Aligning Bayesian Network Classifiers with Medical Contexts

Linda C. van der Gaag

Silja Renooij

Ad Feelders

Arend de Groote

Marinus J.C. Eijkemans

Frank J. Broekmans

Bart C.J.M. Fauser

Technical Report UU-CS-2008-015
June 2008

Department of Information and Computing Sciences
Utrecht University, Utrecht, The Netherlands
www.cs.uu.nl

ISSN: 0924-3275

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

Aligning Bayesian Network Classifiers with Medical Contexts

Linda C. van der Gaag [1], Silja Renooij [1], Ad Feelders [1], Arend de Groote [2], Marinus J.C. Eijkemans [2,3], Frank J. Broekmans [2], and Bart C.J.M. Fauser [2]
[1] Department of Information and Computing Sciences, Utrecht University, P.O. Box 80.089, 3508 TB Utrecht, The Netherlands
[2] Department of Reproductive Medicine and Gynaecology, Utrecht Medical Centre, Heidelberglaan 100, 3584 CS Utrecht, The Netherlands
[3] Department of Public Health, Erasmus University Medical Center, P.O. Box 2040, 3000 CA Rotterdam, The Netherlands

Abstract

While for many problems in medicine classification models are being developed, Bayesian network classifiers do not seem to have become as widely accepted within the medical community as logistic regression models. We compare first-order logistic regression and naive Bayesian classification in the domain of reproductive medicine and demonstrate that the two techniques can result in models of comparable performance. For Bayesian network classifiers to become more widely accepted within the medical community then, we feel that they should be better aligned with their context of application. We describe how to incorporate well-known concepts of clinical relevance in the process of constructing and evaluating Bayesian network classifiers to achieve such an alignment.

Keywords and Phrases: Bayesian network classifiers, Naive Bayesian classifier, Learning Bayesian classifiers, Medical alignment, Logistic regression, Accuracy, Area under the ROC curve.

1 Introduction

Bayesian network classifiers are stochastic models that describe the relationship between one or more feature variables and a class variable, and provide for establishing posterior probabilities of the various classes for a given instance of the feature variables. Numerous applications of Bayesian network classifiers exist. Yet, within the medical field where most diagnostic problems can be considered classification problems, Bayesian classifiers are hardly ever used. Stated informally, in a diagnostic medical problem, patients have to be assigned to one of a usually small number of distinct diagnostic classes based upon the patient's characteristics. A similar observation also holds for many problems that are prognostic in nature. In the domain of reproductive medicine, for example, patients

are classified as elective or non-elective for single embryo transfer upon in vitro fertilisation. To support physicians in taking classification decisions about individual patients, the most commonly employed models in the medical community, are based on the technique of *logistic regression*. Logistic regression serves to construct, from a set of available patient data, a model that describes the relationship between the various feature variables involved and a class variable. The model equally provides for establishing the posterior probabilities of the classes, based upon which a decision is recommended.

Bayesian network classifiers have a number of advantages over logistic regression models, which should render them attractive alternatives for the medical field. A major advantage of Bayesian network classifiers lies in their ability to give reliable classifications even if evidence is available for only a subset of the feature variables. Bayesian network classifiers moreover provide a graphical representation of the independences between the modelled variables, which allows for transparency and ease of interpretation of the models and their parameters. Bayesian network classifiers further range from the simplest type of model, the *naive Bayesian classifier* which makes strong independence assumptions concerning the feature variables involved, through the slightly more sophisticated *TAN classifier* allowing restricted dependences between the feature variables, to full Bayesian networks modelling the intricate dependence structure that actually holds in the application domain. Classifiers of varying complexity can thus be modelled within a single framework.

Even though it is known in theory that first-order logistic regression models perform at least as good as naive Bayesian classifiers for larger data sets, many researchers have reported comparable or even better performance of the Bayesian network classifier for smaller data sets [Ng and Jordan (2002), Twardy et al. (2006)]. In this paper we describe our first steps aimed at the adoption of a Bayesian network classifier in the domain of reproductive medicine. At our disposal we had a small data set from patients undergoing single embryo transfer upon in vitro fertilisation. From this data set, a first-order logistic regression model had been constructed for the problem of predicting ongoing pregnancies [Verberg et al. (2007)]. From the small data set, we equally constructed a naive Bayesian classifier and studied its performance compared to that of the logistic regression model.

Logistic regression models, developed to support physicians in making patient-specific classification decisions, are typically evaluated using well-known concepts of clinical relevance such as the *area under the ROC curve*, or AUC, and *sensitivity* and *specificity*. The AUC gives an indication of quality, averaged over all possible decision thresholds for assigning an instance to a particular class. For use in practice, a fixed decision threshold is chosen based upon knowledge of the consequences of misclassification. Given this threshold, the model has an associated sensitivity and specificity, where the sensitivity is the percentage of true positives predicted by the model and the specificity is the percentage of true negatives.

