The impact of modeling the dependencies among patient findings on classification accuracy and calibration

Stefano Monti[†]
[†]Intelligent Systems Program
University of Pittsburgh
901M CL, Pittsburgh, PA - 15260
smonti@isp.pitt.edu

Gregory F. Cooper[†‡]
[‡]Center for Biomedical Informatics
University of Pittsburgh
8084 Forbes Tower, Pittsburgh, PA - 15261
gfc@cbmi.upmc.edu

ABSTRACT

We present a new Bayesian classifier for computer-aided diagnosis. The new classifier builds upon the naive-Bayes classifier, and models the dependencies among patient findings in an attempt to improve its performance, both in terms of classification accuracy and in terms of calibration of the estimated probabilities. This work finds motivation in the argument that highly calibrated probabilities are necessary for the clinician to be able to rely on the model's recommendations. Experimental results are presented, supporting the conclusion that modeling the dependencies among findings improves calibration.

INTRODUCTION

The development of medical diagnostic decision-support systems is a field of research that has received considerable attention in the past thirty years, and numerous systems have been developed to this end [1]. When developing a diagnostic model, statistical or otherwise, we are faced with the problem of assessing its performances. This assessment is usually based on a validation process whereby the model is applied to a test set (i.e., to the solution of a set of clinical cases for which the actual outcome is known), and the diagnosis provided by the model for each of the cases in the test set is compared with the actual diagnosis.

Summary statistics of interest, such as the proportion of cases correctly classified, positive predictive value (PPV), negative predictive value (NPV), and ROC curve area [2], are collected and possibly compared with the corresponding statistics for competing models.

A problem with the adoption of summary statistics measuring classification accuracy only, such as PPV and NPV, or ROC curve area, is that these statistics do not evaluate whether the degree of confidence with which each diagnosis provided by a model is produced is valid. With statistical diagnostic models, the probability attributed to a given diagnosis is the natural measure of the degree of confidence in that diagnosis. A probability p of a given outcome o, is considered well *calibrated* when cases assigned a probability p of yielding outcome o, actually yield outcome o approximately $100p\%$ of the times. Accordingly, whether a probability is well calibrated tells us whether the degree of confidence assigned to a given diagnosis is valid.

In this paper, we address the issue of a model's calibration and of its classification accuracy in the context of Bayesian models for decision-theoretic computer-aided diagnosis. These models operate by specifying a probability distribution over the set of possible diagnostic outcomes, conditioned on a set of relevant clinical findings. The diagnostic decision is then based on the provided probability distribution.

Recent results [3, 4] show that, for classification purposes, the calibration of the probabilities produced by a classifier is not necessary to achieve high classification accuracy as measured by means of a zero/one loss (i.e., by measuring the proportion of cases correctly classified by the model). These conclusions can be intuitively understood by noticing that if the actual probability of the correct classification for a given case is p, the classifier will be equally accurate if it assigns to that classification any probability equal to or higher than p. That is, if the classifier consistently errs in the direction of higher probabilities for the favored classification.

These results can be naturally applied to the analysis of diagnostic models. In fact, provided we assume the set of possible diagnoses to be an exhaustive set of mutually exclusive outcomes, the diagnostic task can be interpreted as a classification task, whereby a set of clinical findings are to be assigned to the correct diagnosis.

In computer-aided diagnosis, the calibration of the probabilities output by the model may be as important as the classification accuracy for the clinician to be able to rely on the system's recommendations. This claim is based on a decision analyti-

1091—8280/98/$5.00

cal view of the diagnostic task, with the clinician acting as decision maker. We apply these considerations to the analysis of the *naive-Bayes* (NB) model [5], also known as simple Bayes, or independence Bayes. We provide evidence supporting the conclusion that the assumption of independence among patient's findings adversely affects the calibration of the model's probability estimates, and we propose a new Bayesian classifier that models the findings' dependencies. We present experimental results aimed at comparing the classification accuracy and the calibration of the new model with the NB model. For the comparison, we use a medical database of pneumonia patients collected over several medical institutions in North America. The results show that, while modeling the probabilistic dependencies of the findings does not improve diagnostic accuracy, it positively affects the calibration of the estimated probabilities.

BACKGROUND

In this section, we briefly describe the Bayesian classifiers that are the building blocks of the new model proposed in the next section.

