Robustness of Multi-dimensional Bayesian Network Classifiers

Janneke H. Bolt Silja Renooij

Department of Information and Computing Sciences, Utrecht University
P.O. Box 80.089, 3508 TB Utrecht, The Netherlands

Abstract

Multi-dimensional Bayesian network classifiers (MDCs) generalise the popular robustly performing one-dimensional classifiers (ODCs) to application domains that require an instance to be classified into a combination of classes. In previous work we compared the sensitivity of MDC and ODC output probabilities to small parameter inaccuracies. In this paper we extend our analyses and study the robustness of the classification performance of MDCs.

1 Introduction

Bayesian networks are powerful tools for supporting decisions under uncertainty. A Bayesian network defines a joint probability distribution over a set of stochastic variables by combining a directed acyclic graph and a set of (conditional) probability distributions [10]. Bayesian networks are often used in the context of classification, where an input instance is classified into one of several distinct classes. For such classification tasks, one-dimensional Bayesian network classifiers (ODCs) are very popular [8]. An ODC is a Bayesian network of restricted topology, consisting of a single class variable and several feature variables. More recently, multi-dimensional Bayesian network classifiers (MDCs) were introduced to generalise ODCs to application domains that require an instance to be classified into a combination of classes [7, 13], represented by a set of class variables. MDCs have gained a growing interest as tool for multi-dimensional classification [1, 3].

Classification performance of ODCs is known to be rather good. This claim is supported by experimental results and further substantiated theoretically in among others [5, 11]. In this paper we address the robustness of MDCs. We extend our previous paper [2] in which we studied the effects of local parameter changes on the output probabilities of MDCs and argued that MDCs in general can be expected to be even more robust with respect to such changes than ODCs. Since, in classification tasks, we are often more interested in the most likely combination of classes as output than in exact output probabilities, in this paper we study the robustness of the classification output of MDCs. To this end we use sensitivity functions for MDCs, thereby providing an insightful alternative to [4] where arithmetic circuits where used to assess classification robustness. We express intervals of parameter values for which an original classification remains unchanged in just a few meaningful probabilities and we show that a classification can change at most once given an increasing or decreasing parameter value.

The paper is organised as follows. In Section 2 we provide some preliminaries on Bayesian networks for multi-dimensional classification and on sensitivity analysis. In Section 3 we review our sensitivity functions for MDCs from [2] and include our proofs for their validity. In section 4.1 we review our conclusions with respect to the relative sensitivity of MDCs and in Section 4.2 we give intervals for admissible deviation for their feature parameters and root class parameters. We end our paper with an example and a concluding section.

2 Preliminaries

2.1 Bayesian Networks, ODCs and MDCs

A *Bayesian network* is a graphical model of a joint probability distribution $\Pr$ over a set of stochastic variables $\mathbf{V} = \{V_i, \ldots, V_n\}$. We will denote some value assignment to V_i by v_i and a joint value assignment to $\mathbf{V}$ by $\mathbf{v}$. In the sequel we will use V_i and $\mathbf{V}$ also to indicate the set of possible assignments to V_i and $\mathbf{V}$, respectively. In a Bayesian network, each variable of the modelled distribution is represented by a node in a directed acyclic graph[1]. Independences between the variables are, as far as possible, captured by the digraph's set of arcs according to the d-separation criterion [10]. Moreover, for each variable V_i, conditional probability distributions $\Pr(V_i \mid \pi_{V_i})$ are specified, where π_{V_i} denotes a joint value assignment to the set of parents of V_i in the digraph. These (conditional) probabilities, or network parameters, together define the joint probability distribution

$$\Pr(\mathbf{V}) = \prod_{V_i \in \mathbf{V}} \Pr(V_i \mid \pi_{V_i})$$

where each assignment to V_i and π_{V_i} is compatible with the joint assignment to $\mathbf{V}$ under consideration. We will use the signs $\sim$ and $\nsim$ to indicate compatibility and non-compatibility of variable assignments, respectively. For example, $abc \sim ab$ and $ab \nsim a\bar{b}c$. Furthermore we will use x_0 to indicate original parameter values and $\Pr_0$ for probabilities computed with the original values of all parameters involved.

An MDC is a Bayesian network in which the variables are divided into a set of class variables $\mathbf{C} = \{C_1, \ldots, C_n\}$ and a set of feature variables $\mathbf{F} = \{F_1, \ldots, F_m\}$. For a variable V_i, we use $\pi_{\mathbf{F_i}}$ to denote those parents of V_i that are in $\mathbf{F}$, and $\pi_{\mathbf{C_i}}$ to denote those parents of V_i that are in $\mathbf{C}$; instantiations to these sets are indicated by $\pi_{\mathbf{f_i}}$ and $\pi_{\mathbf{c_i}}$, respectively. In an MDC the relationship between the set of class variables and the set of feature variables is restricted in the sense that class variables are not allowed to have feature parents [7, 13], that is, $\pi_{\mathbf{F_i}} = \emptyset$ for $V_i \in \mathbf{C}$. In case of a *naive* classifier feature variables do not have feature parents either, that is, $\pi_{\mathbf{F_i}} = \emptyset$ for $V_i \in \mathbf{F}$ as well and class variables do not have class parents, that is $\pi_{\mathbf{C_i}} = \emptyset$ for $V_i \in \mathbf{C}$. In this paper we consider MDCs in which no further assumptions are made concerning the relationships among the class variables, or among the feature variables. An ODC is an MDC with just a single class variable.

