HOW RELIABLE IS A MEASURE OF MODEL RELIABILITY? BOOTSTRAP CONFIDENCE INTERVALS OVER VALIDATION RESULTS

Marcin Kozniewski[1,3], Mario A. Cypko[2], Marek J. Druzdzel[1,3]

[1] School of Information Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

[2] The Innovation Center for Computer Assisted Surgery, University of Leipzig, Leipzig, Germany

[3] Faculty of Computer Science, Bialystok University of Technology, Bialystok, Poland

Abstract: A researcher testing a model will frequently question the reliability of the test results, understanding well the intuition that verification performed on a handful of cases is less reliable than verification based on very large numbers of cases. Because a limited number of verification cases happens pretty often in very specific domains, a question of practical importance is, thus, how reliable is a reported reliability measure.
We propose a methodology based on deriving confidence intervals over various measures of accuracy of Bayesian network models by means of bootstrap confidence intervals. We evaluate our approach on ROC and calibration curves derived for a model derived from an UC Irvine Machine Learning Repository data set and a sizeable (over 300 variables) practical model constructed using expert knowledge and evaluated on merely 66 accumulated real patient cases. We show how increasing the number of test cases impacts the width of confidence intervals and how this can aid in estimating a reasonable number of verification cases that will increase the confidence in model reliability.

Keywords: Bayesian networks, bootstrap confidence intervals, validation

1. Introduction

Bayesian networks (BNs) [13] allow for intuitive, flexible, yet theoretically sound, modeling of uncertain domains. They are acyclic directed graphs, in which nodes represent random variables and edges represent direct dependencies between pairs of variables. These dependencies are expressed numerically by means of conditional

Advances in Computer Science Research, vol. 13, pp. 27-41, 2016.

probability distributions of every node conditional on its direct predecessors (parents) in the graph.

BN models readily combine a variety of available information sources, such as data and expert opinion. A variety of methods for building Bayesian network models have been devised. When data are readily available, networks can be learned automatically [14,4,16]. When no data are available, networks can be entirely elicited from experts (e.g., [5], described later in this paper). Combinations of the two approaches can be used in all cases when some data are available.

BN models are compact representations of the joint probability distribution (JPD) over the variables that they represent. Given an observation of a subset of BN's variables (evidence), it is possible to calculate the conditional posterior probability distribution over the remaining variables. This probability can be used for risk assessment, diagnosis, prognosis, and other tasks. Before a Bayesian network model can be embedded into a decision support system, one would like to know its reliability in terms of the accuracy of its results. The problem is of particular importance in all applications where a potential system error may cause serious practical implications, such as in most medical applications. Arguably the strongest, objective way of assessing the accuracy of a system is to test the system on a set of cases that it has never seen. Several statistical methods are used for assessing different aspects of model quality, such as accuracy, sensitivity and specificity, receiver operating characteristic (ROC) curves, area under the ROC curve (AUC), calibration curve, etc. [8,9,12,6,7].

When the network has been learned from data, one can set aside a subset of data and use that subset for the purpose of verification after learning the network from the remaining (training) records. When a system has been constructed from expert opinion, one can collect independently real cases and use these for verification. In both cases, the number of verification cases is usually limited, either because of lack of data (for example, one might hope that there are not too many data records for serious airplane malfunctions; a limited number of patients is seen with some rare disorders) or because setting aside data records for verification purposes takes them away from the learning set and reduces the quality of the model at the outset. A researcher testing a model on a limited number of cases will frequently question the reliability of the verification result, understanding well the intuition that tests performed on a handful of cases are less reliable than those based on very large numbers of cases. After all, a limited number of cases will rarely cover all important elements of the model. Practical questions of vital importance are, thus, how reliable is a reported reliability measure and is there a need for more cases for reliable verification.

In this paper, we address these questions by deriving confidence intervals over various measures of accuracy of Bayesian network models. Our approach is based on

producing bootstrap confidence intervals from the available verification data. These intervals are going to be broad when the number of verification cases is small and narrow when their number is large. Too broad confidence intervals will indicate the need for more verification cases and will help in evaluating the usefulness of a new decision aid. Such approach is well established for ROC curves [15]. Furthermore, we propose to analyze the size of confidence intervals for a subset of the testing set. If the size of a confidence interval does not change too much for a smaller subset of test cases, it means that bringing more data will not increase significantly the reliability of the evaluation of the model.

