Towards Being Discrete in Naive Bayesian Networks

Roel Bertens Silja Renooij Linda C. van der Gaag

Department of Information and Computing Sciences, Utrecht University
P.O. Box 80.089, 3508 TB, The Netherlands
{R.Bertens, S.Renooij, L.C.vanderGaag}@uu.nl

Abstract

Bayesian networks are often used in problem domains that include variables of a continuous nature. For capturing such variables, their value ranges basically have to be modelled as finite sets of discrete values. While the output probabilities and conclusions established from a network are dependent of the actual discretisations used for its variables, the effects of choosing alternative discretisations are largely unknown as yet. This paper describes the first steps of a study into the effects of changing discretisations on the probability distributions computed from a Bayesian network. We focus more specifically on the feature variables of a naive Bayesian network and demonstrate how insights from the research area of sensitivity analysis can be exploited for studying how the network's output is affected by alternative discretisations.

1 Introduction

Bayesian networks are being used in an ever increasing range of problem domains. A Bayesian network in essence is a compact representation of a joint probability distribution over a set of stochastic variables. Modern Bayesian-network tools provide various algorithms for probabilistic inference with a network, which basically allow computing any prior or posterior probability of interest over the network's variables [1, 2]. Most of these inference algorithms assume the variables of a network to be discrete. The knowledge in a problem domain under study, however, may involve variables which are of a continuous nature. For capturing such variables in a network, their value ranges should be modelled as finite sets of discrete values. Several different methods are available for automated discretisation of variables in general, each of which requires for its input the number of intervals into which a value range is to be split, and possibly also the boundaries of these intervals; for an overview of methods, we refer to [3]. For Bayesian-network modelling, these general methods unfortunately tend to yield unsatisfactory results [4]. Yet, while the output probabilities and conclusions established from a network are dependent of the actual ways in which its continuous variables are discretised [5], the effects of choosing alternative discretisations are largely unknown.

As a first step in studying the effects of changing discretisations on the posterior probability distributions computed from a Bayesian network, we focus in this paper on the class of naive Bayesian networks. Naive Bayesian networks are networks of highly restricted topology, composed of a single class variable of interest and multiple feature variables which are directly connected with this class variable and unconnected otherwise. For ease of exposition, we will consider binary discretisations only for the feature variables, splitting their value ranges into two intervals. Choosing an alternative discretisation for a variable then amounts to choosing a different threshold value to separate the two intervals. We will argue that a change of threshold value will result in changes in the values of all parameter probabilities specified for the variable at hand. Building upon this observation, we will then demonstrate how recent insights from the research area of sensitivity analysis of Bayesian networks in general [6, 7] can be used to study the effects of different discretisations on the output probability distributions established from a naive Bayesian network. We further establish conditions under which a change in discretisation can affect the conclusions drawn from the network. Throughout the paper, we will illustrate our findings using real-life data from which various naive Bayesian networks were constructed [8].

The paper is organised as follows. In Section 2, we provide some preliminaries on Bayesian networks and on sensitivity analysis. In Section 3, we establish functions that describe the effects of changing the discretisation of a feature variable on the probability distribution computed from a naive Bayesian network. Whether or not choosing another discretisation can lead to a different conclusion for the class variable is discussed briefly in Section 4. The paper ends with our concluding observations in Section 5.

2 Preliminaries

We briefly review concepts from Bayesian networks and insights from sensitivity analysis.

2.1 Bayesian networks

A Bayesian network is a probabilistic graphical model, which describes a joint probability distribution $\Pr(\mathbf{V})$ over a set $\mathbf{V}$ of discrete stochastic variables. The variables and their interrelationships are modelled as nodes and arcs respectively, in an acyclic directed graph. Associated with each variable V in this graph is a set of parameter probabilities $p(V \mid \pi(V))$ from the distribution $\Pr$ which describe the influence of the possible values of the parents $\pi(V)$ of V on the probabilities over V itself; this set is commonly termed the conditional probability table of V. For computing prior and posterior probabilities over the separate variables of a Bayesian network, efficient algorithms are available [2].