The naive-Bayes model

As discussed in the introduction, recent results [3, 4] show that the calibration of the probabilities produced by a classifier is not necessary to achieve high classification accuracy. A notable example supporting these results is the naive-Bayes (NB) model [5]. The NB model is one of the most popular classification methods used for computer-aided diagnosis [7-9]. One of its first applications to a diagnostic task was explored more than thirty years ago by Warner et al. [10], and it is still today the object of active research [3, 11]. Part of its success is due to its simplicity and interpretability. Furthermore, several empirical studies have shown that it is very accurate, and can often outperform more complex and sophisticated classification methods, despite the strong assumptions on which it is based [3]. The main assumption necessary to the application of the NB model is that all the findings are mutually independent conditioned on the outcome variable. A graphical representation of the NB model is shown in Figure 1.a. The outcome variable O is defined as the common *parent* of the findings $\boldsymbol{F} = \{F_1, \ldots, F_n\}$, and each of the findings F_i is a *child* of the outcome variable O. The independence assumptions implied by the model allow for the following factorization of the conditional probability $\mathrm{P}(O \mid \boldsymbol{F})$ of the outcome variable given a set of findings:

$$\begin{aligned} \mathrm{P}(O \mid \boldsymbol{F}) &= \frac{\mathrm{P}(O, \boldsymbol{F})}{\sum_{O'} \mathrm{P}(O', \boldsymbol{F})} \\ &= \frac{\mathrm{P}(O) \prod_{i=1}^{n} \mathrm{P}(F_i \mid O)}{\sum_{O'} \mathrm{P}(O') \prod_{i=1}^{n} \mathrm{P}(F_i \mid O')}. \end{aligned} \tag{1}$$

It follows that for the specification of the model, we need to estimate the prior probability $\mathrm{P}(O)$, and the conditional probabilities $\mathrm{P}(F_i \mid O)$ of each finding F_i given the outcome variable O. These probabilities can be easily estimated from data for both discrete and continuous variables.

The conditional independence assumption is often violated by the data, but both empirical results and theoretical analysis suggest that this violation does not necessarily affect classification accuracy [3, 4, 11-13], a point to which we will return.

The finite mixture model

An alternative to the NB model that allows for the relaxation of the conditional independence assumption is the *finite mixture* (FM) model [6]. In a FM model, all the dependencies between observed variables, both the findings and the outcome variable, are assumed to be modeled by a single discrete latent (i.e., unobserved) variable. In a FM model, the outcome variable is itself a child node, and the common parent is a latent variable. With reference to Figure 1.b, the parent node L represents an unmeasured, discrete variable, which models the interaction among the findings $\{F_i\}$, as well as the interaction between the findings and the outcome variable O. Based on the conditional independencies implied by the FM model, the conditional probability $\mathrm{P}(O \mid \boldsymbol{F})$ can be factored as follows:

$$\begin{aligned} \mathrm{P}(O \mid \boldsymbol{F}) &= \frac{\sum_L \mathrm{P}(L)\mathrm{P}(O, \boldsymbol{F} \mid L)}{\sum_L \mathrm{P}(L) \sum_{O'} \mathrm{P}(O', \boldsymbol{F} \mid L)} \\ &= \frac{\sum_L \mathrm{P}(L)\mathrm{P}(O \mid L) \prod_i \mathrm{P}(F_i \mid L)}{\sum_L \mathrm{P}(L) \prod_i \mathrm{P}(F_i \mid L)}. \end{aligned} \tag{2}$$

It follows that for the complete specification of the FM model, we need to estimate the prior probability $\mathrm{P}(L)$, and the conditional probabilities $\mathrm{P}(O \mid L)$ and $\mathrm{P}(F_i \mid L)$.