We evaluate our approach on ROC and calibration curves derived for three Bayesian network models: (1) two models learned automatically from the *Cover Type* data set [1] available from the Irvine Machine Learning Repository, and (2) a sizeable (over 300 variables) practical model for cancer therapy planning, constructed using experts' knowledge for representing the tumor-, lymph nodes- and metastasis staging (TNM staging for laryngeal cancer)[17]. In (2), the number of patient records available for validation is small (only 66 prior patient cases), so the question of model reliability is of vital practical interest. In all cases, we show how increasing the number of test cases impacts the width of confidence intervals.

The remainder of this paper is structured as follows. Section 2. explains two widely used methods for the assessment of model accuracy: ROC curves and calibration curves. Section 3. describes our approach to deriving confidence intervals over model validation measures based on bootstrap confidence intervals. Section 4. describes out experiments testing our method in practice.

2. Model Quality Validation Methods

In this section, we review two important measures of accuracy of probabilistic systems: (1) ROC curves, and (2) calibration curves.

2.1 Receiver Operating Characteristic (ROC) Curves

Several measures of accuracy have been proposed for systems performing tasks such as classification, prediction, or diagnosis. The most straightforward method is accuracy, which, unfortunately does not give sufficient insight into system's performance. When the asymmetry among classes is very large, a system betting always on the prevalent class will perform well, while it may be of practical importance to identify correctly unlikely classes (e.g., rare but serious disorders in the domain of medicine) at the expense of overall accuracy. Much better measure of accuracy are per class

parameters known as sensitivity and specificity, which are expressing the system's ability to identify the class correctly when present and when absent, respectively. Most researchers report evaluation results in a *confusion matrix*, which is a table listing the total number of correctly and incorrectly identified instances for each of the classes.

Better yet in characterizing the quality of a system is a plot known as *receiver operating characteristic* (ROC) curve [8,9], which shows the tested system's ability to identify a class. The ROC curve plots the true positive rate (i.e., sensitivity) as a function of the false positive rate ($1-$specificity) and shows the ability of a model to distinguish a class for a continuum of decision criteria. The closer an ROC curve is to the upper left corner (0,1) (i.e., to perfect values of sensitivity=1.0 and specificity=1.0), the better the model. When the decision criterion (e.g., a probability threshold) for choosing a class is changed, the sensitivity and specificity for this class also change, within the constraints shown by the ROC curve. The ROC curve can be seen as a exhaustive collection of confusion matrices, as each point on the ROC curve corresponds to one possible matrix, resulting from the modeler's decision when to identify a class and what sensitivity to choose. The ROC curve shows clearly the compromise that the designer of a decision support system has to make by choosing a threshold (and fixing the combination of sensitivity and specificity values), that optimizes the utility of a decision. Figure 1 shows ROC curves for a collection of models predicting the day of female ovulation in the context of a model for fertility awareness [11]. The curves illustrate the idea that the exact values of sensitivity and specificity are the result of a designer's choice. When the model is used by a couple seeking pregnancy, the optimal point and the optimal model are different than when it is used by couples who want to avoid pregnancy. In the latter case, one would want to choose a model with near perfect sensitivity, i.e., with almost-zero false negatives.

Similarly to sensitivity and specificity, the ROC curve is meaningful only in expressing the system's ability to detect a single class. Whenever a system focuses on detecting multiple classes, or multiple grades of a single class, the ROC curve is plotted for a single class and a single value with all remaining classes or values lumped together as a complement of the class in focus.

2.2 Calibration curve

A less popular but not less important measure of quality of probabilistic systems, such as those based on Bayesian networks, is their accuracy in probability estimates. When a system derives the probability of cancer to be 0.07, for example, one would

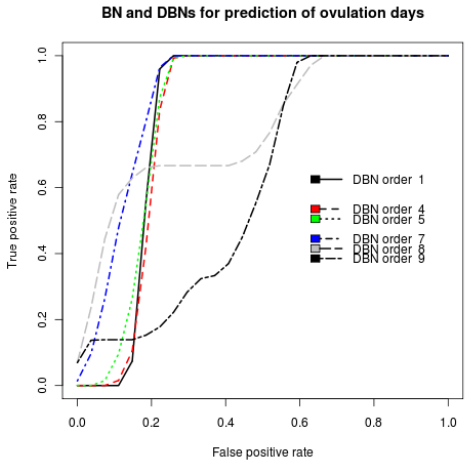

Fig. 1. A collection of ROC curves for a collection of models predicting the day of ovulation [11].

like the actual prevalence of cancer among patients with similar characteristics to be close to 7%. The main reason for importance of precision in probability estimates is that these are fundamental in decision making. No prediction is certain and in order to make a rational decision, a decision maker needs to know the probability distributions over the possible states of the world. Decision theory prescribes an optimal decision to be one that maximizes the expected utility [2]. Utility is a measure of desirability of outcomes that can be combined in the same way as probabilistic expectation. The more precise the probability estimates, the better quality of the resulting decisions.