An MDC can be used to assign some instance $\mathbf{f}$, that is, a joint value assignment to the feature variables, to a most likely combination of classes. As such, it computes $\mathrm{argmax}_{\mathbf{c}} \Pr(\mathbf{c} \mid \mathbf{f})$; note that this is not necessarily the combination of most likely c_i's from the marginals $\Pr(C_i \mid \mathbf{f})$.

2.2 Sensitivity analysis

The parameters of a network are elicited from data or experts and are inevitable inaccurate. To investigate the effects of inaccuracies in its parameters, a Bayesian network can be subjected to a sensitivity analysis. In a sensitivity analysis, network parameters are varied and some probability of interest as a function of the varied parameters is computed. If just one parameter $x_i = \Pr(v \mid \pi_V)$ for a variable V is varied, the effects on a probability of interest $\Pr(y \mid \mathbf{e})$ are captured by a function of the form:

$$f_{\Pr(y|\mathbf{e})}(x_i) = \frac{a \cdot x_i + b}{c \cdot x_i + d}$$

where the constants a, b, c and d are constructed from the non-varied parameters. Note that the other parameters $x_j \neq x_i$ of the same conditional distribution over V need to be co-varied to ensure that the distribution sums to 1. In this paper we will, as is standard, co-vary these parameters proportionally, that is, $x_j = x_j^o \cdot (1 - x_i)/(1 - x_i^o)$. Moreover we will assume that deterministic parameters are not varied and that varied parameters will not adopt deterministic values.

From a sensitivity function various properties can be computed. The *sensitivity value* [9] is the absolute value of the first derivative of the sensitivity function at the original assessment x_0 of the parameter, that is, $\left|\frac{\delta f}{\delta x}(x_0)\right|$. High values indicate a high sensitivity of the outcome of the probability of interest to parameter changes. Sensitivity values above 1 are considered to express a higher sensitivity. The *admissible deviation* [6] describes the interval $[\alpha, \beta]$ of parameter values for which the original most

[1] From now on, the terms node and variable will be used interchangeably.

likely (joint) value remains unchanged. Not only the classification of some instance may be of interest, it may also be of importance how well the classifier can discriminate between different outcomes. Change in discrimination ability can also be expressed by a sensitivity function [12].

3 One-way Sensitivity functions of MDCS

In previous work we introduced, without proofs, one-way sensitivity functions for outcomes $\Pr(\mathbf{c} \mid \mathbf{f})$ of MDCs [2]. Here we review the results necessary for our further analyses and include the proofs.

Proposition 1. Let $\Pr(\mathbf{c} \mid \mathbf{f})$ be an outcome probability of an MDC($\mathbf{C}, \mathbf{F}$), that is, an MDC with class variables $\mathbf{C}$ and feature variables $\mathbf{F}$ and let $x = \Pr(f_i \mid \pi_{\mathbf{f_i}} \pi_{\mathbf{c_i}})$ be a feature parameter with $\pi_{\mathbf{f_i}} \sim \mathbf{f}$[2]. The sensitivity function $f_{\Pr(\mathbf{c}|\mathbf{f})}(x)$ has one of the following forms

	$\pi_{\mathbf{c_i}} \sim \mathbf{c}$	$\pi_{\mathbf{c_i}} \nsim \mathbf{c}$
$f_i \sim \mathbf{f}$	$\frac{x \cdot \Pr_0(\mathbf{c}\mid\mathbf{f})}{(x-x_0)\cdot\Pr_0(\pi_{\mathbf{c_i}}\mid\mathbf{f})+x_0}$	$\frac{x_0 \cdot \Pr_0(\mathbf{c}\mid\mathbf{f})}{(x-x_0)\cdot\Pr_0(\pi_{\mathbf{c_i}}\mid\mathbf{f})+x_0}$
$f_i \nsim \mathbf{f}$	$\frac{(1-x) \cdot \Pr_0(\mathbf{c}\mid\mathbf{f})}{(x_0-x)\cdot\Pr_0(\pi_{\mathbf{c_i}}\mid\mathbf{f})+1-x_0}$	$\frac{(1-x_0) \cdot \Pr_0(\mathbf{c}\mid\mathbf{f})}{(x_0-x)\cdot\Pr_0(\pi_{\mathbf{c_i}}\mid\mathbf{f})+1-x_0}$