Bayesian network classifiers generally are not evaluated using the concepts of clinical relevance mentioned above, but using *classification accuracy* as an indication of quality instead. Classification accuracy refers to the percentage of instances that are correctly classified by the model. The importance of communicating clinical relevance of constructed models, however, should not be underestimated: such concepts help convey to the physician a detailed assessment of the quality and relevance of patient-specific decisions based upon the model. We feel that the smaller acceptance of Bayesian network classifiers in the medical community can be attributed to at least some extent to their un-

familiar underlying concepts. For Bayesian network classifiers to become more widely accepted, we feel more specifically that they should be better aligned with the medical contexts in which they are to be used. In this paper, we describe how failure to use concepts of clinical relevance results in medically unacceptable Bayesian network classifiers. On the contrary, we show that incorporating such concepts both in the process of constructing and of evaluating Bayesian network classifiers, helps us to achieve such an alignment. In fact, we can report comparable results for the logistic regression model and a naive Bayesian network classifier only after alignment.

The paper is organised as follows. In Section 2, we review Bayesian network classifiers and compare them, theoretically, to logistic regression models. In Section 3 we describe our domain of application and the data that we had available for our alignment study; in addition, we describe the different concepts relevant for alignment. In Section 4 we elaborate on the incorporation of the concepts of clinical relevance in the process of constructing and evaluating Bayesian network classifiers. The results for the naive Bayesian classifiers constructed from our data set are presented in Section 5. We end with our concluding observations in Section 6.

2 Bayesian network classifiers and logistic regression

Quite a number of stochastic classification paradigms exist; we refer to Larrañaga et al. (2006) for an overview. In this paper, however, we focus on Bayesian network classifiers and logistic regression models. We begin by reviewing different types of Bayesian network classifier and compare them to logistic regression models.

2.1 Bayesian network classifiers

Stochastic classifiers in general provide for addressing problems in which an instance of a set of feature variables has to be assigned to a value of the class variable. These classifiers in essence establish the conditional probability distribution over the class variable given the instance, from which they decide upon the class for that instance using a decision rule.

Bayesian network classifiers build upon a Bayesian network for establishing the probability distribution over their class variable. Such a network is a concise representation of a joint probability distribution over the set of variables involved. For the purpose of classification, this set is divided into a set of feature variables, the class variable, and possibly some intermediate, or hidden, variables. Bayesian network classifiers vary in complexity from general models posing no restrictions on the dependences between the variables, to very simple models with highly constrained dependency structures. Two well-known simple Bayesian network classifiers are the *naive Bayesian classifier* and the *TAN classifier* [Friedman, Geiger and Goldszmidt (1997)]. These models both assume an empty set of hidden variables. The naive Bayesian classifier in addition assumes mutual independence of the feature variables given the class variable; the TAN classifier, or tree augmented network classifier, allows a tree-like dependency structure over its feature variables. Because of their simplicity, naive Bayesian classifiers are being developed for a wide range of application domains and, despite their simplicity, often very good performance is reported [Friedman, Geiger and Goldszmidt (1997), Domingos and Pazzani (1997)].

Throughout this paper, we assume that the class variable Y is a binary variable, with a

positive class value denoted by y and a negative class value denoted by $\overline{y}$; we use y' to refer to either class. The set of feature variables is denoted by $\mathbf{X}$; $\mathbf{x}$ is used to denote a specific instance of this set. The naive Bayesian network classifier now explicitly models the joint probability distribution $\Pr(\mathbf{X}, Y)$ over its variables in terms of parameters $p(X_i \mid Y)$ specified for its feature variables $X_i \in \mathbf{X}$, and $p(Y)$ specified for the class variable Y. Its independence assumptions result in the following parametrisation:

$$\Pr(\mathbf{X}, Y) = p(Y) \cdot \prod_i p(X_i \mid Y)$$

Bayesian network classifiers in general are often constructed automatically from a data set. Algorithms for this purpose include a measure to decide upon the dependences between the variables to be included in order to optimise the model's quality. Examples of such measures are a model's accuracy and its minimum description length (MDL). The quality of a model in view of the data can only be established if the model is fully specified, that is, if it includes estimates for all numerical parameters involved. These parameters are estimated as simple frequency counts, which serve to maximise the log-likelihood of the model given the data. The quality measure that is used as an optimisation criterion upon constructing the model often is also exploited for comparing different classifiers.

Upon learning Bayesian network classifiers, the quality of a model is not optimised just by including appropriate dependences, but also by including only the most relevant feature variables. Data sets often contain more variables than are strictly necessary for the classification task at hand and the more or less redundant variables could result in an undesirable bias [Langley and Sage (1994)]. The process of *feature selection* now carefully selects from the data set the variables that serve to improve the model's quality the most. For this purpose, various feature-selection methods exist; we again refer to Larrañaga et al. (2006) for an overview. Here we focus on the so-called wrapper approach to feature selection and assume that a greedy forward-selection method is used for choosing the feature variables to be included. In this approach, feature variables are iteratively added to an initially empty model until its quality given the data no longer increases.

