: Example sensitivity functions for the parameter $x = p(e'_v | c')$ with $x_0 = 0.2$, $p' = 0.4$, where $e_v = e'_v$ (left) or $e_v \neq e'_v$ (right).

From the values of the asymptotes found in the proof above, we observe that for $e_v = e'_v$ the sensitivity function $f_{\Pr(c'|e)}(x)$ is a fragment of a IVth-quadrant hyperbola branch; all other functions $f_{\Pr(c|e)}(x)$ with $c \neq c'$ are fragments of Ist-quadrant branches. For $e_v \neq e'_v$ we find IIIrd- and IInd-quadrant branches, respectively. These properties are illustrated in Figure 2. To intuitively explain why the function $f_{\Pr(c'|e)}(x)$ has a different shape than the other functions, we observe that varying a parameter $p(e_i | c')$ has a direct effect on the probability $\Pr(c' | e)$, whereas the probabilities $\Pr(c | e)$, $c \neq c'$, are only indirectly affected to ensure that the distribution sums to one. Due to the constrained form of these functions, moreover, all $f_{\Pr(c|e)}(x)$, $c \neq c'$, have the same shape. The function $f_{\Pr(c'|e)}(x)$ therefore *must* be deviant.

We illustrate the functional forms derived above with an example. The example shows more specifically that as a result of their constrained form, any sensitivity function can be established from very limited information.

Example 1. We consider a naive Bayesian network with a class variable modelling the stages I, IIA, IIB, III, IVA and IVB of cancer of the oesophagus; the feature variables of the network capture the results from diagnostic tests. For a given patient, the available findings are summarised in the input instance e , given which the following posterior distribution is computed over the class variable S :

S	I	IIA	IIB	III	IVA	IVB
$\Pr(S e)$	0.01	0.19	0.01	0.07	0.61	0.11

We further consider the feature variable *CT-loco*, modelling the observed presence or absence of loco-regional lymphatic metastases. The network includes the following assessments for the parameters for this variable:

S	I	IIA	IIB	III	IVA	IVB
<i>CT-loco</i> yes	0.02	0.02	0.48	0.48	0.48	0.27
no	0.98	0.98	0.52	0.52	0.52	0.73

Now suppose that we are interested in the effects of inaccuracies in the parameter $x = p(\text{CT-loco} = \text{no} \mid S = \text{IVA})$ on the posterior probabilities $\Pr(S \mid e)$ for our patient who has no evidence of loco-regional metastases. These effects are captured by six functions with the vertical asymptote $s_1 = 0.52 - \frac{0.52}{0.61} = -0.33$. Without having to perform any further computations, we now find that

$$f_{\text{IVA}}(x) = \frac{x}{x + 0.33}$$

and

$$f_S(x) = \Pr(S \mid e) \cdot \frac{0.85}{x + 0.33}$$

for any $S \neq \text{IVA}$. We further find that for the complement of the parameter x all functions have their vertical asymptote at $s_2 = 1.33$. \square

3.2 Sensitivity properties

For a naive Bayesian network, any sensitivity function is exactly determined by the assessment for the parameter being varied and the original value of the probability of interest. From the function now any sensitivity property of interest can be computed. We study the properties of sensitivity value, vertex proximity and admissible deviation.

Sensitivity value and vertex proximity

For the various possible sensitivity functions for a naive Bayesian network, the sensitivity values are readily established: for $e_v = e'_v$ we find that

$$\left| \frac{\partial f_c}{\partial x}(x_0) \right| = (1 - p') \cdot \frac{p'}{x_0} \geq p_c \cdot \frac{p'}{x_0} = \left| \frac{\partial f_c}{\partial x}(x_0) \right|.$$

For $e_v \neq e'_v$, we find a similar result by replacing x_0 by $1 - x_0$. Note that for a binary class variable, the sensitivity values for its two class values are the same. This property holds for any binary output variable in any Bayesian network.

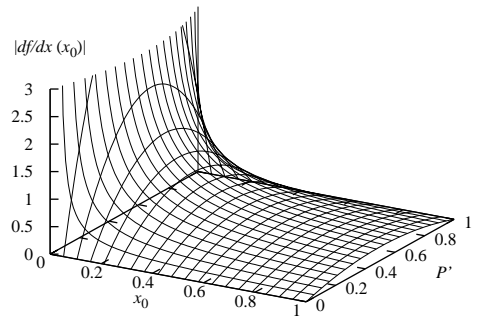


Figure 3: The sensitivity value for $f_{\Pr(c'|e)}(x)$ as a function of x_0 and p' , for $e_v = e'_v$.

