Probabilistic Networks as Probabilistic Forecasters*

Linda C. van der Gaag and Silja Renooij

Institute of Information and Computing Sciences, Utrecht University
P.O. Box 80.089, 3508 TB Utrecht, The Netherlands
{linda,silja}@cs.uu.nl

Abstract. To establish its clinical value, a probabilistic network is typically subjected to an evaluation study using real patient data from the field of application. The results of such a study are often summarised in the percentage of correctly predicted outcomes. In this paper, we propose the use of a *forecasting score* as an alternative way of expressing the clinical value of a network. Such a score takes not just the predicted outcome into consideration but also the associated distribution of uncertainty. We illustrate the use and interpretation of the Brier forecasting score for a real-life probabilistic network in oncology.

1 Introduction

An increasing number of decision-support systems are being designed that aim at supporting the tasks of medical diagnosis and prognostication. More and more of these systems build upon a probabilistic network for capturing and reasoning about the uncertainties involved in these tasks. A probabilistic network is a concise representation of a joint probability distribution and provides for efficiently computing any probability of interest over its variables [1].

To establish the clinical value of a probabilistic network that is developed for a medical field of application, it is typically subjected to an evaluation study using real patient data. Such a study amounts to entering the data available for each patient into the network, computing the most likely diagnosis or prognosis, and comparing this outcome against a given standard of validity. The percentage of correctly predicted outcomes is then taken to convey the clinical value of the network. For example, a percentage correct of 85% is taken to indicate that the network establishes the correct outcome for 85 out of every 100 patients. A percentage correct cannot be interpreted just like that, however, as it pertains to a specific data collection. Each data collection is likely to include errors, to reflect biases, and to show the effects of random variation. These factors affect the percentage correct for the network under study, yet the percentage does not express the extent to which they do so.

* This research is (partly) supported by the Netherlands Organisation for Scientific Research (NWO).

M. Dojat, E. Keravnou, and P. Barahona (Eds.): AIME 2003, LNAI 2780, pp. 294–298, 2003.

While for computing a network's percentage correct a single outcome per patient is established, the network in essence does not yield a single, deterministic outcome. Instead, it produces a posterior probability distribution for the outcome variable. Since the percentage correct only considers the most likely outcome, it disregards the uncertainty expressed by the posterior distribution. To incorporate this uncertainty in the assessment of a network's clinical value, we propose the use of a *forecasting score* from the field of statistical forecasting. We illustrate the use and interpretation of such a score by means of an evaluation study of a real-life probabilistic network in the field of oesophageal cancer.

The paper is organised as follows. In Sect. 2, we briefly describe the oesophagus network and the available patient data. Sect. 3 presents the results from an evaluation study of the network in terms of its percentage correct. Sect. 4 introduces the Brier score as an alternative way of summarising the results from the study. The paper ends with our concluding observations in Sect. 5.

2 The Oesophagus Network and the Patient Data

With the help of two experts in gastrointestinal oncology from the Netherlands Cancer Institute, Antoni van Leeuwenhoekhuis, we constructed a probabilistic network in the field of oesophageal cancer. The network details the characteristics of an oesophageal tumour and captures the pathophysiological processes associated with its growth. The advance of the cancer is summarised in its *stage*, which can be either I, IIA, IIB, III, IVA, or IVB, in progressive order. The network currently includes 42 statistical variables and almost 1000 (judgmental) probabilities [2], and provides for computing the most likely stage of a patient's cancer based upon his or her symptoms and test results.

For studying the clinical value of the oesophagus network, the medical records of 156 patients diagnosed with oesophageal cancer were available from the Antoni van Leeuwenhoekhuis; these data had not been used in the construction of the network. For each patient between 6 and 21 different symptoms and test results are available. Also recorded is the stage of the patient's cancer as established by the attending physician. In our evaluation study, we take these stages for the standard of validity to compare the outcomes of our network against.

3 The Percentage Correct and Its Shortcomings

Using the available patient data, we conducted an evaluation study of the oesophagus network. We entered, for each patient, all symptoms and test results available and computed the most likely stage for the patient's cancer; we then compared this stage against the one mentioned in the patient's medical record. The results are summarised in the table of Fig. 1, on the left. We find that the network establishes the correct stage for 133 of the 156 patients, that is, we find a percentage correct of 85%.