Bayesian network classifiers use Bayes' rule for establishing the posterior probability distribution $\Pr(Y \mid \mathbf{X})$ over the class variable that is used for the actual classification:

$$\Pr(Y \mid \mathbf{X}) = \frac{\Pr(Y, \mathbf{X})}{\Pr(\mathbf{X})} = \frac{\Pr(\mathbf{X} \mid Y) \cdot \Pr(Y)}{\sum_{y'} \Pr(\mathbf{X} \mid y') \cdot \Pr(y')}$$

The decision rule that is commonly used with a naive Bayesian classifier is the *winner-takes-all rule*, which for a binary class variable amounts to assigning an instance to the class whose posterior probability exceeds the threshold probability of 0.5.

To conclude, if the performance of the constructed model is evaluated against the same data set as that from which the model is learned, its performance will tend to be overestimated as a result of overfitting the model to the data. To correct for this effect of overfitting and estimate the model's performance on unseen data, often ten-fold cross validation is used.

2.2 Logistic regression

Logistic regression models are much more commonly used within the medical community than Bayesian network classifiers, even though there are quite a number of similarities be-

tween these types of model. A logistic regression model, like a Bayesian network classifier, is a model over a class variable Y and a set of feature variables $\mathbf{X}$. The model captures the conditional probability distribution over the class variable directly as a function of the feature variables $X_i \in \mathbf{X}$. Logistic regression models again range from simple models imposing a linear function on the feature variables, to more complex models involving higher-order terms to describe interactions between the feature variables. The first-order logistic regression model captures the conditional probability distribution $\Pr(Y \mid \mathbf{X})$ in terms of a linear function of the feature variables $X_i \in \mathbf{X}$ through

$$\Pr(y \mid \mathbf{X}) = (1 + \exp(-\beta_0 - \textstyle\sum_i \beta_i \cdot X_i))^{-1}$$

in which β_i denote the model's parameters.

Logistic regression models are also constructed automatically from data. The log-likelihood of the model given the available data then is maximised by obtaining appropriate estimates for the parameters β_i. While for the parameters of a Bayesian network classifier, a closed-formula solution exists, the optimisation problem involved in finding the parameters for a logistic regression model does not have such a solution. The parameter values therefore are established using an iterative method. As for Bayesian network classifiers, furthermore, upon constructing a logistic regression model methods for feature selection and for correcting for the effect of overfitting are applied.

A logistic regression model provides for directly computing the posterior probabilities $\Pr(Y \mid \mathbf{x})$ for the class variable given an instance $\mathbf{x}$, by filling in the values for the feature variables. The decision rule used with the model is based upon a threshold probability t for this posterior probability. The value of this *decision threshold* t is based upon knowledge of the consequences of the different types of misclassification in the domain of application.

2.3 A theoretical comparison

Naive Bayesian classifiers and first-order logistic regression models essentially index the same set of conditional probability distributions, in the sense that for any combination of parameter values of a first-order logistic regression model there exists a combination of parameter values for a naive Bayesian classifier that describes the same distribution $\Pr(Y \mid \mathbf{X})$, and vice versa (provided that $\Pr(Y, \mathbf{X})$ is strictly positive) [McLachlan (1992)]. Yet, given a particular data set, naive Bayesian classification and logistic regression will typically not result in the same estimated distribution $\widehat{\Pr}(Y \mid \mathbf{X})$, because the parameter values for the Bayesian network classifier are chosen so as to maximise the log-likelihood of the *joint* probability distribution over the variables whereas the parameter values for the logistic regression model are chosen so as to optimise the log-likelihood of the *conditional* distribution. If in learning a naive Bayesian classifier, its parameter values are computed iteratively so as to maximise the log-likelihood of the conditional distribution, called discriminative learning, the resulting model would in essence capture the same distribution as a first-order logistic regression model learned from the data.

Several researchers have argued that discriminative learning is more appropriate for classification purposes than generative learning in which the log-likelihood of the joint distribution is maximised, since we are interested in predicting the class for a given instance and not in their joint probability [Friedman, Geiger and Goldszmidt (1997)]. Others, however, argue that such a conclusion may be premature [Ng and Jordan (2002)].

For models fitted to infinite data sets, the asymptotic classification accuracy of a first-order logistic regression model is never smaller than that of a naive Bayesian classifier. This basically implies that given a large enough data set the naive Bayesian classifier will not outperform the first-order logistic regression model. The regression model would typically do better when the independence assumption underlying the naive Bayesian classifier does not hold in the data set, that is, when there are strong associations among the feature variables [Anderson (1982)]. On the other hand, although naive Bayesian classification asymptotically converges to a lower accuracy, it does so significantly faster than logistic regression. For smaller data sets, therefore, naive Bayesian classifiers can be expected to outperform first-order logistic regression models, as has been largely confirmed experimentally [Ng and Jordan (2002)].

Similar observations in essence hold for TAN classifiers on the one hand and logistic regression models with interaction terms for pairs of feature variables on the other hand. The observations, however, cannot be extended to Bayesian network classifiers involving more complex dependency structures over their feature variables. Such Bayesian network classifiers index essentially different sets of conditional probability distributions than logistic regression models with higher-order interaction terms and as a consequence may theoretically as well as effectively outperform any such regression model.