Based upon the general form of the sensitivity value derived above, Figure 3 depicts the sensitivity value of $f_{\Pr(c'|e)}(x)$ given $e_v = e'_v$ for different combinations of values for x_0 and p' . We observe that high sensitivity values can only be found if the assessment for the parameter under study is small and the original posterior probability p' is non-extreme. More formally, we have that $\left| \frac{\partial f_{\Pr(c'|e)}(x_0)}{\partial x} \right| > 1$ if and only if $x_0 < p' - p'^2 \leq 0.25$. For $e_v \neq e'_v$, high sensitivity values can only be found if the assessment for the parameter is larger than 0.75 and the original posterior probability p' is non-extreme. Note that high sensitivity values can thus only be found for the parameters that describe the *least likely* value of a feature variable given a particular class value. This property was noticed before for general Bayesian networks (Van der Gaag & Renooij, 2006).

In view of naive Bayesian classifiers, we observe that input instances which occur the more frequently given a particular class are also likely to be the more probable given that class in the network underlying the classifier. Since high sensitivity values can only be found for the parameters that describe the least likely value of a feature variable, the posterior probability of the corresponding class value for such an input instance will not be very sensitive to inaccuracies in the various parameters. In fact, the vertical asymptote of an associated sensitivity function, and hence the x -value of the function's vertex, will be quite distant from the parameter's assessment, resulting in a rather flat function in

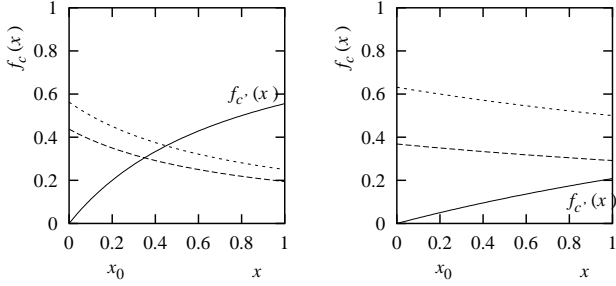


Figure 4: Examples of sensitivity functions for the parameter $x = p(e'_v | c')$, where $e_v = e'_v$ and c' is *not* the most likely value. Variation either can make c' the most likely value (*left*) or will not change the most likely value (*right*).

the broader vicinity of x_0 . The most likely class value will then hardly ever change upon small shifts in a parameter. The above argument thus corroborates the empirically observed robustness of classifiers to parameter inaccuracies.

Admissible deviation In view of classification, the property of admissible deviation is especially of interest. The admissible deviation for a parameter gives the amount of variation that is allowed before an instance is classified as belonging to a different class. The following proposition now gives a general form for all possible deviations in a naive Bayesian network. The proposition more specifically shows that in such a network, the most likely class value can change *at most once* upon varying a parameter.

Proposition 2. *Let $p^\top = \max\{p_c | c \neq c'\}$, let c^\top be a value of C for which $\Pr(c^\top | e) = p^\top$, and let $f^\top(x) = f_{\Pr(c^\top|e)}(x)$. Then,*

- we have $f_{\Pr(c|e)}(x) \leq f^\top(x)$ for all $c \neq c'$;
- for $e_v = e'_v$ the admissible deviation for the parameter x is:

condition	admissible deviation
$p^\top < p'$	$(x_0 - x_s, \rightarrow)$
$p^\top = p'$	$(0, 0)$
$p^\top > p'$	$(\leftarrow, x_s - x_0)$ if $x_0 < \frac{p'}{p^\top}$
	$(\leftarrow, \rightarrow)$ otherwise

where $x_s = p^\top \cdot \frac{x_0}{p'}$;

- for $e_v \neq e'_v$ then the above admissible deviations hold with each occurrence of x_0 and x_s replaced by $1 - x_0$ and $1 - x_s$, respectively.

Proof. The first property follows from the observation that all functions $f_{\Pr(c|e)}$ with $c \neq c'$ have the same horizontal and vertical asymptotes and therefore do not intersect. As a consequence, either c' or c^\top is the most likely value of C , regardless of the value of x . If the two functions $f_{\Pr(c'|e)}(x)$ and $f^\top(x)$ intersect at x_s , then we have that $x_s \in [0, \infty)$ for $e_v = e'_v$ and $x_s \in \langle -\infty, 1]$ for $e_v \neq e'_v$. The second and third properties stated above now follow immediately from the general forms of the functions. \square

We observe from the above proposition that, for any parameter, an admissible amount of variation different from \leftarrow or \rightarrow is possible only to the left or to the right of the parameter's assessment but never in both directions. From this observation we have that upon varying the parameter, the most likely class value can change only once. Figures 2 and 4 support this observation.