Accuracy of probability estimates is expressed by *calibration curves*, which are also known in the area of weather forecasting as *reliability diagrams* [12,6]. While the term *reliability diagram* testifies to the importance of accuracy in probability estimates in practical systems, we find it somewhat too general and prefer the term *calibration curve* instead.

The calibration curves express the relationship between the estimated probabilities (horizontal axis) and the observed frequencies (vertical axis) in the data set. In practice, calibration curves are constructed by dividing records with similar probability estimates $\Pr(E)$ produced by the system into bins. For each bin, the frequency $f(E)$ of the event E in the data (e.g., the prevalence of the class in question) is calculated. Calibration curve is a plot of $f(E)$ as a function of $\Pr(E)$. A model that is perfectly calibrated will have a calibration curve that is a diagonal line from the point

(0,0) to point (1,1). Every probability estimate of the model corresponds to identical frequency in the data.

Figure 2 shows an example calibration curve for a system's estimate of the tumor state *t4a* from a TNM staging model. The curve departs from the diagonal line and is also rugged, which is caused by the small size of the test data set (only 66 patient records).

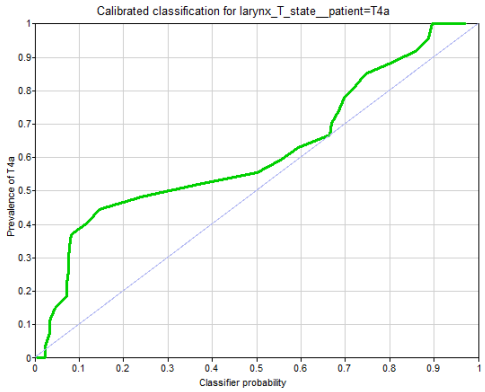

Fig. 2. The calibration curve for the probability of $T4a$ state produced by a TNM staging model.

The calibration curve is capable of expressing overconfidence (or underconfidence) of a system, which provides important information for model builders.

3. Bootstrap Confidence Intervals for Evaluation Statistics

A simple statistical technique for deriving confidence intervals over an estimate can be based on bootstrap resampling [7,3]. Given a representative sample, bootstrap resampling allows to create a large collection of samples with similar statistical properties. Bootstrap resampling has been shown to provide very good results when the original sample is representative for the population. If all records in the sample are indeed independent and identically distributed and originate from the same joint probability distribution, the sample has a very high chance of being representative, i.e., reflect the statistical properties of the distribution. In our work, we start with the sample used for model validation. We assume that it is representative and subsequently use bootstrap resampling to derive the confidence intervals over any statistic based on the original sample. While bootstrap resampling can be used to derive any statistic,

we demonstrate the power of this technique on two statistics of interest in the context of model validation: (1) confidence intervals over the ROC curves, and (2) confidence intervals over the calibration curves.

In the remainder of this section, we describe how confidence intervals and confidence areas for validation curves are obtained. Then, we sketch a technique for capturing the change in the confidence intervals as a function of the size of the test data set.

3.1 Bootstrap confidence intervals

Bootstrap confidence intervals are constructed from statistics calculated for artificial samples drawn from the original sample with replacement. The general procedure is as follows.

i. Having a data set D with n records, create m bootstrap samples by drawing n elements with replacement from D.
ii. For each bootstrap sample D_j, $j = 1, \ldots, m$, calculate the desired statistic for each sample.
iii. Sort the calculated statistics to get $a^{(1)} \leq a^{(2)} \leq \cdots \leq a^{(m)}$.
iv. Take the $(m(\alpha/2))$th element $(a^{(m(\alpha/2))})$ as the lower bound and the $(m(1-\alpha/2))$th element $(a^{(m(1-\alpha/2))})$ as the upper bound of the $100(1-\alpha)\%$ confidence interval.