3 The medical context

In this section we describe the medical concepts relevant for our case study of aligning Bayesian network classifiers with a medical context. We briefly introduce the domain of reproductive medicine in which we conducted the case study, together we the data set and logistic regression model that we had available.

3.1 Clinical performance

Classification models in medicine are often evaluated using concepts of clinical relevance such as area under the ROC curve, and sensitivity and specificity. Given the importance of these concepts for aligning Bayesian network classifiers with medical contexts, we briefly review them here.

A *Receiver Operator Characteristic*, or ROC, curve visualises a classifier's performance by plotting its sensitivity against one minus its specificity for all possible values of the decision threshold t. Some example ROC curves are shown in Figure 1. The decision threshold serves for classifying an instance $\mathbf{x}$ as belonging to class y only if the posterior probability $\Pr(y \mid \mathbf{x})$ computed for the instance is at or above the threshold. Given this threshold, the model has an associated sensitivity and specificity. The *sensitivity* of the model is the probability that it correctly classifies a positive instance $\mathbf{x}^+$, that is, it is the percentage of such instances for which the classifier predicts that $\Pr(y \mid \mathbf{x}^+) \geq t$. The *specificity* of a classifier is the probability that it correctly classifies a negative instance $\mathbf{x}^-$, that is, it is the percentage of such instances for which $\Pr(y \mid \mathbf{x}^-) < t$ is predicted.

The *area under the ROC curve*, or AUC, in essence measures the classifier's ability to discriminate between the different classes [Hanley and McNeil (1982)]. More specifically, it captures the probability that a randomly chosen positive instance and a random negative instance are correctly ranked, that is, it is the percentage of such pairs of in-

stances for which the classifier predicts that $\Pr(y \mid \mathbf{x}^+) > \Pr(y \mid \mathbf{x}^-)$. The area under the curve gives an indication of a classifier's quality, averaged over all possible decision thresholds for assigning an instance to a particular class. Upon using the model in a practical setting, however, a clinician typically has to base a decision upon the posterior probability computed for a patient. For this purpose, a fixed decision threshold is chosen, based upon domain knowledge of the consequences of misclassification. In view of such a fixed decision threshold, the quality of a classification model is captured by a single point on the ROC curve. The area under the curve may then no longer be an appropriate indication of the model's performance, which is then expressed directly by the sensitivity and specificity implied by the fixed decision threshold.

3.2 In vitro fertilisation

In vitro fertilisation, or IVF, is an assisted reproductive technique of embryo transfer used to help infertile couples conceive a child. There are many factors that determine whether or not IVF treatment results in an ongoing pregnancy, including the age of the patient, the quality of the embryo, and the receptivity of the uterus. To increase the probability of pregnancy, it used to be common practice for IVF programmes to transfer multiple embryos. With the increasing success of the treatment, however, multiple embryo transfer involves an increased risk of multiple pregnancy, associated with pregnancy loss, obstetrical complications, prematurity, and neonatal morbidity with long term damage.

Single embryo transfer is now being used as a means of reducing the risks involved with multiple pregnancy. Applying single embryo transfer without any selection based on patient characteristics and embryo quality, however, has been shown to lead to a reduction of the probability of an ongoing pregnancy per transfer. A patient may therefore need to undergo multiple treatments for a pregnancy to persist. Multiple treatments involve additional costs as well as physical and emotional discomfort for the patient. To guide appropriate use of single embryo transfer, therefore, a patient-specific assessment of the expected result of the transfer should be available.

In a recent study, a prognostic model was developed for establishing the probability of an ongoing pregnancy after single embryo transfer [Verberg et al. (2007)]. The data used for constructing this logistic regression model were derived from a randomised controlled trial on the effectiveness of in vitro fertilisation, in which 201 women with an indication for IVF treatment were randomised to a mild stimulation protocol [Heijnen et al. (2007)]. For constructing the model, a subset of these data including only women with at least two embryos suitable for transfer was used. The subset includes the data of 152 women who underwent single embryo transfer. In 42 of these women (28%), the treatment resulted in an ongoing pregnancy.

In the data set, patient characteristics, treatment details, and embryo quality related factors are recorded. The feature variables include such patient characteristics as female age, previous pregnancy, cause and duration of infertility, and body mass index. Further independent variables are related to the treatment and include the number of dominant follicles, the number of oocytes retrieved, the proportion of fertilised oocytes, the duration of the stimulation, the amount of administered recFSH per retrieved oocyte, and endometrial thickness. The remaining feature variables are related to embryo quality and include the grade of fragmentation, whether there was a top-quality embryo available for transfer, and whether there were embryos available for cryopreservation. The number of