The numbers of correctly and incorrectly staged patients, as shown in Fig. 1, do not convey any information about the uncertainty in the outcomes computed

		network						*network*					
		I	IIA	IIB	III	IVA	IVB	I	IIA	IIB	III	IVA	IVB
	I	**2**	0	0	0	0	0	**0.21**	–	–	–	–	–
	IIA	0	**37**	0	1	0	0	–	**0.28**	–	1.52	–	–
data	IIB	0	1	**0**	3	0	0	–	1.17	–	0.98	–	–
	III	1	10	0	**36**	0	0	1.40	0.89	–	**0.26**	–	–
	IVA	0	0	0	4	**35**	0	–	–	–	0.75	**0.08**	–
	IVB	0	0	0	3	0	**23**	–	–	–	0.87	–	**0.06**

Fig. 1. Results from the evaluation study: the numbers of correctly and incorrectly staged patients (left) and the average Brier scores (right)

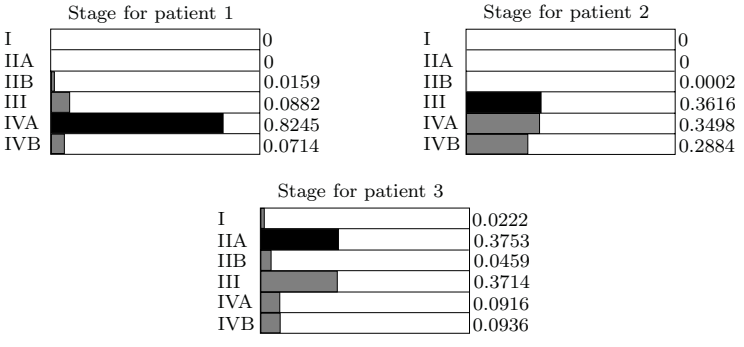

Fig. 2. The posterior distributions over the six possible stages for three patients; the medical records state stage IVA for patient 1 and stage III for patients 2 and 3

from the oesophagus network. We recall that the network yields, for each patient, a posterior probability distribution over the possible stages of his or her cancer; as an example, Fig. 2 shows the probability distributions that are yielded for three real patients. Now, such a computed distribution may clearly point to a single most likely stage. The medical record of patient 1, for example, mentions stage IVA for his cancer. Stage IVA is indeed yielded by the network as the most likely stage; moreover, it is predicted with high probability, indicating that there is little doubt as to the true stage of this patient's cancer. The computed posterior distribution, however, may also reveal considerable uncertainty. The medical record of patient 2, for example, mentions stage III. The network indeed finds III for the most likely stage, but not without considerable doubt: it assigns relatively high probabilities to the stages IVA and IVB as well. For patient 3, the medical record also states stage III, yet the network yields stage IIA. The probability computed for stage III, however, is almost equal to the probability of stage IIA. The percentage correct reported for the network does not express these distributions of uncertainty over the various different stages. For the patients shown in Fig. 2, the network's predictions are classified simply as correct for the first two patients and as incorrect for patient 3.

4 The Forecasting Score

As illustrated in the previous section, the percentage correct as a summary of evaluation results does not take the uncertainties of a network's predictions into account. We feel that for assessing the clinical value of a real-life probabilistic network, not just the most likely outcome but also the posterior distribution over all possible outcomes should be studied. To this end, we observe that probabilistic networks in essence are probabilistic *forecasters*. For the oesophagus network, for example, the posterior distribution over the six possible stages that is computed for a specific patient, can be viewed as a forecast for the true stage of this patient's cancer. An alternative way of establishing the clinical value of a probabilistic network now is to assess its quality as a forecaster.