This method extends readily to confidence intervals over a curve by calculating the confidence intervals over every point on the curve. For each x value we obtain m y-values, which creates a sample to produce a confidence interval of y for each value of x. The extended method can be described schematically as follows.

i. Create m bootstrap samples by drawing n elements with replacement from D.
ii. For each bootstrap sample, create a curve c_j (i.e., ROC or calibration curve).
iii. Iterate through the representative set of values of $x^* \in [0,1]$ (e.g., 100 values from range $[0,1]$ with step of 0.01):
   (a) Generate the set of y values that reflects all the points from generated curves with $x = x^*$,
   (b) Sort the y values
   (c) Create a confidence interval $[y_L, y_U]$ over the y values by taking the $(m(\alpha/2))$th and the $(m(1-\alpha/2))$th elements.
iv. This procedure results in construction of two curves: (1) one curve representing the lower bound of the confidence interval, constructed by points of the y_L value for a given x, and (2) second curve as the upper bound is constructed by the y_U values.

By connecting all lower bounds and all upper bounds of these intervals we obtain the lower and upper bound of the confidence region. Figure 3 shows an example of the confidence region over a ROC curve. In iii we use an iteration step of 0.01.

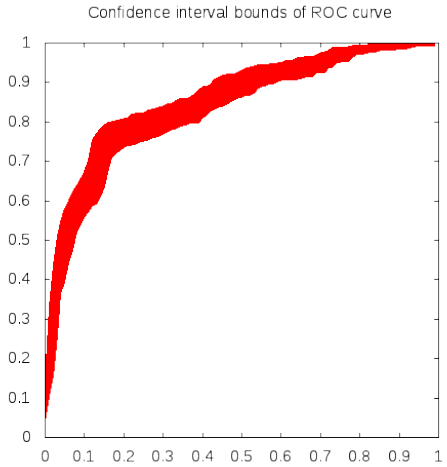

Fig. 3. Sample confidence region over a ROC curve

Confidence regions express the uncertainty about the statistic derived for the model, conditional on the test set used. A plot of the confidence region is useful in determining whether the given data set is sufficient for model assessment with a given validation method.

3.2 Dynamics of the Area of a Confidence Region

A plot of the confidence region suggests a summary statistic, which is the total area of the confidence region (ACR). For both ROC and calibration curves, ACR ranges between 0 and 1. When no validation records are available, there is maximum uncertainty about the ROC and calibration curves and ACR takes the value of 1. We expect, that as the size of the validation set approaches infinity, the area of the confidence region will approach a constant value.

A useful extension of ACR, showing the dynamics of the confidence regions, is a plot of the ACR as a function of the size of the validation set. The larger the validation data set, the smaller the area of the confidence region. We can simulate it by creating subsets of the original dataset. Plotting the area of the confidence region

as a function of the validation set size may be helpful in finding the optimal test set size. We will show the ACR plots in our experiments in Section 4..

4. Evaluation

We evaluate our approach on ROC and calibration curves derived for three Bayesian network models: (1) two models learned automatically from the *Cover Type* data set [1] available from the Irvine Machine Learning Repository, and (2) a sizeable (over 300 variables) practical model for cancer therapy planning, constructed using experts' knowledge for representing the tumor-, lymph nodes- and metastasis staging (TNM staging for laryngeal cancer). The number of validation records (real patient cases) available for the cancer therapy planning model is very small (only 66 prior patient cases). We treat the analysis of this model as an inspiration for a practical application of our work.

4.1 Cover Type Data Set and Models

Our first two models are derived by means of a Bayesian network learning algorithm from the Cover Type data set [1], available from the UC Irvine Machine Learning Repository[4]. The data set consists of 581,012 records over 55 variables. We discretized the continuous variables using uniform interval width discretization method with three intervals. We extracted two disjoint data sets from the original data set for the purpose of our analysis: (1) training data set, one consisting of 81,012 records, and (2) validation data set of 15,000 records for our tests. For the purpose of our experiments, we created a family of validation data sets from (2) by taking subsets of the first 100 records (D_1), the first 200 records (D_2), etc., in such a way that $D_1 \subset D_2 \subset \ldots \subset D_k \subset D$. These subsets simulate the natural process of harvesting more data records for the validation data set.