Over the years, several restricted types of Bayesian network have been distinguished, among which is the rather popular naive Bayesian network. A naive Bayesian network consists of a single class variable C and one or more feature variables E_i. In the network's graphical structure, the feature variables are all connected directly with the class variable and not connected otherwise; the feature variables thereby are modelled as being mutually independent given the class variable. Naive Bayesian networks are commonly used for computing posterior probability distributions $\Pr(C \mid \mathbf{E})$ over the various values of the class variable, given joint values for the set $\mathbf{E}$ of feature variables. When used as a classification model, the most likely value of the class variable given a specific joint value $\mathbf{e}$ of $\mathbf{E}$ is established from the computed probability distribution $\Pr(C \mid \mathbf{e})$ and returned as the model's output. In the sequel, we will focus on naive Bayesian networks and consider binary-valued class variables C only, with values c and $\bar{c}$.

2.2 Sensitivity analysis

Sensitivity analysis is a general technique for studying the effects of parameter variation on the output of a mathematical model. For Bayesian networks, tailored techniques for sensitivity analysis have been developed, which provide for investigating the effects of varying the values of one or more parameter probabilities on an output probability of interest; for an overview of recent insights, we refer to [9]. The research area of sensitivity analysis of Bayesian networks so far focused mainly on one-way analyses in which a single parameter probability p is varied. The effects of varying this parameter are captured by a mathematical function which describes the output probability as a function of the parameter. When a parameter probability p is being varied as x in such an analysis, its complement $1-p$ from the same distribution varies as $1-x$. Given this co-variation, a one-way sensitivity function $f(x)$ for a marginal or joint probability of interest is linear in the parameter being varied. For a conditional probability of interest, the effects of parameter variation are described by a fraction of two linear functions. This sensitivity function $f(x)$ essentially is a fragment of one of the branches of a rectangular hyperbola [9]. We note that both the parameter under study and the probability of interest are restricted to values from the range $[0, 1]$, thereby effectively constraining the two-dimensional space of feasible points to the so-termed unit window. For computing sensitivity functions from a Bayesian network, efficient algorithms are available [10].

In the sequel, we will use a higher-order sensitivity analysis in which multiple parameter probabilities are being varied simultaneously. In general, in an n-way sensitivity analysis in which n parameters are varied, a marginal or joint probability of interest is described by a multi-linear function in these parameters; for a conditional probability, the sensitivity function is a fraction of two such functions. For example, a two-way sensitivity function that expresses some posterior probability of interest $\Pr(c \mid \mathbf{e})$ in terms of two parameter probabilities which are being varied as x and y, has the following form:

$$f_{\Pr(c|\mathbf{e})}(x, y) = \frac{f_{\Pr(c,\,\mathbf{e})}(x, y)}{f_{\Pr(\mathbf{e})}(x, y)} = \frac{a_1 \cdot x \cdot y + a_2 \cdot x + a_3 \cdot y + a_4}{b_1 \cdot x \cdot y + b_2 \cdot x + b_3 \cdot y + b_4}$$

where the constants $a_i, b_i, i = 1, \ldots, 4$, are built from the non-varied parameters of the network under study. The two parameter probabilities and the output probability of interest again are restricted to the $[0, 1]$-range, which defines a three-dimensional space of feasible points called the unit cube.

For studying the effects of changing a variable's discretisation in a naive Bayesian network, we will consider in the sequel the parameters from the conditional probability table of a single feature variable only. We will the argue that the parameter probabilities to be varied stem from different conditional distributions, that is, conditioned on different values of the class variable. We now observe that the $x \cdot y$ terms in a general two-way sensitivity function describe the interaction effects of the two parameters being varied as x and y. If the two parameters stem from different conditional probability distributions however, as in our study, the constants a_1 and b_1 in the interaction terms are equal to zero. For the parameters in our study, we can thus simplify the general form of a two-way sensitivity function to [11]:

$$f_{\Pr(c|\mathbf{e})}(x, y) = \frac{a_2 \cdot x + a_3 \cdot y + a_4}{b_2 \cdot x + b_3 \cdot y + b_4}$$

As long as ambiguity cannot arise, we will write $f(x)$ instead of $f_{\Pr(c|\mathbf{e})}(x)$, for short. In our analyses, we will further assume that none of the parameter probabilities specified in a naive Bayesian network equal zero. We assume moreover that the parameter probabilities for the feature variable under study are not varied to be equal to 0 or 1, that is, we assume that $x, y \in \langle 0, 1 \rangle$: although in theory it is possible to generate a parameter probability equal to 0 or 1 by discretisation, such a discretisation would not be very useful in practice.