Proof. We first detail the relation between $x = \Pr(f_i \mid \pi_{\mathbf{f_i}} \pi_{\mathbf{c_i}})$ and joint probabilities $\Pr(\mathbf{c}^*\, \mathbf{f})$, $\pi_{\mathbf{f_i}} \sim \mathbf{f}$:

$$\Pr(\mathbf{c}^*\, \mathbf{f})(x) = \begin{cases} \Pr_0(\mathbf{c}^*\, \mathbf{f}) & \text{for any} \quad \mathbf{c}^* \nsim \pi_{\mathbf{c_i}} \\ \Pr_0(\mathbf{c}^*\, \mathbf{f}) \cdot x/x_0 & \text{for any} \quad \mathbf{c}^* \sim \pi_{\mathbf{c_i}},\ \mathbf{f} \sim f_i \\ \Pr_0(\mathbf{c}^*\, \mathbf{f}) \cdot (1-x)/(1-x_0) & \text{for any} \quad \mathbf{c}^* \sim \pi_{\mathbf{c_i}},\ \mathbf{f} \nsim f_i \end{cases}$$

where the bottom result is due to proportional co-variation.

Now consider the case where $\pi_{\mathbf{c_i}} \sim \mathbf{c}$ and $f_i \sim \mathbf{f}$. We find

$$f_{\Pr(\mathbf{c}|\mathbf{f})}(x) = \frac{\Pr_0(\mathbf{c}\, \mathbf{f}) \cdot x/x_0}{\sum_{\mathbf{c}^* \sim \pi_{\mathbf{c_i}}} \Pr_0(\mathbf{f}\, \mathbf{c}^*) \cdot x/x_0 + \sum_{\mathbf{c}^* \nsim \pi_{\mathbf{c_i}}} \Pr_0(\mathbf{f}\, \mathbf{c}^*)} \tag{1}$$

Using $\sum_{\mathbf{c}^* \nsim \pi_{\mathbf{c_i}}} \Pr(\mathbf{f}\, \mathbf{c}^*) = \Pr(\mathbf{f}) - \sum_{\mathbf{c}^* \sim \pi_{\mathbf{c_i}}} \Pr(\mathbf{f}\, \mathbf{c}^*)$, and using that $\sum_{\mathbf{c}^* \sim \pi_{\mathbf{c_i}}} \Pr(\mathbf{f}\, \mathbf{c}^*) = \Pr(\mathbf{f} \pi_{\mathbf{c_i}})$ which marginalises out all class variables outside subset $\pi_{\mathbf{C_i}}$, and then dividing all terms by $\Pr_0(\mathbf{f})$ gives

$$\begin{aligned} f_{\Pr(\mathbf{c}|\mathbf{f})}(x) &= \frac{\Pr_0(\mathbf{c}\, \mathbf{f}) \cdot x/x_0}{x/x_0 \cdot \Pr_0(\mathbf{f} \pi_{\mathbf{c_i}}) + \Pr_0(\mathbf{f}) - \Pr_0(\mathbf{f} \pi_{\mathbf{c_i}}) \cdot x_0/x_0} \\ &= \frac{\Pr_0(\mathbf{c}\, \mathbf{f}) \cdot x/x_0}{(x-x_0)/x_0 \cdot \Pr_0(\mathbf{f} \pi_{\mathbf{c_i}}) + \Pr_0(\mathbf{f})} \\ &= \frac{\Pr_0(\mathbf{c} \mid \mathbf{f}) \cdot x/x_0}{(x-x_0)/x_0 \cdot \Pr_0(\pi_{\mathbf{c_i}} \mid \mathbf{f}) + 1} \\ &= \frac{\Pr_0(\mathbf{c} \mid \mathbf{f}) \cdot x}{(x-x_0) \cdot \Pr_0(\pi_{\mathbf{c_i}} \mid \mathbf{f}) + x_0} \end{aligned}$$

For the case where $\pi_{\mathbf{c_i}} \sim \mathbf{c}$, yet $f_i \nsim \mathbf{f}$, we observe that x/x_0 in Equation (1) needs to be replaced by $(1-x)/(1-x_0)$ both in the numerator and in the denominator. For the cases where $\pi_{\mathbf{c_i}} \nsim \mathbf{c}$, the numerator in Equation (1) remains constant. Subsequently following similar steps as above serves to prove the remaining results stated in Proposition 1. □

The following proposition provides the sensitivity function for a root class parameter, that is, a parameter of a class variable without parents. Special cases of MDCs that assume class variables to be independent contain only root class variables; MDCs in general contain at least one such variable.

[2] If $\pi_{\mathbf{f_i}} \nsim \mathbf{f}$ then $\Pr(\mathbf{c} \mid \mathbf{f})$ is not affected by a change of $\Pr(f_i \mid \pi_{\mathbf{f_i}} \pi_{\mathbf{c_i}})$ and remains constant.