We created two models from the Cover Type data set (we will refer to them as A and B). For the first model (A), we used all variables and the Bayesian search algorithm [4] implementation in *GeNIe* software[5]. The model consisted of 55 variables connected by a total of 175 arcs. We obtained the second model (B) by mutilating model A. Our mutilation amounted to removing five strongest arcs in model A. Arc strength is a standard function of GeNIe and its theoretical background is described in [10].

[4] `https://archive.ics.uci.edu/ml/datasets/Covertype`

[5] Available free of charge for academic research and teaching use at `http://www.bayesfusion.com/`

We also created a modification of the validation set that contained 50% missing values (these were selected randomly, using uniform distribution).

We ran three experiments for the models A and B:

- model A with original validation data set,
- model B with original validation data set, and
- model A with validation data set with missing values.

In each experiment, we generated regions of confidence as described in Section 3.2 with the size of the validation data set ranging from 100 to 15,000. For each confidence region, we reported the area of the region and the longest confidence interval among x values.

4.2 The TNM-Staging Model

The third model used in our experiments is a detailed representation of the NM staging of laryngeal cancer, created manually by clinical experts from the Leipzig University Hospital (Universitätsklinikum Leipzig, UKL) [17]. The model consists of 303 vertices connected by 334 arcs. Model variables have between 2 and 27 states with average of 4 states. The model is specified by a total of 78,606 parameters obtained from experts. Even though the UKL hospital is highly specialized in head and neck cancer, because challenging head and neck cancers are not common, it receives only around 80 patients per year. The medical records covering more than ten years worth of treatment at UKL are stored, although they are incomplete, unstructured, and recorded mainly in free text format. So far, only 66 patient records have been coded and are suitable for use by the system developers. Many values in these 66 records are missing. The number of observations (values of evidence variables) in these cases vary between 35 and 157 (with the average of 78) per patient record. This is an ongoing project and the data from previous years is successively being added to the validation data set. New patient cases are encoded directly in the destination format.

In our experiment, we derived the confidence regions of the ROC curve and the calibration curve for one of the key decision variables (patient larynx T-state) in the TNM staging model for laryngeal cancer.

4.3 Results

Figure 4 shows the areas of confidence region (ACR) as a function of the validation data set size for each of the states of the *Cover Type* variable. To build this plot, we

created a confidence interval plot for every subset D_i of the validation data set (see Section 4.1) and calculated the area of this plot. Analyzing the plots subjectively, we can see that ACR decreases as the number of validation records becomes larger, reaching a plateau for roughly 2,000 records. The plots show the dynamics of this process and suggests that roughly 2,000 validation records is a reasonable number to get an idea of the accuracy of the model's reliability.

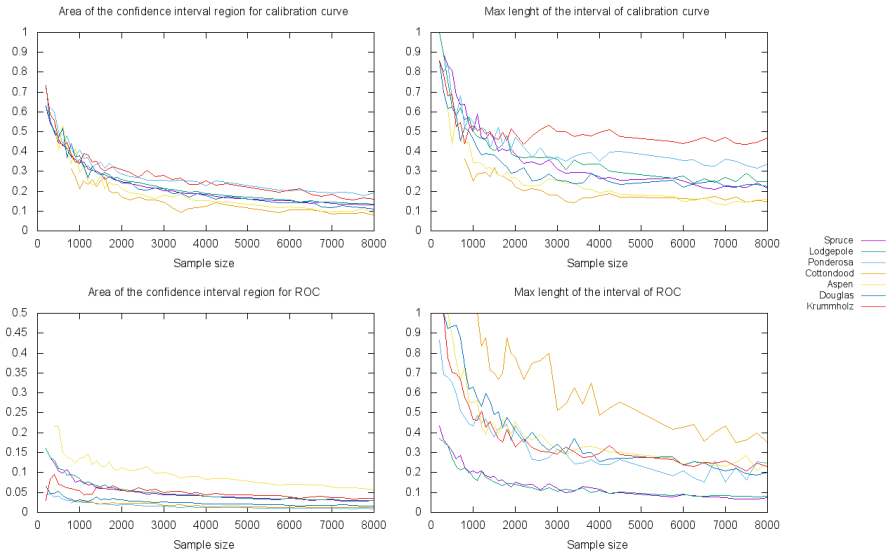

Fig. 4. Changes of areas of confidence regions (ACR) and maximum confidence interval of the ROC and the calibration curve as a function of the number of data records (model A and the original data set).