3 Studying the Effects of Discretisation

The basic formalism of Bayesian networks requires all included variables to be discrete. Upon modelling domain knowledge, variables which take their value from an intrinsically continuous value range will therefore have to be discretised before they can be captured in a network. Such a discretisation in essence amounts to splitting the variable's value range into two or more intervals and associating each interval with a value of a (newly defined) discrete variable. We consider a single continuous variable E_i and address its discretisation. For ease of exposition, we assume that the value range of this variable is split into two intervals by means of a threshold value t; note that choosing a different discretisation for E_i amounts to changing this threshold value. Slightly abusing notation, we will write $\bar{e}_i \equiv E_i < t$ and $e_i \equiv E_i \geq t$ for the two values of the (now discretised) variable E_i; we will write e'_i to indicate either of the two values.

Upon including a discretised variable E_i as a feature variable in a naive Bayesian network, a conditional probability table is constructed which includes the parameter probability $p(E_i < t \mid c)$ and its complement $p(E_i \geq t \mid c)$, as well as the probabilities $p(E_i < t \mid \bar{c})$ and $p(E_i \geq t \mid \bar{c})$. It is now readily seen that changing the discretisation of E_i by choosing a different threshold value t, will affect *all* parameters from the probability table for E_i. Since these parameter probabilities do not stem all from the same conditional distribution, we must conclude that we cannot study the effects of changing E_i's discretisation by conducting a one-way sensitivity analysis in terms of the parameters involved. But it is not necessary to use a four-way sensitivity analysis either. We observe that by varying the parameter probability $p(E_i < t \mid c)$, the variation of $p(E_i \geq t \mid c)$ is covered by standard co-variation; similarly, variation of $p(E_i \geq t \mid \bar{c})$ is handled by varying $p(E_i < t \mid \bar{c})$. A two-way sensitivity analysis thus in essence suffices for studying the effects of changing the discretisation for E_i on the output probabilities computed from a naive Bayesian network.

In Section 2, we reviewed the general form of a two-way sensitivity function. In previous research, we showed that the independence properties underlying a naive Bayesian network allow for simplifying the general form of a one-way sensitivity function [12]. We now show that naive Bayesian networks also yield two-way sensitivity functions of a constrained form.

Proposition 1. *Let C be the class variable of a binary naive Bayesian network which further includes the set $\mathbf{E}$ of feature variables. Let $\Pr(c \mid \mathbf{e})$ be the network's probability of interest, for a joint combination of values $\mathbf{e}$ for $\mathbf{E}$. Now, let $x = p(e''_i \mid c)$ and $y = p(e''_i \mid \bar{c})$ be parameter probabilities for some value e''_i of the feature variable $E_i \in \mathbf{E}$, and let e'_i be the actually observed value for E_i in $\mathbf{e}$. Then, the two-way sensitivity function expressing $\Pr(c \mid \mathbf{e})$ as a function of x and y has the following form:*

$$f_{\Pr(c|\mathbf{e})}(x, y) = \begin{cases} \dfrac{a \cdot \Pr(c) \cdot x}{a \cdot \Pr(c) \cdot x + a' \cdot \Pr(\bar{c}) \cdot y} & \text{if} \quad e'_i = e''_i \\ \dfrac{a \cdot \Pr(c) \cdot (1 - x)}{a \cdot \Pr(c) \cdot (1 - x) + a' \cdot \Pr(\bar{c}) \cdot (1 - y)} & \text{if} \quad e'_i \neq e''_i \end{cases}$$

where a and a' are constants.

Proof. We consider the probability of interest $\Pr(c \mid \mathbf{e})$. Using Bayes' theorem and exploiting the independence properties of a naive Bayesian network, we find that

$$\begin{aligned}\Pr(c \mid \mathbf{e}) &= \frac{\Pr(\mathbf{e} \mid c) \cdot \Pr(c)}{\Pr(\mathbf{e} \mid c) \cdot \Pr(c) + \Pr(\mathbf{e} \mid \bar{c}) \cdot \Pr(\bar{c})} \\ &= \frac{\prod_{e'_k \text{ in } \mathbf{e}} \Pr(e'_k \mid c) \cdot \Pr(c)}{\prod_{e'_k \text{ in } \mathbf{e}} \Pr(e'_k \mid c) \cdot \Pr(c) + \prod_{e'_k \text{ in } \mathbf{e}} \Pr(e'_k \mid \bar{c}) \cdot \Pr(\bar{c})}\end{aligned}$$

The result now follows immediately with $a = \prod_{e'_k \text{ in } \mathbf{e}, k \neq i} \Pr(e'_k \mid c)$ and $a' = \prod_{e'_k \text{ in } \mathbf{e}, k \neq i} \Pr(e'_k \mid \bar{c})$. □

We note that the constants a and a' in the sensitivity function stated above are readily computed from the parameter probabilities of the feature variables in the naive Bayesian network; the two-way sensitivity function can in fact be established without the need of any propagation. We further note that if the probability of interest pertains to the value $\bar{c}$ of C, a similar property holds, with c and $\bar{c}$ interchanged.