Proposition 2. Let $\Pr(\mathbf{c} \mid \mathbf{f})$ be an outcome probability of an MDC$(\mathbf{C}, \mathbf{F})$ as before and let $x = \Pr(c_i)$ be a parameter of a root class variable. The sensitivity function $f_{\Pr(\mathbf{c}|\mathbf{f})}(x)$ has one of the following forms:

$$f_{\Pr(\mathbf{c}|\mathbf{f})}(x) = \frac{x \cdot (1 - x_0) \cdot \Pr_0(\mathbf{c} \mid \mathbf{f})}{(x - x_0) \cdot \Pr_0(c_i \mid \mathbf{f}) + (1 - x) \cdot x_0}, \ \text{ if } c_i \sim \mathbf{c}$$

$$f_{\Pr(\mathbf{c}|\mathbf{f})}(x) = \frac{(1 - x) \cdot x_0 \cdot \Pr_0(\mathbf{c} \mid \mathbf{f})}{(x - x_0) \cdot \Pr_0(c_i \mid \mathbf{f}) + (1 - x) \cdot x_0}, \ \text{ if } c_i \not\sim \mathbf{c}$$

Proof. We first detail the relation between $x = \Pr(c_i)$ and joint probabilities $\Pr(\mathbf{c}^* \, \mathbf{f})$:

$$\Pr(\mathbf{c}^* \, \mathbf{f})(x) = \begin{cases} \Pr_0(\mathbf{c}^* \, \mathbf{f}) \cdot x/x_0 & \text{for any} \quad \mathbf{c}^* \sim c_i \\ \Pr_0(\mathbf{c}^* \, \mathbf{f}) \cdot (1 - x)/(1 - x_0) & \text{for any} \quad \mathbf{c}^* \not\sim c_i \end{cases}$$

where the bottom result is due to proportional co-variation.

Now consider the case where $c_i \sim \mathbf{c}$. We find

$$f_{\Pr(\mathbf{c}|\mathbf{f})}(x) = \frac{\Pr_0(\mathbf{c}\,\mathbf{f}) \cdot x/x_0}{\Pr_0(\mathbf{f}\,c_i) \cdot x/x_0 + \sum_{c_i^* \neq c_i} \Pr_0(\mathbf{f}\,c_i^*) \cdot (1 - x)/(1 - x_0)} \tag{2}$$

Using $\sum_{c_i^* \neq c_i} \Pr(\mathbf{f}\,c_i^*) = \Pr(\mathbf{f}) - \Pr(\mathbf{f}\,c_i)$, subsequent simplification and division of all terms by $\Pr_0(\mathbf{f})$ gives

$$\begin{aligned} & f_{\Pr(\mathbf{c}|\mathbf{f})}(x) = \\ = \ & \frac{\Pr_0(\mathbf{c}\,\mathbf{f}) \cdot x/x_0}{\Pr_0(\mathbf{f}\,c_i) \cdot x/x_0 + (\Pr_0(\mathbf{f}) - \Pr_0(\mathbf{f}\,c_i)) \cdot (1 - x)/(1 - x_0)} \\ = \ & \frac{\Pr_0(\mathbf{c}\,\mathbf{f}) \cdot x/x_0}{\Pr_0(\mathbf{f}\,c_i) \cdot (x/x_0) \cdot (1 - x_0)/(1 - x_0) + (\Pr_0(\mathbf{f}) - \Pr_0(\mathbf{f}\,c_i)) \cdot (1 - x)/(1 - x_0) \cdot x_0/x_0} \\ = \ & \frac{\Pr_0(\mathbf{c}\,\mathbf{f}) \cdot x \cdot (1 - x_0)}{(x \cdot (1 - x_0) - x_0 \cdot (1 - x)) \cdot \Pr_0(\mathbf{f}\,c_i) + x_0 \cdot (1 - x) \cdot \Pr_0(\mathbf{f})} \\ = \ & \frac{\Pr_0(\mathbf{c} \mid \mathbf{f}) \cdot x \cdot (1 - x_0)}{(x - x_0) \cdot \Pr_0(c_i \mid \mathbf{f}) + x_0 \cdot (1 - x)} \end{aligned}$$

For the case where $c_i \not\sim \mathbf{c}$ the proof is analogous: we just need to replace x/x_0 by $(1 - x)/(1 - x_0)$ in the numerator of Equation (2). □

With $\mathbf{C}$ reduced to a single variable, the propositions above apply for ODCs. The functions in that case are equal to the functions found in [11] for *naive* ODCs. Our proofs show that in fact the results in [11] apply to ODCs in general.

4 Sensitivity properties of MDCs

In Section 4.1 we shortly review our results with respect to the sensitivity value of MDCs from [2]. In Section 4.2 we provide intervals of admissible deviation for the feature parameters and the root class parameters of an MDC.

4.1 Sensitivity Value

As mentioned in Section 2.2, a sensitivity value > 1 is considered to express a higher sensitivity of some outcome to a local parameter change. From the sensitivity functions given in the previous section we derived expressions for the sensitivity value $\left|\frac{df}{dx}(x_0)\right|$. We used these expressions to establish for which proportion of combinations of the terms x_0, $\mathrm{Pr}_0(\mathbf{c} \mid \mathbf{f})$, and $\mathrm{Pr}_0(\pi_{\mathbf{c_i}} \mid \mathbf{f})$ or $\mathrm{Pr}_0(c_i \mid \mathbf{f})$ a sensitivity value > 1 will be found and we compared MDCs in general to the special case of ODCs in this respect.