Learning an FM model from data consists of two steps: 1) determination of the number of values of the latent variable L; and 2) parameter estimation. The determination of the number of values of the latent variable is the most difficult step. Exact parameter estimation is not feasible in general, because of the presence of the latent variable, and ap-

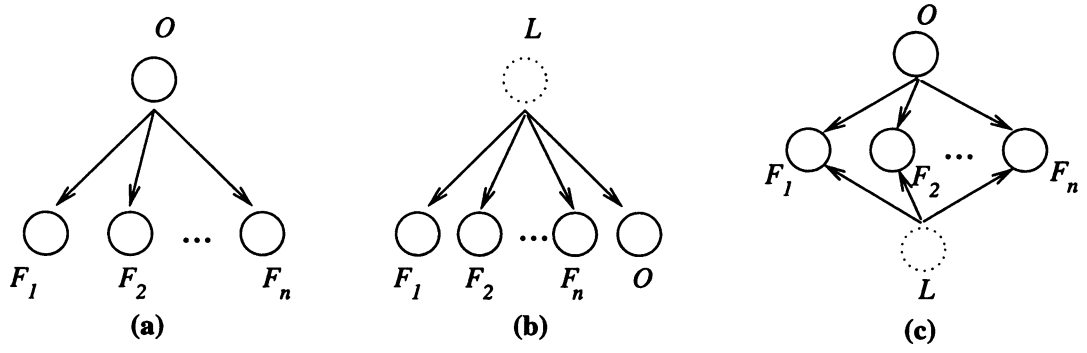

Figure 1: Bayesian classifiers: a) the naive-Bayes (NB) model with outcome variable O and set of findings $\{F_i\}$, which are independent given O; b) the finite mixture (FM) model, where the hidden variable H models the dependencies among all the observed variables; and c) the finite-mixture-augmented naive-Bayes (FAN) model, obtained by superimposing an FM model on the set of feature variables of an NB model.

proximate iterative methods are usually adopted. For lack of space, we refer the interested reader to a detailed discussion of these topics in [15].

METHODS

The classifier that we describe in this section combines the two models described in the previous section, while relaxing the assumptions on which they are based. We call the new classifier the Finite-mixture-Augmented Naive Bayes (FAN) model.

As shown in Figure 1.c, the proposed classifier is obtained by superimposing a FM model on the set of findings of a NB model. That is, the latent variable L is introduced to model the residual probabilistic dependencies between the findings $\{F_i\}$ that are not captured by the outcome variable O. At the same time, in an attempt to improve over the FM model, the FAN model reduces the burden on the latent variable L by modeling part of the dependencies among findings through the outcome variable O. Notice that the NB model is subsumed by the FAN model, since it corresponds to a trivial FAN model with a one-valued latent variable.

Based on the conditional independencies implied by the FAN model, the conditional probability $\mathrm{P}(O \mid \boldsymbol{F})$ can be factored as follows:

$$\mathrm{P}(O \mid \boldsymbol{F}) = \frac{\sum_L \mathrm{P}(L)\mathrm{P}(O, \boldsymbol{F} \mid L)}{\sum_L \mathrm{P}(L) \sum_{O'} \mathrm{P}(O', \boldsymbol{F} \mid L)} = \frac{\mathrm{P}(O) \sum_L [\mathrm{P}(L) \prod_i \mathrm{P}(F_i \mid O, L)]}{\sum_{O'} \{\mathrm{P}(O') \sum_L [\mathrm{P}(L) \prod_i \mathrm{P}(F_i \mid O', L)]\}}. \quad (3)$$

The method for learning a FAN model from data is a straightforward adaptation of the corresponding method for learning an FM model from data.

Experimental design

For the experimental evaluation, aimed at comparing the three models described in the previous sections, we used the pneumonia PORT database [16, 17]. This database contains the results of an observational study of outpatients and inpatients with community-acquired pneumonia, conducted at five medical institutions in North America. The database has a total of 2287 cases, with each case corresponding to a different patient. In our experiments, each case is described by a total of 159 variables, including the outcome variable, and 158 predictors, corresponding to demographics, symptoms, physical findings, and test results. The outcome variable used in the experiments is `DIREOUT`, which registers the occurrence or non-occurrence of any of the following events: a severe complication, death within 30 days of being seen initially, or ICU admission for respiratory failure, respiratory or cardiac arrest, or shock/hypotension. The dataset was randomly partitioned into a training set of 1601 cases and a test set of 686 cases, both containing approximately the same proportion of positive cases of `DIREOUT`. For each of the three models considered, we built the model based on the training set, and collected the relevant summary statistics on the test set.