We illustrate the two-way sensitivity function derived above by means of a naive Bayesian network for a real-life problem in dairy-farming practice.

Example 1. To distinguish between false positive (*fp*) and true positive (*tp*) alerts for clinical mastitis (an udder infection) in cows issued by an automatic milking system, various naive Bayesian networks were constructed [8]. These networks use information from milkings and information about the cows themselves to classify the issued alerts. Several of the variables involved are continuous by nature, and were discretised into two intervals each based upon domain knowledge. Among these variables is the somatic cell count (SCC), which indicates the number of white blood cells per ml milk. Now suppose that we construct a naive Bayesian network with just the feature variable SCC, and further suppose that we are interested in the output probability $\Pr(tp \mid SCC < t)$ for some threshold value t. Given that the automatic milking system issues a true positive alert with a prior probability of $\Pr(tp) = 0.014$, we can now readily determine the two-way sensitivity function that describes how the probability of interest is affected by varying the two parameter probabilities $x = p(SCC < t \mid tp)$ and $y = p(SCC < t \mid fp)$:

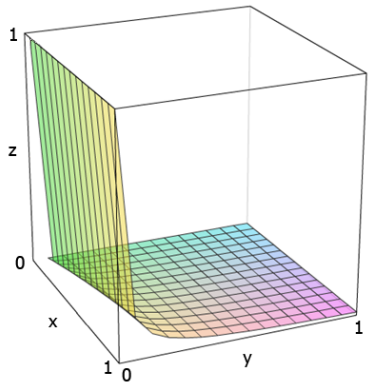
$$f_{\Pr(tp \mid SCC<t)}(x, y) = \frac{0.014 \cdot x}{0.014 \cdot x + 0.986 \cdot y}$$

Figure 1(a) shows the part of the function $f_{\Pr(tp \mid SCC<t)}(x, y)$ that lies within the unit cube; the function $f_{\Pr(fp \mid SCC<t)}(x, y)$ describing the effects of varying the same parameter probabilities x and y on the complementary output probability $\Pr(fp \mid SCC < t)$ is shown in Figure 1(b). From Figure 1(a), we read for example that a large probability $\Pr(tp \mid SCC < t)$ of an alert being truly positive given a small SCC value can be found only for relatively small values of the parameter y. □

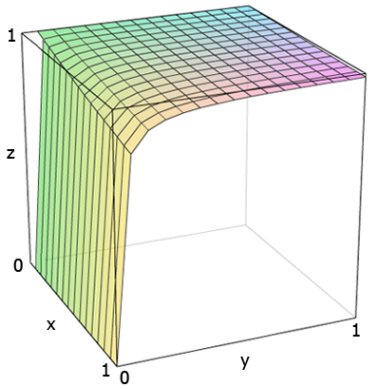
In Proposition 1, we derived the general form of a two-way sensitivity function which expresses an output probability computed from a naive Bayesian network in terms of two parameter probabilities from the conditional probability table of one of the network's feature variables. This two-way function specifies a value for the output probability of interest for each combination of values for the two parameters. We now recall that our aim is to use two-way sensitivity analysis as a means for studying the effects of changing the discretisation of a feature variable. In view of a variable's discretisation, the two parameters under study are not unrelated, as is assumed in a two-way sensitivity function in general. Since varying the threshold value t in a discretisation affects all parameter probabilities for a feature variable, the two parameters under study are dependent of t and are actually varied as $x(t)$ and $y(t)$; through the threshold value t moreover, the parameters are related as $y(t) = g(x(t))$ for some function g. Studying the effects of changing a discretisation thus requires the function

$$f_{\Pr(c \mid \mathbf{e})}(x(t), g(x(t))) = \frac{a_1 \cdot x(t) + a_2 \cdot g(x(t)) + a_3}{b_1 \cdot x(t) + b_2 \cdot g(x(t)) + b_3}$$

where the constants a_i, b_i, $i = 1, \ldots, 3$, again are built from the non-varied parameters of the network under study. Note that while this function essentially is a function in a single parameter probability, the effects of a change in discretisation cannot be described by a one-way sensitivity function in one of the parameters under study. To simplify our notations in the sequel, we will omit the explicit dependence of the parameter probabilities $x(t)$ and $y(t)$ on t and once again write x and y for short.