For feature parameters we found that, for MDCs in general approximately 8% of the independently chosen feasible combinations of terms has a sensitivity value > 1, whereas for the special case of ODCs this percentage is approximately 17%. For root class parameters we found percentages of respectively approximately 20% and 50%. For MDCs in general these percentages thus are considerably lower than for ODCs. The percentages given above hold if terms are chosen independently. The terms, however, are in fact related. We argued that although the dependencies between the terms will change the actual percentages, the percentages will remain smaller for MDCs. All in all we concluded that MDCs will in general be less sensitive to feature parameter changes and root class parameter changes than ODCs.

4.2 Admissible deviation

In view of the classification task of an MDC, the property of admissible deviation is of importance. The admissible deviation gives the amount of variation that is allowed before an instance is classified as belonging to a different class. In [4] a method is given to determine the robustness of a most probable explanation (MPE) to parameter change. A Bayesian network is transformed into an arithmetic circuit from which two constants are computed that are used to assess for which parameter values the MPE may change. Since classification with an MDC is a special case of the MPE-problem, this method can provide for establishing admissible deviations for an MDC. Here we present an alternative approach in which the admissible deviations for an MDC are derived from the intersections of the appropriate sensitivity functions. Our approach is more simple since a transformation of the network is not required and more insightful since it yields formulas that express the admissible deviations in just a few meaningful probabilities. In Lemma 1 we first prove that a classification can change at most once given a changing x. We use this lemma in Proposition 3 in which the intervals of admissible deviation are given.

Lemma 1. Let MDC($\mathbf{C}, \mathbf{F}$) be an MDC as before, let $\mathbf{f}$ be an instance, let $\mathbf{c}^{\mathbf{max}} = \mathrm{argmax}_{\mathbf{c} \in \mathbf{C}} \Pr(\mathbf{c} \mid \mathbf{f})$ be the classification of $\mathbf{f}$ and let $x = \Pr(f_i \mid \pi_{\mathbf{f_i}} \pi_{\mathbf{c_i}})$ be a feature parameter with $\pi_{\mathbf{f_i}} \sim \mathbf{f}$. Let $\mathbf{C}^{\sim} = \{\mathbf{c} \mid \mathbf{c} \sim \pi_{c_i}\}$ and $\mathbf{C}^{\nsim} = \{\mathbf{c} \mid \mathbf{c} \nsim \pi_{c_i}\}$ be the instantiations of the class variables, respectively compatible and incompatible with x and let $\mathbf{c}^{\mathbf{max}\sim}$ be $\mathrm{argmax}_{\mathbf{c} \in \mathbf{C}^{\sim}} \Pr(\mathbf{c} \mid \mathbf{f})$ and $\mathbf{c}^{\mathbf{max}\nsim}$ be $\mathrm{argmax}_{\mathbf{c} \in \mathbf{C}^{\nsim}} \Pr(\mathbf{c} \mid \mathbf{f})$. The classification $\mathbf{c}^{\mathbf{max}}$ can change at most once given a decreasing or an increasing x; either from $\mathbf{c}^{\mathbf{max}\sim}$ to $\mathbf{c}^{\mathbf{max}\nsim}$ or from $\mathbf{c}^{\mathbf{max}\nsim}$ to $\mathbf{c}^{\mathbf{max}\sim}$. For a root class parameter the lemma is analogous, but now $\mathbf{C}^{\sim} = \{\mathbf{c} \mid \mathbf{c} \sim c_i\}$ and $\mathbf{C}^{\nsim} = \{\mathbf{c} \mid \mathbf{c} \nsim c_i\}$.