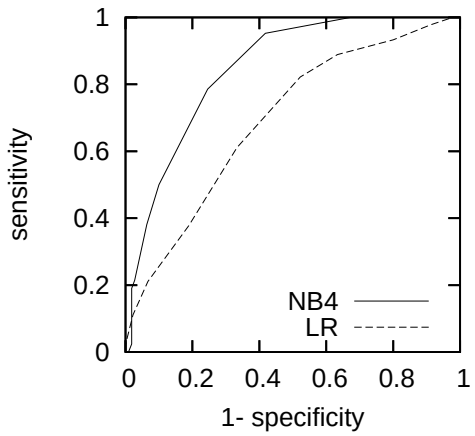

Figure 1: ROC curves for the logistic regression model (dashed) and the naive Bayesian network classifier (solid) with four variables each (uncorrected).

independent variables equals 17, of which 11 variables are continuous, 3 are binary, and 3 are multi-categorical. For two of the variables, data were not complete, with 4% and 6% of the values missing, respectively. For these variables, single imputation was used by filling in the predictive mean after regression on all other variables. The variable designated as the class variable in the data set captures whether or not single embryo transfer results in an ongoing pregnancy.

We briefly review the performance characteristics of the first-order logistic regression model constructed from the data. The model includes four feature variables, which was imposed as the maximum number of variables to be included. The variables of the model are the patient's *body-mass index*, the *total amount of administered follicle stimulating hormone*, the *number of retrieved oocytes*, and *whether there was a top-quality embryo available for transfer*. An ROC curve for the model is shown in Figure 1; its area under the curve equals 0.68, or 0.60 after correcting for the effects of overfitting. Using a decision threshold of 0.2, the model has a sensitivity of 0.90, or 0.86 after correction, and a specificity of 0.37, or 0.14 after correction. These characteristics are summarised in Tables 1 and 2.

4 Aligning Bayesian classifiers

We recall that building stochastic classification models typically involves selecting a subset of appropriate feature variables and that this subset is often construed using a greedy forward-selection approach, in which feature variables are iteratively added to an initially empty model, until its performance no longer increases. Upon constructing a Bayesian network classifier, performance is often measured by *classification accuracy*, which refers to the percentage of correctly classified instances. In this section, we discuss why classification accuracy is an unacceptable measure of performance in our domain of application and show how concepts of clinical relevance can instead be used for this purpose.

Table 1: Performance characteristics of the first-order logistic regression model and of the naive Bayesian network classifier, with four selected variables each.

	AUC (corrected)	*sensitivity* (corrected)	*specificity* (corrected)
logistic regression model:	0.68 (0.60)	0.90 (0.86)	0.37 (0.14)
naive Bayesian classifier:	0.85 (0.58)	0.95 (0.66)	0.55 (0.50)

Table 2: The variables included in the first-order logistic regression model and in the naive Bayesian network classifier with four selected variables each.

logistic regression model	*naive Bayesian classifier*
–	*duration of infertility* (discretised)
number of retrieved oocytes	*number of retrieved oocytes* (discretised)
top-quality embryo available	*top-quality embryo available*
–	*endometrial thickness* (discretised)
administered follicle stimulating hormone	–
body-mass index	–

4.1 Classification accuracy and its problems

Classification accuracy refers to the probability of *correctly* classifying an arbitrary instance $\mathbf{x}$ of the feature variables involved, where $\mathbf{x}$ is considered correctly classified if the class y' to which $\mathbf{x}$ is assigned, corresponds to its true class y^*. We recall that the assignment of a class value depends upon the threshold probability t that is used with the classifier's decision rule: an instance $\mathbf{x}$ is assigned to a class y' whenever $\Pr(y' \mid \mathbf{x}) \geq t$. Alternatively, a classifier's accuracy can be interpreted as the percentage of randomly chosen pairs of a positive instance $\mathbf{x}^+$ and negative instance $\mathbf{x}^-$, for which $\Pr(y \mid \mathbf{x}^+) \geq t$ and $\Pr(\bar{y} \mid \mathbf{x}^-) \geq t$. The winner-takes-all decision rule commonly employed by Bayesian network classifiers implies a threshold probability of $t = 0.5$.

While the metric of classification accuracy is widely used within the Bayesian network community, it is hardly ever used with logistic regression models in medical contexts. To elaborate on why classification accuracy often is inappropriate for measuring performance in medicine, we begin by observing that the accuracy of a classifier is highly dependent upon the threshold probability that is used with the classifier's decision rule. If instances $\mathbf{x}$ are assigned to a class y' whenever $\Pr(y' \mid \mathbf{x}) \geq t$, then changing the value of the threshold probability t will change the number of instances assigned to class y'. As a result, the sensitivity and the specificity of the classifier also change.