(a) $f_{\Pr(tp|SCC<t)}(x,y) = \frac{0.014 \cdot x}{0.014 \cdot x + 0.986 \cdot y} = z$



(b) $f_{\Pr(fp|SCC<t)}(x,y) = \frac{0.986 \cdot y}{0.986 \cdot y + 0.014 \cdot x} = z$

Figure 1: Two-way sensitivity functions for the alert variable given an SCC value smaller than t, for the parameter probabilities $x = p(SCC < t \mid tp)$ and $y = p(SCC < t \mid fp)$ assuming functional independence.

We now observe that the sensitivity function in x and $g(x)$ described above can only be detailed if we can formally specify the function g which describes the relation between the parameters x and y through t. From the way in which discretisations are formalised, we have that the function g cannot be any arbitrary function: we know that the function is either monotonically non-decreasing or monotonically non-increasing.

Lemma 1. *Let E_i be a binary discretised feature variable in a naive Bayesian network with the class variable C. Now, let $x = p(e_i' \mid c)$ and $y = p(e_i'' \mid \bar{c})$ be parameter probabilities for the values e_i', e_i'' of E_i, and let g be the function with $y = g(x)$. Then, g is monotonically non-decreasing if $e_i' = e_i''$ and monotonically non-increasing otherwise.*

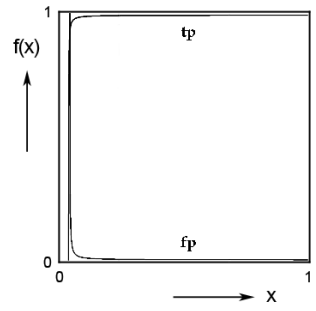
Proof. The property stated in the lemma is closely related to the well-known interdependence of test characteristics from the field of epidemiology [13]. The property can be easily verified by observing that as the threshold t is shifted to larger values of the continuous variable E_i, the probability $p(E_i < t \mid C)$ cannot decrease regardless of the value of C; similarly, the probability $p(E_i \geq t \mid C)$ cannot increase. □

From Lemma 1 we have that the function g which relates the two parameter probabilities x and y through the discretisation under study, is monotonically non-decreasing if x and y pertain to the same feature value; the function g is monotonically non-increasing otherwise. The precise form of the function g is defined by further knowledge of the problem domain at hand.

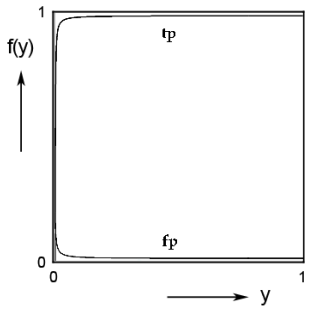
Example 2. We consider again our problem of classifying alerts for clinical mastitis in cows. For constructing our naive Bayesian networks, we had available a large number of real data from milkings on dairy farms. These data show that the true function which relates the parameter probabilities $x = p(SCC < t \mid tp)$ and $y = p(SCC < t \mid fp)$ through t, can be approximated by a linear function. By means of linear regression of y on x, we found that the function $y = 1.1700 \cdot x - 0.0478$ fitted the available data best. We now recall that the surface $f_{\Pr(tp|SCC<t)}(x,y)$ from Figure 1(a) described the probability of interest $\Pr(tp \mid SCC < t)$ in terms of the two parameters x and y under the assumption of functional independence. By intersecting this surface with the plane $y = 1.1700 \cdot x - 0.0478$, we can now find the function that expresses $\Pr(tp \mid SCC < t)$ in terms of x taking its actual relationship with y into consideration. The intersection curve thus describes the sensitivity of the probability of interest $\Pr(tp \mid SCC < t)$ to changes occasioned in x as a result of varying the discretisation threshold value t. The curve is defined as:

$$f_{\Pr(tp|SCC<t)}(x) = \frac{0.014 \cdot x}{0.014 \cdot x + 0.986 \cdot (1.17 \cdot x - 0.0478)}$$