Proof. Consider a feature parameter $x = \Pr(f_i \mid \pi_{\mathbf{f_i}} \pi_{\mathbf{c_i}})$ with $f_i \sim \mathbf{f}$. From Proposition 1 we have that for all $\mathbf{c} \in \mathbf{C}^{\sim}$, the sensitivity functions are given by $f_{\Pr(\mathbf{c}|\mathbf{f})} = \frac{x \cdot \mathrm{Pr}_0(\mathbf{c}|\mathbf{f})}{(x - x_0) \cdot \mathrm{Pr}_0(\pi_{\mathbf{c_i}}|\mathbf{f}) + x_0}$. These functions only differ in $\mathrm{Pr}_0(\mathbf{c} \mid \mathbf{f})$ and only intersect for $x = 0$. Since $x \in \langle 0, 1 \rangle$, $\mathbf{c}^{\mathbf{max}\sim}$ will remain the same upon varying x. Likewise, $\mathbf{c}^{\mathbf{max}\nsim}$ will not change upon varying x. We moreover observe that for all $\mathbf{c} \in \mathbf{C}^{\sim}$ the sign of the first derivative of the sensitivity functions $f_{\Pr(\mathbf{c}|\mathbf{f})}(x)$ equals the sign of $\mathrm{Pr}_0(\mathbf{c} \mid \mathbf{f}) \cdot \Pr(1 - \Pr(\pi_{\mathbf{c_i}} \mid \mathbf{f})) \cdot x_0$ and these functions thus always increase with increasing x. For for all $\mathbf{c} \in \mathbf{C}^{\nsim}$ the sign of the first derivative of the sensitivity functions $f_{\Pr(\mathbf{c}|\mathbf{f})}(x)$ equals the sign of $-\Pr(\mathbf{c} \mid \mathbf{f}) \cdot \Pr(\pi_{\mathbf{c_i}} \mid \mathbf{f})) \cdot x_0$ and these functions thus always decrease with increasing x. The functions $f_{\Pr(\mathbf{c}^{\sim \mathbf{max}}|\mathbf{f})}(x)$ and $f_{\Pr(\mathbf{c}^{\nsim \mathbf{max}}|\mathbf{f})}(x)$ can therefore intersect at most once. These two results together imply that $\mathbf{c}^{\mathbf{max}}$ can change at most once with changing x and then changes from $\mathbf{c}^{\mathbf{max}\sim}$ to $\mathbf{c}^{\mathbf{max}\nsim}$ or vice versa. For feature parameters with $f_i \nsim \mathbf{f}$ and root class parameters similar arguments apply. $\square$

Proposition 3. Let MDC($\mathbf{C}, \mathbf{F}$), $\mathbf{c}^{\mathbf{max}}$, $\mathbf{c}^{\mathbf{max}\sim}$ and $\mathbf{c}^{\mathbf{max}\nsim}$ be as before. Let $\gamma = \frac{\mathrm{Pr}_0(\mathbf{c}^{\mathbf{max}\nsim}|\mathbf{f})}{\mathrm{Pr}_0(\mathbf{c}^{\mathbf{max}\sim}|\mathbf{f})}$ and let x be a feature parameter with $\pi_{\mathbf{f_i}} \sim \mathbf{f}$ or a root class parameter. We then find the following intervals of admissible deviation for x with respect to the classification $\mathbf{c}^{\mathbf{max}}$:

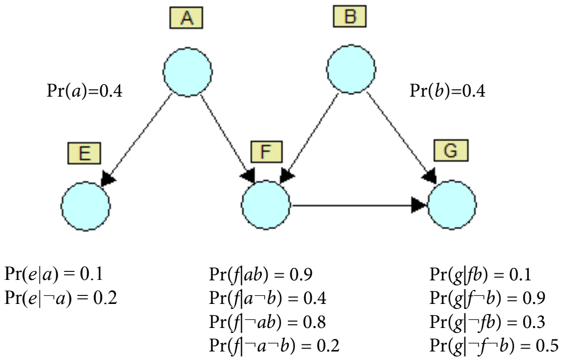

Figure 1: A small example MDC with class variables A, B and feature variables E, F, G.

x	$\mathbf{c^{max}} = \mathbf{c^{max\sim}}$	$\mathbf{c^{max}} = \mathbf{c^{max\nsim}}$
$\Pr(f_i \mid \pi_{\mathbf{f_i}}\pi_{\mathbf{c_i}}),\, f_i \sim \mathbf{f}$	$\langle \gamma \cdot x_0, 1 \rangle$	$\langle 0, \gamma \cdot x_0 \rangle$
$\Pr(f_i \mid \pi_{\mathbf{f_i}}\pi_{\mathbf{c_i}}),\, f_i \nsim \mathbf{f}$	$\langle 0, 1 - \gamma \cdot (1 - x_0) \rangle$	$\langle 1 - \gamma \cdot (1 - x_0), 1 \rangle$
$\Pr(c_i)$	$\left\langle \frac{\gamma \cdot x_0}{1 - x_0 \cdot (1-\gamma)}, 1 \right\rangle$	$\left\langle 0, \frac{\gamma \cdot x_0}{1 - x_0 \cdot (1-\gamma)} \right\rangle$

Proof. From lemma 1 we have that the intervals of admissible deviation are determined by the intersections of $f_{\Pr(\mathbf{c^{max\sim}} \mid \mathbf{f})}(x)$ and $f_{\Pr(\mathbf{c^{max\nsim}} \mid \mathbf{f})}(x)$. For a feature parameter with $f_i \sim \mathbf{f}$, this intersection is found from $x \cdot \Pr_0(\mathbf{c^{max\sim}} \mid \mathbf{f}) = x_0 \cdot \Pr_0(\mathbf{c^{max\nsim}} \mid \mathbf{f})$. In case $\mathbf{c^{max\sim}}$ is originally more likely than $\mathbf{c^{max\nsim}}$, for example, the classification will remain unchanged as long as $x \in \langle \gamma \cdot x_0, 1 \rangle$. All other intervals of admissible deviation are derived analogously. □