Now, by writing classification accuracy as

$$\mathit{accuracy}(t) = \mathit{sensitivity}(t) \cdot p(y) + \mathit{specificity}(t) \cdot p(\bar{y})$$

two problems of using this metric as a performance measure become apparent [Bradley

(1997)]. The first problem is that its dependence on the choice of decision threshold makes classification accuracy an inappropriate measure of performance in general. In fact, the threshold probability $t = 0.5$ that is implied by the winner-takes-all rule is defendable only if the prior distribution of the class variable is close to uniform. In our domain of in vitro fertilisation for example, this property does not hold: single embryo transfer results in an ongoing pregnancy in only 28% of the patients. Moreover, patients with a small probability of an ongoing pregnancy will receive multiple embryos and, as a consequence, will be exposed to the risks of multiple pregnancy. For this reason, a decision threshold of 0.2 for predicting ongoing pregnancy was chosen for the logistic regression model, and in fact a decision threshold of 0.5 may not result in medically acceptable behaviour. Since the decision threshold generally is not a parameter of the learning process, it is questionable whether feature selection based upon classification accuracy as a performance measure would result in an acceptable model. Note that this particular problem of the metric of classification accuracy is only technical and could be resolved by fitting the choice of decision threshold to the prior class distribution. A procedure to this end is described in Lachiche and Flach (2003).

The second problem with the metric of classification accuracy for measuring performance is that it assigns fixed importance weights to the sensitivity and the specificity of a model, dictated by the prior probability distribution $\Pr(Y)$ over the class variable. Thereby, it assigns a fixed weight to the costs of the two types of misclassification. A uniform class distribution, for example, entails that the cost of misclassification is independent of the predicted class. For many medical contexts, however, the consequences of false positive errors may be very different from those of false negative errors. Moreover, for non-uniform priors, either the sensitivity or the specificity is automatically weighed more heavily, independent of any medical considerations. In our domain of in vitro fertilisation, for example, the prior distribution over the class variable would assign a higher weight to the model's specificity, that is, to correctly predicting non-implantation, than to the sensitivity, that is, to predicting ongoing pregnancies upon transferring a single embryo. Experts in reproductive medicine, however, indicate that the consequences of acting upon a false negative prediction are more severe than for false positive predictions. A high sensitivity therefore is considered more important than a high specificity. This second problem may very well be the reason why classification accuracy is not used as a measure of performance for logistic regression models in medicine.

The inappropriateness of classification accuracy as a performance measure has been recognised in other domains as well. In the machine learning community, the area under the ROC curve has been used for some time now as a measure both for comparing classifiers and for constructing them. Since the area under the curve is not dependent on the decision threshold chosen and is invariant to the prior distribution over the class variable, it is more generally applicable as a performance measure for classification models than the classification accuracy. In fact, it has been shown that Bayesian network classifiers constructed to maximise AUC, provide better ranking and probability estimates for the instances to be classified, and in addition even score better on classification accuracy than those optimised for that purpose [Ling, Huang and Zhang (2003)].

4.2 Clinical alignment

Bayesian network classifiers will only become an accepted alternative to logistic regression models in medicine, if their quality is at least comparable and communicated in terms of clinical relevance. Although this observation does not necessarily affect the construction of Bayesian network classifiers, it has been recognised that if classifiers are evaluated using some quality measure, then it makes sense to optimise that very measure during construction. For this reason, on top of the problems with classification accuracy mentioned above, we incorporate measures of clinical relevance in the learning process of our Bayesian network classifiers.

Upon learning Bayesian network classifiers from data, the area under the curve can be readily incorporated in a greedy forward-selection approach to feature selection. We recall that in this approach feature variables are iteratively added to an initially empty model until its performance no longer increases. We further recall that with this approach, in each iteration, for each (remaining) feature variable, the increase it incurs in the classifier's accuracy is computed, using the decision threshold 0.5 of the winner-takes-all rule. In using the area under the curve as optimisation criterion, we now compute instead for each feature variable the increase it incurs in the classifier's AUC. To this end, for each feature variable, the sensitivity $\mathit{sensitivity}(t)$ and the specificity $\mathit{specificity}(t)$ of the classifier, for n different values of the decision threshold t between zero and one is determined. From the n points thus obtained of the classifier's ROC curve, the area under the curve can be approximated by constructing trapezoids under the curve between every two consecutive points. It can be readily shown that this approximation equals

$$\frac{1}{2} \cdot \sum_{i=1,\ldots,n-1} \Big(\mathit{sensitivity}(t_i) + \mathit{sensitivity}(t_{i+1}) \Big) \cdot \Big(\mathit{specificity}(t_{i+1}) - \mathit{specificity}(t_i) \Big)$$

where t_i is the decision threshold that resulted in the ith sensitivity-specificity pair. We then select for inclusion in the classifier, the feature that results in the largest increase in AUC, if any. We would like to note that for establishing the n points of the ROC curve, we have to compute the posterior distribution over the class variable only once.

5 Experimental results

Our first step into building a Bayesian network classifier for the domain of reproductive medicine has been to learn a collection of naive Bayesian classifiers from the available data. Based upon the theoretical results reviewed in Section 2, we could expect similar performance of the naive Bayesian classifier and the logistic regression model constructed from the data. In fact, since our data set is relatively small, we could even expect slightly better performance of the Bayesian network classifier. We used our *Dazzle* toolbox [Schrage, Van IJzendoorn and Van der Gaag (2005)], for constructing various naive Bayesian network classifiers. Before doing so, however, we had to discretise the continuous variables from the data set. For this purpose, knowledge was elicited from the domain experts who had been involved in the collection of the data. We would like to note that the resulting discretisation might not be the best situated for our classifiers. For the purpose of feature selection, we employed the greedy forward-selection approach outlined above, using the area under the curve for our optimisation criterion. In this section, we

review the results that we obtained. For each constructed model, we report the area under the curve, as well as the sensitivity and specificity characteristics that result from using a decision threshold of 0.2 on the entire data set; we further report corrected performance characteristics obtained using ten-fold cross validation.