Figure 5 shows the results of testing the modified model (top) and testing with validation records with missing values (bottom). The results are qualitatively similar to those for model A and the original validation set. Here also 2,000 records seem sufficient for the ACR to reach a plateau beyond which improvement is rather small.

For the TNM staging model, which is a real clinical model, we had only 66 validation records, so the dynamic of the ACR measure is hard to derive. We report only the calibration curves with their 90% confidence intervals. Figure 6 (left) shows the confidence interval of the calibration curve plot for stage T2 laryngeal cancer. The curve seems to suggest that the model is underconfident – the probabilities of the T2 stage of laryngeal cancer tend to be concentrated around moderate probabilities,

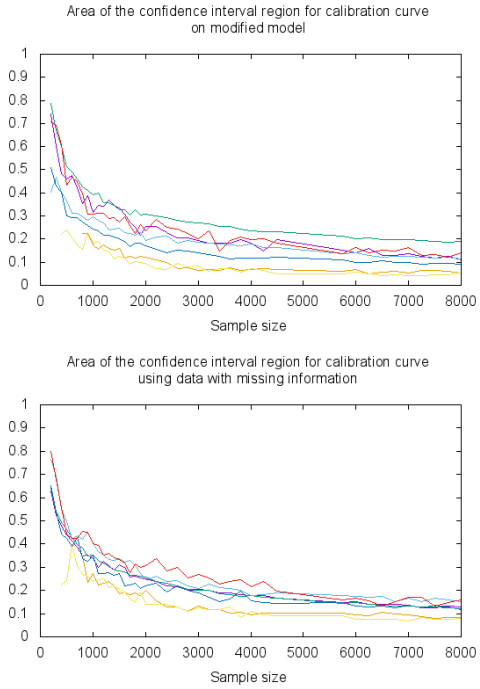

Fig. 5. Changes of areas of confidence region (ACR) as a function of the number of validation data records. Model B tested on the original data set (top). Model A and the data set with missing values (bottom).

while the corresponding frequencies are more extreme. Figure 6 (right) shows an identical plot for the TNM staging model for stage T1a of laryngeal cancer. The model seems to be better calibrated but the confidence intervals are wide and more data are needed to assess the model response more precisely.

5. Conclusion

This paper proposed a methodology for deriving confidence intervals over various measures of accuracy of a Bayesian network model by means of bootstrap resampling. While any validation statistic is amenable for confidence interval analysis, we applied our method to deriving confidence intervals over ROC curves and calibration curves. We have proposed capturing the uncertainty over confidence intervals by means of areas of confidence region (ACR) and have shown how these change as a function of the number of records in the validation data sets. Such plots allow

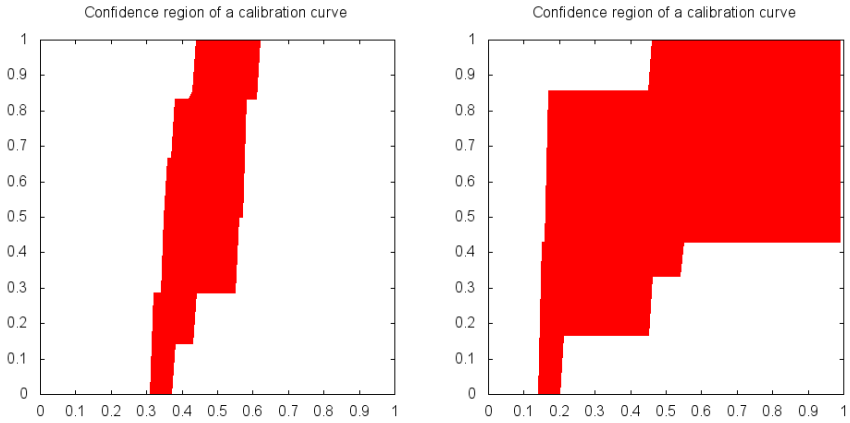

Fig. 6. The 90% confidence region over the calibration curve for the tumor state T2 (left) and state T1a (right) in the TNM-staging model.

for finding a number of records that are necessary to gain reasonable confidence in validity of the model.

The areas of confidence region (ACR) as a function of the number of records curves seems to follow a systematic shape. It may be worth to investigate whether this shape can be characterized theoretically. If so, the number of records needed for a reasonable confidence in evaluation can be predicted a-priori, based on a limited number of validation records.