Note that the intersection of $f_{\Pr(\mathbf{c^{max\sim}} \mid \mathbf{f})}(x)$ and $f_{\Pr(\mathbf{c^{max\nsim}} \mid \mathbf{f})}(x)$ maybe at inadmissible values for x. In that case $\mathbf{c^{max}}$ cannot be changed by just changing x. Note furthermore that it might be that $\Pr_0(\mathbf{c^{max\sim}} \mid \mathbf{f}) = 0$ which implies that computing γ requires division by zero. Since gamma is just used for notational reasons there is no objection. Note that in case $\Pr_0(\mathbf{c^{max\sim}} \mid \mathbf{f}) = 0$, all sensitivity functions are constant and $\Pr(\mathbf{c^{max}} \mid \mathbf{f}) = \Pr(\mathbf{c^{max\nsim}} \mid \mathbf{f})$.

The formulas for the intervals of admissible deviation are equal for ODCs and MDCs. However, in an ODC there is just one instantiation of $\mathbf{C}$ compatible with x, whereas in an MDC in general there are multiple compatible instantiations. This may affect γ in general. In future research we want to investigate experimentally if, and if so how, this affects the size of the intervals of admissible deviation of ODCs compared to MDCs.

For some applications, for example when decisions based on a outcome have considerable consequences, not only the robustness, but also the reliability of a classification is crucial. We are then interested in how well a network can discriminate between the different classes. A measure for this reliability is the absolute difference in probability between two classifications. The sensitivity of this discrimination ability to parameter changes can be easily computed from the appropriate sensitivity functions [12]. By providing full sensitivity functions, we thus enable detecting changes in discrimination ability for shifts in feature and root class parameters of an MDC.

5 Example

Figure 1 shows a small example MDC with class variables, A and B, and feature variables, E, F and G. We consider an instance efg and the output probabilities $\Pr(AB \mid efg)$. In the current network we find $\Pr(ab \mid efg) = 0.054$, $\Pr(a\bar{b} \mid efg) = 0.321$, $\Pr(\bar{a}b \mid efg) = 0.143$ and $\Pr(\bar{a}\bar{b} \mid efg) = 0.482$. The current classification thus is $\bar{a}\bar{b}$. Moreover we find $\Pr(a \mid efg) = 0.375$ and $\Pr(b \mid efg) = 0.196$. With this information the sensitivity functions given the observations efg can be constructed for all parameters. As an example, Figures 2 and 3, respectively, show the sensitivity functions $f_{\Pr(AB \mid efg)}(x)$

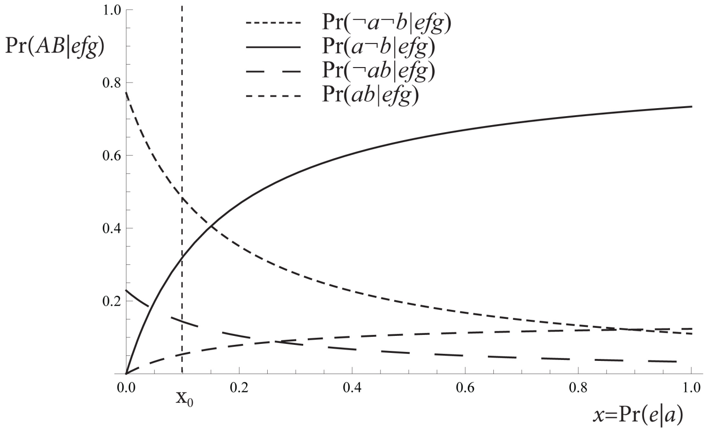

Figure 2: $\Pr(AB \mid efg)$ as a function of $\Pr(e \mid a)$ given the network from Figure 1.

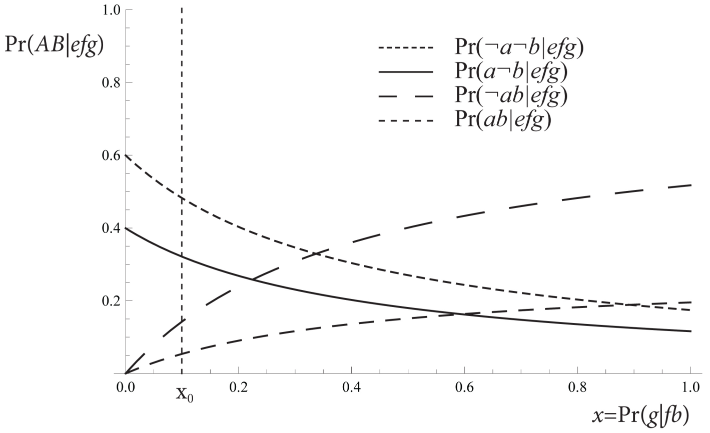

Figure 3: $\Pr(AB \mid efg)$ as a function of $\Pr(g \mid fb)$ given the network from Figure 1.

and $f_{\Pr(AB|efg)}(x)$ for $x = \Pr(e \mid a)$ and $x = \Pr(g \mid fb)$. The sensitivity values for the functions related to the current classification $x = \bar{a}\bar{b}$ are 1.81 and 0.95, respectively.