To allow for comparing the performance of the first-order logistic regression model constructed from the available data and that of our naive Bayesian classifiers in detail, we decided to construct network classifiers with different numbers of feature variables. With a maximum of four variables, as was imposed on the regression model, the constructed naive Bayesian classifier includes the feature variables modelling the *duration of the infertility*, the *number of retrieved oocytes*, *endometrial thickness* and *whether there was a top-quality embryo available for transfer*. The area under the curve of this classifier equals 0.85, or 0.58 after correcting for the effects of overfitting. Using a decision threshold of 0.2, the classifier has a sensitivity of 0.95, or 0.66 after correction, and a specificity of 0.58, or 0.50 after correction. These characteristics are summarised in Tables 1 and 2. By comparing the characteristics after correction of the naive Bayesian classifier with those of the first-order logistic regression model, we find that the differences between their area under curve and their sensitivities are not significant; the specificity of the naive Bayesian classifier, however, is significantly larger than that of the regression model (using a Student t distribution with a significance level of $\alpha = 0.05$).

In addition to the naive Bayesian classifier with four variables, we also constructed classifiers with fewer and with more variables. The results from all constructed network classifiers are summarised in Table 3. With the restriction of a single feature variable, the constructed classifier includes just the *duration of the infertility*: the addition of this variable is found to increase the area under the curve of the initially empty classifier the most. When allowed a second feature variable, the learning algorithm includes the *number of retrieved oocytes* in addition to the *duration of the infertility* in the classifier. The feature variables modelling *endometrial thickness* and *whether there was a top-quality embryo available for transfer* are included as the third and fourth variable respectively. The fifth feature variable included in the model is the *total amount of administered follicle stimulating hormone*. If the inclusion of feature variables is continued until the classifier's area under the curve no longer increases, a total of eight variables is included. In addition to the five variables mentioned above, also the feature variables modelling the *grade of fragmentation of the embryo*, the *number of normally fertilised oocytes*, and the patient's *age* are included. The remaining variables are not included into the classifier since they in fact serve to decrease the classifier's area under the curve.

We note that upon constructing a naive Bayesian classifier, the contribution of each feature variable to the area under the curve is studied in view of the entire data set. The uncorrected AUC values reported in Table 3 therefore are the values used upon constructing the model. While the classifier's area under the curve keeps increasing upon including a fourth and even further feature variables when the full data set is considered, the values that have been corrected for the effects of overfitting, also reported in Table 3, reveal a decrease in the expected area under the curve on unseen data. These observations support the conclusion that for our small data set selecting four or more feature variables would result in a naive Bayesian classifier that is overfitted to the data.

When comparing the performance characteristics of the various constructed naive Bayesian classifiers, especially the corrected values for the area under the curve and the sensitivity suggest that the best classifier is the one that includes three feature variables.

Upon comparing the characteristics of this model with those of the first-order logistic regression model constructed from the data, we find that the differences between their area under curve and their sensitivity are not significant; the specificity of the naive Bayesian classifier, however, again is significantly larger than that of the regression model (using a Student t distribution with a significance level of $\alpha = 0.10$).

Table 3: Characteristics of the naive Bayesian network classifiers with different numbers of variables.

# variables	*AUC* (corrected)	*sensitivity* (corrected)	*specificity* (corrected)
0	0.50 (0.50)	1.00 (1.00)	0 (0)
1	0.69 (0.53)	0.93 (0.80)	0.31 (0.22)
2	0.76 (0.63)	0.93 (0.78)	0.45 (0.35)
3	0.80 (0.65)	0.90 (0.80)	0.55 (0.46)
4	0.85 (0.58)	0.95 (0.66)	0.58 (0.50)
5	0.86 (0.56)	0.93 (0.65)	0.63 (0.51)
8	0.89 (0.56)	0.95 (0.57)	0.67 (0.52)

To conclude, we would like to illustrate the inappropriateness of using classification accuracy for measuring performance for our domain of application. We constructed an additional naive Bayesian network classifier from our data set with a maximum of four feature variables; for this classifier we used accuracy for the optimisation criterion. The corrected area under the curve of the classifier is 0.54. With the winner-takes-all rule, the corrected sensitivity is 0.13; the corrected specificity equals 0.84. This classifier would not exhibit medically acceptable performance, as a consequence of its low sensitivity.