Example 2. We consider again the naive Bayesian classifier and the patient information from Example 1. Suppose that we are once more interested in the effects of inaccuracies in the parameter $x = p(CT-loco = no | S = IVA)$ on the posterior probabilities $\Pr(S | e)$ for the patient. We recall that the value IVA corresponds to the most likely stage for this patient, with a probability of 0.61. The second most likely stage for the patient is stage IIA, with probability 0.19. We now find that the most likely stage changes from IVA to IIA at $x_s = 0.19 \cdot \frac{0.52}{0.61} = 0.16$. The admissible deviation for the parameter thus is $(0.36, \rightarrow)$. Note that the most likely stage cannot change into any other value upon varying the parameter under study. \square

In view of naive Bayesian network classifiers, we have from the above proposition that we can expect to find large admissible deviations for mist instances. To support this expectation, we consider a parameter x with $e_v = e'_v$; similar arguments hold for $e_v \neq e'_v$. If $p^\top < p'$, we have from the functional forms that the assessment x_0 for the parameter is unlikely to be very small, as for

functions with an asymptote to the left of the unit window the intersection of the sensitivity functions lies to the left of x_0 . If $p^\top > p'$, on the other hand, we expect an admissible deviation within the unit window *only* for very small values of x_0 . If the instances that are modelled as the more probable given a particular class value, occur the more often in practice, then for these instances we expect that x_0 is not a small value and that $p' > p^\top$. In practice we would therefore expect most instances to result in admissible deviations of (\leftarrow , \rightarrow).

4 Scenario sensitivity

For classification problems, it is generally assumed that evidence is available for every single feature variable. In practical applications, however, this assumption may not be realistic. In the medical domain, for example, a patient is to be classified into one of a number of diseases without being subjected to every possible diagnostic test. The question then arises how much impact additional evidence could have on the probability distribution over the class variable and how sensitive this impact is to inaccuracies in the network's parameters. The former issue is closely related to the notion of value of information. The latter issue involves a notion of sensitivity that differs from the standard notion used in the previous sections in the sense that it pertains not to available evidence but to scenarios of possibly additional evidence. We refer to this notion of sensitivity as *scenario sensitivity* and use the term *evidence sensitivity* to refer to the more standard notion. Although it is applicable to Bayesian networks in general, in this section we discuss scenario sensitivity in the context of naive Bayesian classifiers.

To study the effect of additional evidence on an output probability for the class variable, we consider the ratio $\Pr(c | e^N)$ to $\Pr(c | e^O)$ where e^O and e^N denote the evidence prior to and after receiving the new evidence respectively.

Proposition 3. *Let E^O and E^N be sets of feature variables with $\emptyset \subseteq E^O \subset E^N \subseteq E$ and $E^N - E^O = \{E_1, \dots, E_l\}$, $1 \leq l \leq n$. Let e^O and e^N be consistent instances of E^O and E^N ,*

respectively. Then, for each c , we have that

$$\frac{\Pr(c | e^N)}{\Pr(c | e^O)} = \frac{\prod_{i=1}^l \Pr(e_i | c)}{\sum_{c_j} \prod_{i=1}^l \Pr(e_i | c_j) \cdot \Pr(c_j | e^O)}$$

where e_i is the value of E_i in E^N .

Proof. The property follows immediately by applying Bayes' rule and exploiting the independences that hold in the naive network. \square

The above proposition now allows us to compute the new posterior probability distribution over the class variable from the previous one.

Example 3. We reconsider the naive Bayesian network and the patient information from the previous examples. In addition to the tests to which the patient has already been subjected, a scan of the liver can be performed. We now are interested in the posterior distribution over the various stages if the result of this test were positive. For the feature variable *CT-liver*, the following parameter assessments are specified:

S	I	IIA	IIB	III	IVA	IVB
<i>CT-liver</i> yes	0.05	0.05	0.05	0.05	0.05	0.69
no	0.95	0.95	0.95	0.95	0.95	0.31

From the table and the probability distribution $\Pr(S | e)$ from Example 1, we find that $\sum_S \Pr(\text{CT-liver} = \text{yes} | S) \cdot \Pr(S | e) = 0.1204$. The new posterior probability of $S = \text{IVB}$ now follows directly from

$$\frac{\Pr(\text{IVB} | e^N)}{0.11} = \frac{0.69}{0.1204} = 0.63$$

without requiring any additional computations from the network. Note that the new test result would change the most likely stage. \square