With respect to the intervals of admissible deviation we observe in Figures 2 and 3 that for the parameter $\Pr(e \mid a)$, the outcome probabilities $\Pr(aB \mid efg)$, which are compatible with a, increase with an increase of the parameter, whereas the outcome probabilities $\Pr(\bar{a}B \mid efg)$ that are not compatible with a decrease. For the parameter $\Pr(g \mid fb)$ on the other hand, the outcome probabilities $\Pr(Ab \mid efg)$ increase and the outcome probabilities $\Pr(A\bar{b} \mid efg)$ decrease with an increase of the parameter. The classification may thus change into $\Pr(aB^{\mathbf{max}} \mid efg) = \Pr(a\bar{b} \mid efg)$ with an increase of $\Pr(e \mid a)$, and into $\Pr(A^{\mathbf{max}}b \mid efg) = \Pr(\bar{a}b \mid efg)$ with an increase of $\Pr(g \mid fb)$. For $\Pr(e \mid a)$ we find that $\gamma = \frac{0.482}{0.321}$ and for $\Pr(g \mid fb)$ that $\gamma = \frac{0.482}{0.143}$ The intervals of admissible deviation are respectively $[0, 0.15]$ and $[0, 0.34]$.

6 Conclusions

Multi-dimensional classifiers were introduced to generalise one-dimensional classifiers to application domains that require an instance to be classified into a combination of classes. In this paper we investigated the sensitivity of MDCs to shifts in their parameters by studying the sensitivity functions we proposed in [2]. We first gave proofs for the validity of our sensitivity functions and then used these functions to derive intervals in which the admissible deviation for a parameter is expressed in just a few meaningful probabilities. We thereby provided an more simple and insightful alternative to [4], in which a Bayesian network is converted to an arithmetic circuit in order to find the robustness conditions.

The sensitivity functions, moreover give insight into the reliability of a classification. We also argued that a classification can change at most once given an increasing or decreasing parameter value. In [2], we concluded that MDCs can be expected to be, in general, less sensitive to small parameter changes than ODCs. From the formulas for the intervals of admissible deviation, however, we could not conclude straightforwardly whether to expect larger intervals of admissible deviation in MDCs or ODCs. In future research, therefore, we want to compare experimentally MDCs and ODCs with respect to the size of the intervals of admissible deviation for their parameters.

Acknowledgement

This research was supported by the Netherlands Organisation for Scientific Research (NWO).

References

[1] C. Bielza, G. Li, and P. Larrañaga. 2011. Multi-Dimensional Classification with Bayesian Networks. In: *International Journal of Approximate Reasoning*, 52, 705-727.

[2] J.H. Bolt and S. Renooij. 2014. Sensitivity of multi-dimensional Bayesian classifiers. In: T. Schaub, G. Friedrich and B. O'Sullivan (eds.): *Proceedings of the 21st European Conference on Artificial Intelligence*, 971-972.

[3] H. Borchani, C. Bielza, C. Toro, and P. Larrañaga. 2013. Predicting human immunodeficiency virus inhibitors using multi-dimensional Bayesian network classifiers. In: *Artificial Intelligence in Medicine*, 57(3), 219-229.

[4] H.C. Chan and A. Darwiche. 2006. On the robustness of Most Probable Explanations. In: *Proceedings of the Twenty Second Conference on Uncertainty in Artificial Intelligence*, 63-71.

[5] P. Domingos and M. Pazzani. 1997. On the optimality of the simple Bayesian classifier under zero-one loss. In: *Machine Learning*, 29, 103-130.

[6] L.C. van der Gaag and S. Renooij. 2001. Analysing sensitivity data. In: J. Breese and D. Koller (Eds.), *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, 530-537.

[7] L.C. van der Gaag and P.R. de Waal. 2006. Multi-dimensional Bayesian Network Classifiers. In: M Studeny and J Vomlel (Eds.), *Proceedings of the Third European Workshop in Probabilistic Graphical Models*, 107-114.

[8] M. Friedman, D. Geiger and M. Goldschmidt. 1997. Bayesian network classifiers. In: Machine Learning, 29, 131-163.

[9] K.B. Laskey. 1995. Sensitivity analysis for probability assessments in Bayesian networks. In: *IEEE Transactions on Systems, Man and Cybernetics*, 25: 901-909.

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

[11] S. Renooij and L.C. van der Gaag. 2008. Evidence and scenario sensitivities in naive Bayesian classifiers. In: *International Journal of Approximate Reasoning*, 49(2): 398-416.

[12] S. Renooij and L.C. van der Gaag. 2008. Discrimination and its sensitivity in probabilistic networks. In: M. Jaeger and T.D. Nielsen (Eds.), *Proceedings of the Fourth Workshop on Probabilistic Graphical Models*, 241-248.

[13] P.R. de Waal and L.C. van der Gaag. 2007. Inference and learning in multi-dimensional Bayesian network classifiers. In: K. Mellouli (ed). *European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, 501-511.