Figure 2(a) displays this function, along with the function for the complement of the probability of interest. We observe that the depicted functions do not specify a value for the probability of interest for very small values of the parameter x. This finding originates from the functional dependence of y on x: for very small values for x, there are no matching values for y within the range $\langle 0, 1 \rangle$. The finding underlines our previous observation that the depicted functions are no one-way sensitivity functions, but two-sensitivity functions



(a) Functions tp $= f_{\Pr(tp|SCC<t)}(x)$ and fp $= f_{\Pr(fp|SCC<t)}(x)$, with $x = p(SCC < t \mid tp)$.



(b) Functions tp $= f_{\Pr(tp|SCC<t)}(y)$ and fp $= f_{\Pr(fp|SCC<t)}(y)$, with $y = p(SCC < t \mid fp)$.

Figure 2: The dimension-reduced sensitivity functions for the alert variable given an SCC value smaller than t, taking into account the functional dependence between x and y.

instead that are reduced in dimension as a result of the functional dependence of y on x. In Figure 2(b), the two functions are shown again, this time from the perspective of the parameter y; the function describing the dependence of x on y was approximated by linear regression to $x = 0.845 \cdot y + 0.0448$. □

4 Studying the Effects of Discretisation on the Most Likely Class Value

In the previous section, we established sensitivity functions that describe the effects of changing the discretisation of a continuous-valued feature variable on an output probability of interest computed from a naive Bayesian network. In practice now, we are often not so much interested in an exact output probability, but in the most likely value of the output variable instead. Of particular interest then becomes the question whether changing the discretisation of a feature variable can result in a change of the output value. In general, we have that parameter variation can result in such a change only if the sensitivity functions for the different values of the class variable intersect. For the binary class variable and parameters presently under study, we thus have that if the sensitivity functions for the complementary class probabilities intersect within the unit cube, then a shift in value for the parameters can occasion a change in the most likely class value.

To address the question whether choosing an alternative discretisation for a feature variable E_i can result in a change in the most likely value of the class variable C, we consider the two sensitivity functions $f_{\Pr(c|\mathbf{e})}(x, y)$ and $f_{\Pr(\bar{c}|\mathbf{e})}(x, y)$ expressing the complementary probabilities $\Pr(c \mid \mathbf{e})$ and $\Pr(\bar{c} \mid \mathbf{e})$ in terms of the parameters $x = p(e_i' \mid c)$ and $y = p(e_i' \mid \bar{c})$. As before, we disregard at first the functional dependence between the two parameter probabilities. For ease of exposition we further assume that the joint combination of observed values $\mathbf{e}$ includes the value e_i' for E_i. We now observe that the two class probabilities are equal for values of the parameters x and y for which $f_{\Pr(c|\mathbf{e})}(x, y) = f_{\Pr(\bar{c}|\mathbf{e})}(x, y) = 0.5$. From this observation, we have that the intersection of the two sensitivity functions is located in the horizontal plane $f(x, y) = 0.5$. Using Proposition 1, we further find that the two functions intersect by a linear line:

$$f_{\Pr(c|\mathbf{e})}(x, y) = f_{\Pr(\bar{c}|\mathbf{e})}(x, y) \iff \frac{a \cdot \Pr(c) \cdot x}{a \cdot \Pr(c) \cdot x + a' \cdot \Pr(\bar{c}) \cdot y} = \frac{a' \cdot \Pr(\bar{c}) \cdot y}{a' \cdot \Pr(\bar{c}) \cdot y + a \cdot \Pr(c) \cdot x}$$

$$\iff y = \frac{a \cdot \Pr(c)}{a' \cdot \Pr(\bar{c})} \cdot x$$

Recall that we have assumed that the prior probabilities of the class values are non-zero and that $x, y \in \langle 0, 1 \rangle$. Our assumption of non-zero parameter probabilities in general further guarantees that the constants a and a' are greater than zero, from which we conclude that the intersection line has a positive gradient. The following lemma now shows that the intersection line passes through the unit cube.