The above proposition allows for establishing the impact of additional evidence on the posterior probability distribution over the class variable. To capture the sensitivity of this impact to parameter variation, we consider the function $h(x)$ that describes the above probability ratio as a function of a parameter $x = p(e'_v | c')$:

$$h(x) = \left(\frac{\Pr(c | e^N)}{\Pr(c | e^O)} \right) (x) = \frac{\Pr(c | e^N)(x)}{\Pr(c | e^O)(x)}$$

If the parameter x pertains to a variable in the set $E^N - E^O$, then the denominator is a constant with respect to x . The function $h(x)$ then

just is a *scaled* version of the sensitivity function $f_{\Pr(C|e^N)}(x)$. Given the probability distribution $\Pr(C | e^o)$ over the class variable, we can therefore immediately determine the sensitivity of the impact of the additional evidence from the sensitivity function $f_{\Pr(C|e^N)}(x)$. Note that in a naive Bayesian network the latter function is known for each parameter x once the posterior probability distribution $\Pr(C | e^N)$ is available.

Example 4. We reconsider the previous example. We now are interested in the effects of inaccuracies in the parameter $x = p(CT\text{-liver} = \text{yes} | \text{IVB})$ on the ratio $\Pr(\text{IVB} | e^N)$ to $\Pr(\text{IVB} | e^o)$. We recall that the new posterior probability of $S = \text{IVB}$ would be 0.63; the probability given just the available evidence was 0.11. We now establish the sensitivity function $f_{\Pr(\text{IVB}|e^N)}(x) = \frac{x}{x+0.41}$ and find that

$$h_{\text{IVB}}(x) = \frac{1}{0.11} \cdot \frac{x}{x + 0.41}$$

From Example 3 we had that the probability of the class value IVB increased from 0.11 to 0.63 upon a positive liver scan, thereby becoming 5.7 times as likely. We can now in addition conclude that if the parameter x is varied, the class value IVB can become at most 6.4 times as likely as without the additional evidence. \square

5 Conclusions

In this paper, we used techniques from sensitivity analysis to study the effects of parameter inaccuracies on a naive Bayesian network's posterior probability distributions. We showed that the independence assumptions of such a network constrain the functional form of the associated sensitivity functions: these functions are determined solely by the assessment for the parameter under study and the original posterior probability distribution over the class variable. The sensitivity properties following from the functions provided some fundamental arguments which corroborated the empirically observed robustness of classifiers to parameter inaccuracies. More research is required, however, to further substantiate these arguments. We further introduced the novel notion of scenario sensitivity, which describes the effects of pa-

rameter inaccuracies in view of scenarios of additional evidence. We showed that for naive Bayesian networks scenario sensitivity can be readily expressed in terms of the more standard sensitivity functions. In the near future, we would like to study scenario sensitivity in Bayesian networks in general.

References

- E. Castillo, J.M. Gutiérrez and A.S. Hadi. 1997. Sensitivity analysis in discrete Bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics*, 27: 412 – 423.
- H. Chan and A. Darwiche. 2003. Reasoning about Bayesian network classifiers. In *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, pages 107 – 115.
- V.M.H. Coupé and L.C. van der Gaag. 2002. Properties of sensitivity analysis of Bayesian belief networks. *Annals of Mathematics and Artificial Intelligence*, 36: 323 – 356.
- P. Domingos and M. Pazzani. 1997. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29: 103 – 130.
- U. Kjærulff and L.C. van der Gaag. 2000. Making sensitivity analysis computationally efficient. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, pages 317 – 325.
- K.B. Laskey. 1995. Sensitivity analysis for probability assessments in Bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics*, 25: 901 – 909.
- P. Liu, L. Lei and N. Wu. 2005. A quantitative study of the effect of missing data in classifiers. In *IEEE Proceedings of the 5th International Conference on Computer and Information Technology*, pages 28 – 33.
- I. Rish. 2001. An empirical study of the naive Bayes classifier. In *IJCAI Workshop on Empirical Methods in Artificial Intelligence*, pages 41 – 46.
- L.C. van der Gaag and S. Renooij. 2001. Analysing sensitivity data from probabilistic networks. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, pages 530 – 537.
- L.C. van der Gaag and S. Renooij. 2006. On the sensitivity of probabilistic networks to reliability characteristics. *Modern Information Processing: From Theory to Applications*, Elsevier, pages 385 – 405.