6 Concluding observations

While for many problems in medicine classification models are being developed, Bayesian network classifiers do not seem to have become as widely accepted within the medical community as logistic regression models. To promote Bayesian network classifiers as alternatives to logistic regression, it is important that comparison between the two can be done in terms familiar to the medical community. In the medical domain, concepts of clinical relevance are used, such as the area under the curve and sensitivity and specificity. We have argued that for Bayesian network classifiers to become more widely accepted within the medical community, they should be better aligned with their medical contexts by using these concepts of clinical relevance. In addition, we will have to demonstrate that Bayesian network classifiers are acceptable or better alternatives in terms of performance, ease of construction and ease of interpretation.

Given an infinite data set and optimising accuracy, a naive Bayesian classifier cannot outperform a logistic regression model. Comparing a previously constructed regression model with a naive Bayesian network classifier for the problem of selecting patients for single embryo transfer in reproductive medicine, we found that even for a small data set the naive Bayesian classifier can be outperformed by far by the logistic regression model,

that is, if the former is constructed using classification accuracy as a performance measure. We have argued, however, that the metric of classification accuracy may not be appropriate for measuring performance of classification models in the medical domain. Serious problems are associated with using the metric for non-uniform distributions over the class variable and for unequal cost distributions over the different types of misclassification, which may give rise to classification models of unacceptable medical behaviour. We have shown that concepts of clinical relevance can be readily taken into account upon constructing naive Bayesian classifiers from data. For our relatively small data set in reproductive medicine, we have shown that by doing so, naive Bayesian classifiers can result that exhibit at least comparable behaviour to logistic regression models. The promising results from aligning the simplest type of Bayesian network classifier to its medical context, have made our medical experts enthusiastic.

References

[Anderson (1982)] Anderson, J.A. (1982). Logistic discrimination. In: P.R. Krishnaiah, L.N. Kanal (editors). *Classification, Pattern Recognition and Reduction of Dimensionality*, Handbook of Statistics, vol. 2, North-Holland, pp. 169–191.

[Bradley (1997)] Bradley, A.P. (1997). The use of area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30, pp. 1145 – 1159.

[Domingos and Pazzani (1997)] Domingos, P., Pazzani, M.J. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29, pp. 103 – 130.

[Friedman, Geiger and Goldszmidt (1997)] Friedman, N., Geiger, D., Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, 29, pp. 131 – 163.

[Hanley and McNeil (1982)] Hanley, J.A., McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143, pp. 29 – 36.

[Heijnen et al. (2007)] Heijnen, E.M.E.W., Eijkemans, M.J.C., de Klerk, C., Polinder, S., Beckers, N.G.M., Klinkert, E.R., Broekmans, F.J., Passchier, J., Te Velde, E.R., Macklon, N.S., Fauser, B.C.J.M. (2007). A mild treatment strategy for in-vitro fertilisation: a randomised non-inferiority trial. *Lancet*, vol. 369, pp. 743 – 749.

[Lachiche and Flach (2003)] Lachiche, N., Flach, P. (2003). Improving accuracy and cost of two-class and multiclass probabilistic classifiers using ROC curves. *Proceedings of the Twentieth International Conference on Machine Learning*.

[Langley and Sage (1994)] Langley, P., Sage, S. (1994). Induction of selective Bayesian classifiers. *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pp. 399 – 406.

[Larrañaga et al. (2006)] Larrañaga,P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J.A., Armañanzas, R., Santafé, G., Pérez, A., Robles, V. (2006). Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7, pp. 86 – 112.

[Ling, Huang and Zhang (2003)] Ling, C.X., Huang, J., Zhang, H. (2003). AUC: a better measure than accuracy in comparing learning algorithms. In Y. Xiang and B. Chaib-draa (Eds.): *Advances in Artificial Intelligence: 16th Conference of the Canadian Society for Computational Studies of Intelligence, AI 2003*, Springer-Verlag, pp. 329 – 341.

[McLachlan (1992)] McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*, Wiley.

[Ng and Jordan (2002)] Ng, A.Y., Jordan, M. (2002). On discriminative vs. generative classifiers: a comparison of naive Bayes and logistic regression. In: T.G. Dietterich, S. Becker, Z. Ghahramani (editors). *Advances in Neural Information Processing Systems*, 14, pp. 605 – 610. MIT Press, Cambridge, Mass.

[Schrage, Van IJzendoorn and Van der Gaag (2005)] Schrage, M.M., Van IJzendoorn, A., and Van der Gaag, L.C. (2005). Haskell ready to Dazzle the real world. *Proceedings of the 2005 ACM SIGPLAN Workshop on Haskell*, ACM Press, New York, pp. 17 – 26.
(for further information, see `http://www.cs.uu.nl/dazzle`)

[Twardy et al. (2006)] Twardy, C.R., Nicholson, A.E., Korb, K.B., McNeil, J. (2006). Epidemiological data mining of cardiovascular Bayesian networks. *Electronic Journal of Health Informatics*, vol. 1, no. 1.

[Verberg et al. (2007)] Verberg, M.F.G., Eijkemans, M.J.C., Macklon, N.S. Heijnen, E.M.E.W., Fauser, B.C.J.M., Broekmans, F.J. (2007). Predictors of ongoing pregnancy after single-embryo transfer following mild ovarian stimulation for IVF. *Fertility and Sterility*, in press.