Finally, our approach gives no insight into the reasons for possibly low confidence in validation, such as flaws in the model structure or its numerical parameters. To gain insight into the possible reasons for low accuracy, more detailed approaches are needed, focusing on the feedback between model inputs and its outputs.

Acknowledgements

We acknowledge the support the National Institute of Health under grant number U01HL101066-01. Implementation of our work is based on GeNIe and SMILE, a Bayesian modeling environment available free of charge for academic research and teaching use at `http://www.bayesfusion.com/` This research has been done in collaboration with the ICCAS research group ”Digital Patient and Process Modeling” funded by the German Ministry of Research and Education.

References

[1] J. A. Blackard and D. J. Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture*, 24(3):131–151, 1999.

[2] R.T. Clemen and T. Reilly. *Making Hard Decisions: Introduction to Decision Analysis*. Duxbury Press, 2005.

[3] P. R. Cohen. *Empirical Methods for Artificial Intelligence*, volume 139. MIT Press Cambridge, 1995.

[4] G. F. Cooper and E. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.

[5] M. A. Cypko, D. Hirsch, L. Koch, M. Stoehr, G. Strauss, and Denecke K. Webtool to support medical experts in probabilistic modelling using large bayesian networks with an example of rhinosinusitis. *Studies in Health Technology and Informatics*, 216:259–263, 2014.

[6] M. H. DeGroot and S. E. Fienberg. The comparison and evaluation of forecasters. *The Statistician*, 32:12–22, 1983.

[7] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. CRC Press, 1994.

[8] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[9] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. Elsevier, 2011.

[10] J. R. Koiter. Visualizing Inference in Bayesian Networks. Master's thesis, Delft University of Technology, June 2006.

[11] A. Łupińska-Dubicka and M. J. Druzdzel. Modeling dynamic processes with memory by higher order temporal models. In A. Hommersom and P.J.F. Lucas, editors, *Foundations of Biomedical Knowledge Representation: Methods and Applications: Lecture Notes in Artificial Intelligence*, volume 9521, pages 219–232. Springer Verlag, 2015.

[12] A.H. Murphy and R.L. Winkler. Reliability of subjective probability forecasts of precipitation and temperature. *Applied Statistics*, pages 41–47, 1977.

[13] J. Pearl. Probabilistic reasoning in intelligent systems, 1998.

[14] J. Pearl and Th. S. Verma. A theory of inferred causation. In J.A. Allen, R. Fikes, and E. Sandewall, editors, *KR–91, Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pages 441–452, Cambridge, MA, 1991. Morgan Kaufmann Publishers, Inc., San Mateo, CA.

[15] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.Ch. Sanchez, and M. Müller. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12(1), 2011.

[16] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Springer Verlag, New York, 1993.

[17] M. Stoehr, M. Cypko, K. Denecke, H.U. Lemke, and A. Dietz. A model of the decision-making process: therapy of laryngeal cancer. *International Journal of Computer Assisted Radiology and Surgery*, 9(Suppl 1), 2014.

JAK WIARYGODNA JEST MIARA OCENY MODELU? BOOTSTRAPOWE PRZEDZIAŁY UFNOŚCI DLA MIAR DOKŁADNOŚCI MODELU

Streszczenie Przy testowaniu modelu należy zdawć sobie z tego sprawę, że weryfikacja modelu przy pomocy małego zbioru danych jest mniej przekonywująca niż weryfikacja bazująca na dużym zbiorze danych. Często napotyka się sytuację, w której do analizy modelu dysponujemy nieznaczną ilością rekordów. Nasuwa się pytanie o wiarygodność oceny modelu.

Proponujemy w takiej sytuacji przyjrzeć się bootrstrapowym przedziałom ufności różnych miar dokładności modelu. W tej pracy określamy bootstrapowe przedziały ufności dla krzywych ROC i krzywych kalibracji modeli uzyskanych z danych z repozytorium UC Irvine. Czynność powtarzamy dla modelu skonstruwanego na podstawie wiedzy ekspertów (ponad 300 zmiennych) i testowanego na 66 zebranych rekordach pacjentów. Pokazujemy jak wzrost liczby rekordów wpływa na szerokość bootstrapowych przedziałów ufności oraz jak taka analiza może pomóc w określeniu liczby rekordów, która może podwyższyć rzetelność weryfikacji modelu.

Słowa kluczowe: sieci bayesowskie, bootstrapowe przedziały ufności, walidacja