In the field of statistical forecasting, various different scores for expressing the quality of a probabilistic forecaster have been developed, among which the *Brier score* is the best known [3]. We illustrate the basic idea of this score for our oesophagus network. For each patient i, the network yields a forecast that is composed of the posterior probabilities p_{ij} over the stages $j = \text{I}, \ldots, \text{IVB}$. The Brier score B_i of this forecast is defined as

$$B_i = \sum_{j=\text{I},\ldots,\text{IVB}} (p_{ij} - s_{ij})^2$$

where $s_{ij} = 1$ if the medical record of patient i states stage j, and $s_{ij} = 0$ otherwise. If the network would yield the correct stage with certainty, then the associated Brier score would be equal to 0; for an incorrect deterministic forecast, the score would be 2. The Brier score thus ranges between 0 and 2, and the better the forecast, the lower the score.

The Brier scores of the forecasts for the three patients from Fig. 2 are $B_1 = 0.04$, $B_2 = 0.61$, and $B_3 = 0.56$, respectively. These scores reveal that the forecast for patient 1 is of high quality. The forecasts for patients 2 and 3, on the other hand, appear to be of lesser quality. We recall that the forecast for patient 3 is equivocal as a result of two stages being almost equally likely. For patient 2, there is even more uncertainty in the forecast, as there are three almost equally likely stages. These observations are reflected in the associated Brier scores: the score for patient 3 indicates higher quality than the score for patient 2. While, in terms of the numbers of correctly and incorrectly staged patients, the forecast for patient 2 is correct and the forecast for patient 3 is incorrect, the use of the Brier score results in a more balanced quality assessment.

Now, to assess the quality of the oesophagus network as a probabilistic forecaster, we once again conducted an evaluation study using the available patient data. We entered, for each patient, all symptoms and test results available and computed the posterior probability distribution over the possible stages of the patient's cancer; we then computed the Brier score of the resulting forecast, given the stage mentioned in the patient's medical record. The table of Fig. 1 summarises, on the right, the averaged Brier scores. The low scores on the diagonal signify that the associated forecasts are of high quality. The higher scores beside the diagonal indicate forecasts of lesser quality. The relatively poor quality of

these forecasts may have its origin in uncertainty as to which stage is the true one, as for example for the patients 2 and 3 discussed above. A higher score can also result, however, from a forecast that associates a high probability with an incorrect stage and may thus point to a possible modelling error in the network.

The quality of a real-life probabilistic network can now be expressed in an overall score that averages the scores of the separate forecasts yielded for a given collection of patients. For the oesophagus network, we find an overall Brier score of 0.29 for the available patient data. To interpret this number, we compare it against the overall scores found for three more or less uninformed forecasters. The first of these forecasters does not use any domain knowledge: for each patient, it simply returns a uniform probability distribution over the six possible stages. This forecaster has an overall Brier score of 0.83. The second forecaster yields, for each patient, the prior distribution over the possible stages computed from the network. This forecaster has an overall Brier score of 0.80 and is therefore slightly more informed than the uniform forecaster. The third forecaster, to conclude, yields, for each patient, the prior distribution over the stages recorded in the data collection. This forecaster has an overall Brier score of 0.76, which is slightly lower than the overall score of the second forecaster as a result of its bias towards the data. The much lower Brier score of the oesophagus network now conveys that the network builds upon its knowledge of oesophageal cancer to arrive at relatively good forecasts.

5 Conclusions

The clinical value of a probabilistic network that is developed for a medical application, is typically established by subjecting it to an evaluation study using real patient data. We argued that the percentage correct that is generally computed from such a study, hides the distribution of uncertainties over the possible outcomes and consequently hides the network's doubt as to the true outcome. We suggested the use of a forecasting score to yield a more balanced value assessment for a probabilistic network. We showed that such a score takes not just the most likely outcome but all possible outcomes with their associated uncertainties into consideration and thereby provides useful information in addition to the percentage correct.

References

1. F.V. Jensen (1996). *An Introduction to Bayesian Networks*. UCL Press, London.
2. L.C. van der Gaag, S. Renooij, C.L.M. Witteman, B.M.P. Aleman, and B.G. Taal (2002). Probabilities for a probabilistic network: A case-study in oesophageal cancer. *Artificial Intelligence in Medicine*, vol. 25, pp. 123 – 148.
3. H.A. Panofsky and G.W. Brier (1968). *Some Applications of Statistics to Meteorology*. The Pennsylvania State University, University Park, Pennsylvania.