RESULTS

For the comparison of the classification accuracy of the three models, we use the area under the receiver operating characteristic (ROC) curve [2]. The ROC curve plots the true positive rate as a function of the false positive rate as we vary from

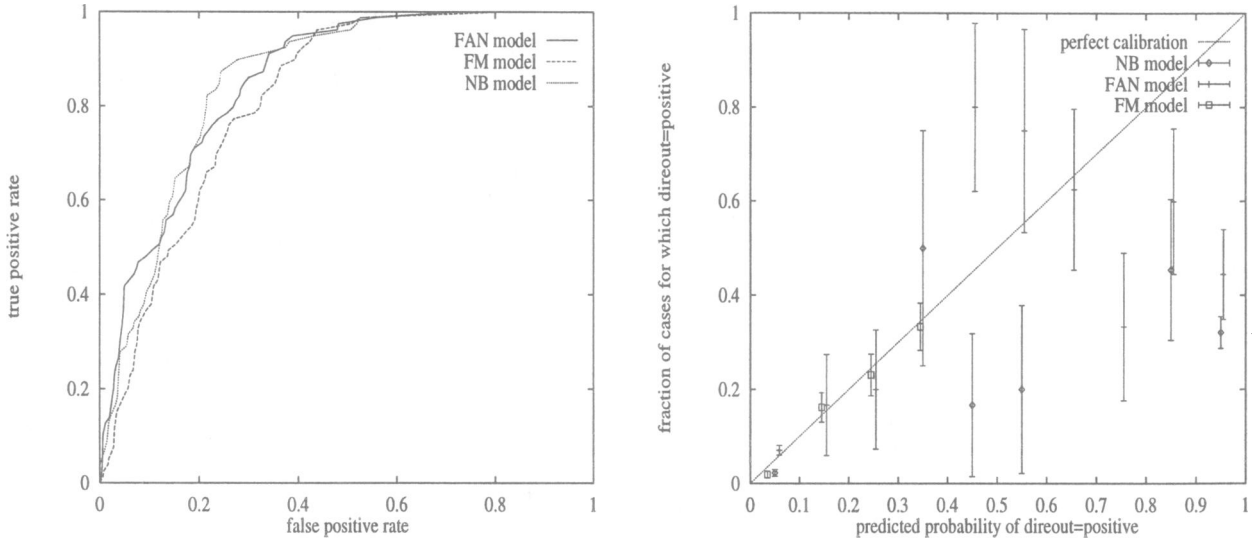

Figure 2: a) ROC curves; and b) calibration plots for the three models being compared.

case as positive. Each point on the curve characterizes the performance of the classifier for a given threshold. The area under the ROC curve can be used as a measure of classifier performance. An area of 1 corresponds to perfect accuracy. An area of .5 corresponds to the performance of a classifier that randomly guesses the outcome.

Table 1 (2nd column) reports the ROC area for the different models we considered. A statistical comparison of the ROC areas [18] shows that there is no statistically significant difference in the ROC area of the NB model and the FAN model ($p > .5$), while the ROC area for the FM model is significantly worse ($p < .05$). By looking at Figure 2.a, we can see that the ROC curves for the NB model and the FAN model are similar, while the ROC curve for the FM model is consistently dominated by the ROC curves for the other two models.

For the comparison of the models' calibration, we present descriptive plots accounting for their calibration. Notice that, if we knew the actual probability of each case's outcome, that is, its asymptotic frequency as the sample size goes to infinity, then for a perfectly calibrated model the plot of the actual probability of each case against the probability assigned by the model to the corresponding case would be the straight line $y = x$. Based on this observation, we specify the calibration plot for a model as follows. We divide the probability range in equal bins of width .1, and for each bin, we report the proportion of actual positive cases that are assigned a probability within the range specified by the bin. For example, if we consider the third bin, corresponding to the probability range between .2 and .3, we compute the proportion of actual positive cases among all the cases that were assigned a probability between .2 and .3 by the model. For a well calibrated model, this proportion should fall within the bin's range.

As shown in Figure 2, the plot for each model is specified by a series of 95% confidence intervals (CIs), one for each of the ten bins, centered at the mid-point of the probability range specified by each bin. No CI is reported if the model does not assign a probability within the bin's range to any case in the test set. The diagonal line corresponding to perfect calibration is also plotted as a reference. The comparison of the calibration plots of Figure 2.b is summarized in Table 1 (3rd column). From Figure 2.b, we can see that the NB model is consistently miscalibrated. In fact, for only one out of six bins the 95% CI intersects the perfect-calibration line. The FAN model is considerably better calibrated, since the CI of 4 out of 9 bins

model	*ROC area*	*Calibration*
NB model	.8505	1 of 6 bins (~17%)
FAN model	.8493	4 of 9 bins (~45%)
FM model	.8125	3 of 4 bins (~75%)

Table 1: Area under the ROC curve, and summary of calibration plots.

intersects the perfect-calibration line. Finally, the FM model is very well calibrated (3 out of 4 bins), but it does not assign any probability above the .4 threshold. That is, it never classifies a case as positive with a large degree of confidence.