Lemma 2. *Let E_i be a binary discretised feature variable in a naive Bayesian network with the class variable C and feature variables $\mathbf{E}$. Let $\mathbf{e}$ be the combination of actually observed values for $\mathbf{E}$ and let e_i' be the value of E_i in $\mathbf{e}$. Now, let $x = p(e_i' \mid c)$ and $y = p(e_i' \mid \bar{c})$ be parameter probabilities for E_i. Then, the sensitivity functions $f_{\Pr(c|\mathbf{e})}(x, y)$ and $f_{\Pr(\bar{c}|\mathbf{e})}(x, y)$ intersect within the unit cube.*

Proof. As argued above, the intersection of the sensitivity functions for the two class probabilities $\Pr(c \mid \mathbf{e})$ and $\Pr(\bar{c} \mid \mathbf{e})$ is a line with the constant $(a \cdot \Pr(c))/(a' \cdot \Pr(\bar{c}))$ for its gradient. We further have seen that the gradient of the intersection line is positive. Since the constants a, a' and $\Pr(c), \Pr(\bar{c})$ involved are all smaller than one, we find that there exists at least one feasible combination of values for the parameter probabilities x and y with $x, y \in \langle 0, 1 \rangle$: an example of such a combination is $x = a' \cdot \Pr(\bar{c})$, $y = a \cdot \Pr(c)$. We thus have that the intersection line has a point within the unit cube from which we conclude that the two sensitivity functions intersect within the cube. □

So far we have not taken into consideration the functional relationship $y = g(x)$ which exists between the parameter probabilities x and y as a result of the discretisation. To study the effects of this relationship, we now consider a specific discretisation of the feature variable E_i under study; we suppose that this discretisation is defined by the threshold value t_0 and has associated the values x_0, y_0 with $y_0 = g(x_0)$ for the parameter probabilities x and y respectively. Without loss of generality, we assume that with the parameter values x_0, y_0 and with the evidence under consideration, the most likely value of the class variable is equal to c. We observe that a change of this value can occur only if there exists a discretisation threshold t_1 with associated parameter values x_1, y_1 with $y_1 = g(x_1)$, with which $\bar{c}$ is the most likely class value. Note that the two pairs of parameter values (x_0, y_0) and (x_1, y_1) would then lie on opposite sides of the intersection line of the two sensitivity functions established above, when projected onto the horizontal plane $f(x, y) = 0.5$ within the unit cube. We now observe that a discretisation threshold t_1 for which $\bar{c}$ is the most likely class value exists only if the intersection of the two sensitivity functions intersects itself with the function $y = g(x)$ within the unit cube, that is, if there exist values $x, g(x) \in \langle 0, 1 \rangle$ for which

$$g(x) = \frac{a \cdot \Pr(c)}{a' \cdot \Pr(\bar{c})} \cdot x$$

If such values exist, then a discretisation threshold t_1 exists, with associated parameter values x_1, y_1, with which a change in the most likely class value will result; otherwise, the most likely class value returned by the naive Bayesian network cannot be changed by choosing a different discretisation. We would like to note that although an appropriate pair x_1, y_1 of parameter values can be readily determined from the various functions involved, a method for effectively establishing an associated discretisation threshold is yet to be designed. The following example moreover shows that existence of an appropriate pair of values in theory may not always make a realistic discretisation in practice.

Example 3. We consider again our problem of classifying alerts for clinical mastitis in dairy cows. In the naive Bayesian network under consideration, we consider the parameter probabilities $x = p(SCC < t \mid tp)$ and $y = p(SCC < t \mid fp)$ for the feature variable SCC. By setting the discretisation threshold for SCC at $t_0 = 500,000$ cells per ml milk, the values for the parameters x and y were established from the available data to be $x_0 = 0.67$ and $y_0 = 0.76$, respectively. With these parameter values, an alert for a milking with $SCC < t$ is most likely to be false, that is, *fp* is the most likely class value. To study whether this most likely value can be changed by choosing another threshold value for the variable SCC in our simple network, we first consider the intersection line of the two sensitivity functions $f_{\Pr(tp|SCC<t)}(x, y)$ and $f_{\Pr(fp|SCC<t)}(x, y)$. This line is found to be equal to

$$y = \frac{\Pr(c)}{\Pr(\bar{c})} \cdot x = \frac{0.014}{0.986} \cdot x$$

located at $f(x, y) = 0.5$ within the unit cube. We now recall that the functional dependence between the parameter probabilities x and y was approximated by the linear function $y = 1.1700 \cdot x - 0.0478$. The intersection line of the two sensitivity functions intersects with this linear function at