Our results are consistent with results presented in similar studies [3, 11-13] that show that modeling the conditional dependencies among the patient findings does not necessarily increase classification accuracy. Significantly however, our results also show that modeling the dependencies among findings improves the calibration of the model.

CONCLUSIONS

We have presented a new Bayesian classifier, which builds upon the naive Bayes model, while relaxing the strong assumptions of probability independence on which that model is based, in an attempt to improve the calibration of its probability estimates. The rationale for the new model is the idea that well calibrated probability estimates are necessary for the clinician to be able to rely on the recommendations provided by a probabilistic diagnostic system. For the comparison of the calibration of different models, we used calibration plots. However, we plan to investigate statistics that measure the calibration of a model, so as to be able to compare competing models, and test the statistical significance of possible differences.

Acknowledgments

We thank the pneumonia PORT project, particularly Dr. W. Kapoor and Dr. M. Fine, for providing access to the data. We thank Mr. S. Obrosky for his help in using the data, and the CEHC project members for providing a stimulating intellectual environment that positively influenced our research. This research was supported in part by grants BES-9315428 and IRI-9509792 from the NSF and by grant LM059291 from the NLM.

References

[1] Miller RA. Medical diagnostic decision support systems – past, present, and future: A threaded bibliography and commentary. *J Am Med Informatics Assoc*, 1(1):8–27, 1994.

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

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

[4] Friedman JH. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1:55–77, 1997.

[5] Good IJ. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* M.I.T. Press, Cambridge, MA, 1965.

[6] Cheeseman P, Stutz J. Bayesian classification (AutoClass): Theory and results. *Advances in Knowledge Discovery and Data Mining*, 1996.

[7] Gorry GA, Barnett GO. Experience with a model of sequential diagnosis. *Comput Biomed Res*, 1:490–507, 1968.

[8] Starmer CF, Lee KL. A mathematical approach to medical deicisions: Application of Bayes' rule to a mixture of continuous and discrete clinical variables. *Comput Biomed Res*, 9:531–541, 1976.

[9] Todd BS, Stamper R. The relative accuracy of a variety of medical diagnostic programs. *Methods Inf Med*, 33:402–416, 1994.

[10] Warner HR, Toronto AF, George Veasey L, Stephenson R. A mathematical approach to medical diagnosis. *J Am Med Assoc*, 177(3):177–183, 1961.

[11] Todd BS, Stamper R. Limits to diagnostic accuracy. *Med Inform*, 18(3):255–270, 1993.

[12] Chard T. The effect of dependence on the performance of the Bayes theorem: An evaluation using a computer simulation. *Comput Methods Programs Biomed*, 29(1):15–19, 1989.

[13] Gammerman A, Thatcher AR. Bayesian diagnostic probabilities without assuming independence of symptoms. *Methods Inf Med*, 30:15–22, 1991.

[14] Dougherty J, Kohavi R, Sahami M. Supervised and unsupervized discretization of continuous features. *Machine Learning: Proocedings of 12th International Conference*, 1995.

[15] Monti S, Cooper GF. A Bayesian network classifier that combines a finite mixture model and a naive Bayes model. Technical Report ISSP-98-01, University of Pittsburgh, March 1998.

[16] Kapoor NK. Assessment of the variation and outcomes of pneumonia. Pneumonia Patient Outcome Research Team (PORT) final report. Agency for HealthCare Policy and Res, 1996.

[17] Fine M, Auble TE, Yealy DM, Hanusa BH et al. A prediction rule to identify low-risk patients with community-acquired pneumonia. *N Engl J Med*, 336(4):243–250, 1997.

[18] Hanley JA, McNeil BJ. A method of comparing the areas under Receiver Operating Characteristics Curves derived from the same cases. *Radiology*, 148:839–843, 1983.