$$\frac{0.0140}{0.9860} \cdot x = 1.17 \cdot x - 0.0478$$

that is, at $x = 0.0414$. Given the functional relationship between x and y, the corresponding value for the parameter y is found to be 0.0006. We conclude that, in theory, it is possible to set a threshold value t, with the associated parameter values x, y with $x < x_0$ and $y < y_0$, that changes the most likely class value from *fp* to *tp*. Actually finding such a threshold value may be a challenge, however: to establish it from data would require a very large dataset in order for such small values of x and y to be estimated reliably. In addition, small values for x and y could be achieved only by setting the threshold value to an unrealistically low level of SCC. We conclude that in our domain of application, there exists no acceptable alternative discretisation for the feature variable SCC that would change the most likely class value in the presence of observed small SCC values. □

5 Conclusions and Further Research

Focusing on binary naive Bayesian networks, we initiated a study into the effects of changing the discretisations of a network's feature variables on the posterior probabilities computed for its class variable. We showed that recent insights from the field of sensitivity analysis of Bayesian networks serve to analytically describe the effects of changing discretisations. We argued more specifically that changing the discretisation of a feature variable occasions shifts in the values of all its parameter probabilities. We showed how the functional relationship that is thus induced between these parameters, can be explicitly taken into account upon establishing the required sensitivity functions. We further described a method for establishing whether or not an alternative discretisation can change the class values established from a naive Bayesian network. This method has theoretical value, but currently is not yet practically applicable. So far moreover, we focused on binary discretisations and binary class variables, which were reasonable restrictions for our domain of application. More generally, we envision a similar approach for studying the effects of discretisations in more than two intervals and for non-binary class variables, which will involve higher-dimensional sensitivity functions. In the near future, we will focus on these issues and hope to arrive at a general, practically applicable method for studying the effects of discretisations in Bayesian networks.

Acknowledgement We are most grateful to Henk Hogeveen and Wilma Steeneveld from Wageningen University, the Netherlands, for supplying us with their dairy-farm data.

References

[1] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

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

[3] J. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretization of continuous features. In: S.J. Russel (editor). *Proceedings of the 12th International Conference on Machine Learning*, pp. 194 – 202, 1995. Morgan Kaufmann.

[4] L. Uusitalo. Advantages and challenges of Bayesian networks in environmental modelling. *Ecological modelling*, vol. 203, pp. 312 – 318, 2007.

[5] P. Myllymäki, T. Silander, H. Tirri, and P. Uronen. B-Course: a web-based tool for Bayesian and causal data analysis. *International Journal of Artificial Intelligence Tools*, vol. 11, pp. 369 – 387, 2002.

[6] K.B. Laskey. Sensitivity analysis for probability assessments in Bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 25, pp. 901 – 909, 1995.

[7] V.M.H. Coupé and L.C. van der Gaag. Properties of sensitivity analysis of Bayesian belief networks. *Annals of Mathematics and Artificial Intelligence*, vol. 36, pp. 323 – 356, 2002.

[8] W. Steeneveld, L.C. van der Gaag, W. Ouweltjes, H. Mollenhorst, and H. Hogeveen. Discriminating between true-positive and false-positive clinical mastitis alerts from automatic milking systems. *Journal of Dairy Science*, vol. 93, pp. 2559 – 2568, 2010.

[9] L.C. van der Gaag, S. Renooij, and V.M.H. Coupé. Sensitivity analysis of probabilistic networks. In: P. Lucas, J. Gámez, and A. Salmerón (editors). *Advances in Probabilistic Graphical Models*, Studies in Fuzziness and Soft Computing, vol. 214, pp. 103 – 124. Springer, 2007.

[10] U. Kjærulff and L.C. van der Gaag. Making sensitivity analysis computationally efficient. *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pp. 317 – 325, 2000. Morgan Kaufmann.

[11] S. Renooij. Bayesian network sensitivity to arc-removal. In: P. Myllymaki, T. Roos, and T. Jaakkola (editors). *Proceedings of the Fifth European Workshop on Probabilistic Graphical Models*, pp. 233 – 240, 2010. HIIT Publications.

[12] S. Renooij and L.C. van der Gaag. Evidence and scenario sensitivities in naive Bayesian classifiers. *International Journal of Approximate Reasoning*, vol. 49, pp. 398 – 416, 2008.

[13] B. Dawson-Saunders and R.G. Trapp. *Basic & Clinical Biostatistics*, 2001. McGraw-Hill, New York.