Evaluation Scores for Probabilistic Networks

Linda C. van der Gaag 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 practical value, a probabilistic network is typically subjected to an evaluation study using real-life data from the domain of application. The results of such a study are often summarised in the percentage of correctly computed outcomes. In this paper, we propose the use of evaluation *scores* as an alternative way of expressing the practical value of a network. These scores take not just the computed outcome into consideration but also the associated distribution of uncertainty. We illustrate the use of such a score for a real-life probabilistic network in oncology and show that it can provide valuable information in addition to a percentage correct.

1 Introduction

In various domains of application, ranging from medicine to meteorology, knowledge-based systems are being developed that build upon a probabilistic network for their knowledge representation. A probabilistic network is a mathematical model comprised of a graphical structure and an associated set of conditional probability distributions [1]. The structure of the network models the statistical variables that are relevant in the domain of application and the influential relationships between them; the probability distributions describe the strengths of the relationships between the variables. A probabilistic network in essence is a concise representation of a joint probability distribution and provides for computing any probability of interest over its variables.

To establish its practical value, a real-life probabilistic network is typically subjected to an evaluation study using data from the domain of application. Such a study amounts to entering the data available for each problem case into the network and computing the most likely outcome. This outcome is then compared against a given standard of validity. The results of the study are often summarised in the percentage of correctly computed outcomes. This *percentage correct* is generally taken to convey the practical value of the network. For a medical diagnostic application, for example, a percentage correct of 85% is taken to indicate that the network is likely to establish the correct diagnosis for 85 out of every 100 patients.

For many applications, this percentage would convey the information that the network performs quite satisfactorily.

Unfortunately, interpretation of the percentage correct of a probabilistic network is not as straightforward as is often suggested. The percentage should be interpreted with respect to a specific data collection. Now, each data collection is likely to include errors and to reflect the biases exhibited by the experts who collected the data. Moreover, the data will include the effects of random variation, especially in domains of a scientific nature. In the medical domain, for example, there is random variation in patient data, arising from biological differences between patients in the progression of pathological processes and from differences in the physicians' interpretation of symptoms and signs [2]. When two outcomes are almost equally likely for a patient, chance determines, to at least some extent, which outcome is entered into the patient's medical record as the most likely one. Random variation may thus affect a network's percentage correct, but the extent to which it does so is not expressed by the percentage.

Probabilistic networks in essence do not yield a deterministic outcome. Rather, they produce a posterior probability distribution for their outcome variable. This distribution reflects the network's doubt as to the most likely outcome. Since the percentage correct of a probabilistic network does not take the computed posterior distribution into consideration, it does not reveal the extent of uncertainty in the outcome. To incorporate the network's doubt in the assessment of its practical value, we propose the use of evaluation *scores* from the field of statistical forecasting. We illustrate the use of such a score by means of an evaluation study of a real-life probabilistic network for the staging of oesophageal cancer. We show that the score computed for the network provides valuable information in addition to its percentage correct.

The paper is organised as follows. In Section 2, we briefly describe the oesophagus network for the staging of oesophageal cancer and the available patient data. In Section 3, we present the results from an evaluation study of the oesophagus network, expressed in terms of a percentage correct; we illustrate the shortcomings of this percentage for expressing the practical value of the network. In Section 4, we introduce the concept of score for summarising evaluation results; we illustrate the use of the Brier score, more specifically, for our network. The paper ends with our concluding observations in Section 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 have constructed a probabilistic network for the staging of oesophageal cancer. As in this paper we use the oesophagus network to illustrate the concept of evaluation score, we briefly describe this network and the patient data that we have available for evaluation purposes.

2.1 The oesophagus network

We have captured the state-of-the-art knowledge about oesophageal cancer in the *oesophagus network*. This network currently includes 42 variables, for which almost 1000 conditional probabilities have been specified. Figure 1 shows the graphical structure of the network, along with the prior probability distributions for the various variables. We briefly review some background knowledge of the domain; for further information on the network and its construction, we refer to [3].

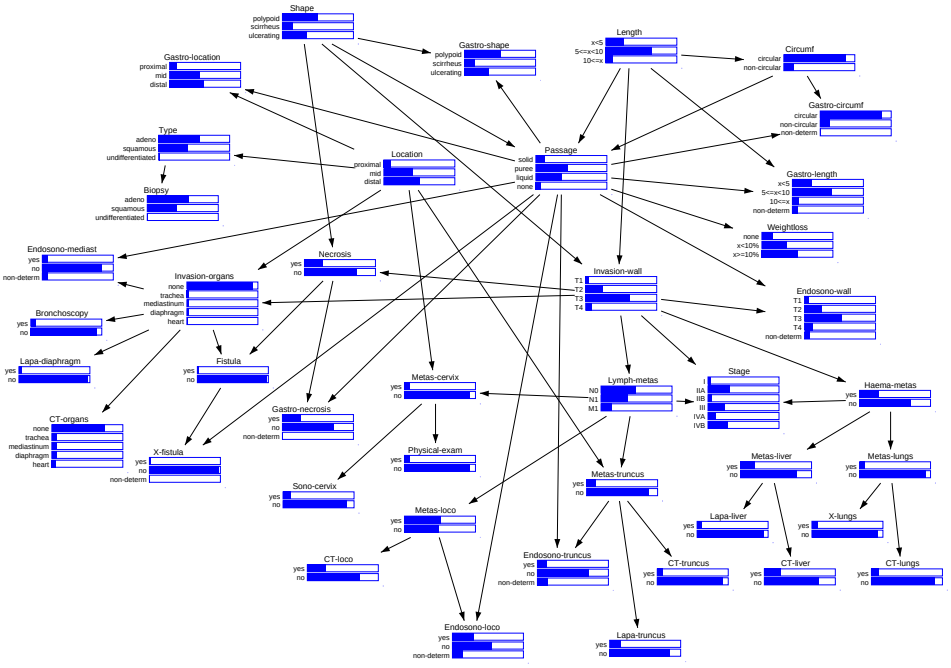

Figure 1: The oesophagus network.

As a consequence of a lesion of the oesophageal wall, for example as a result of frequent reflux or associated with smoking and drinking habits, a tumour may develop in a patient's oesophagus. An oesophageal tumour has various presentation characteristics that influence its prospective growth. These characteristics include the location of the tumour in the oesophagus and its histological type, length, and macroscopic shape. An oesophageal tumour typically invades the oesophageal wall and upon further growth may invade such adjacent organs as the trachea or the diaphragm, dependent upon its location in the oesophagus. In time, the tumour may give rise to lymphatic metastases in distant lymph nodes and to haematogenous metastases in the lungs and the liver. The depth of invasion and extent of metastasis, summarised in the tumour's *stage*, largely influence a patient's life expectancy and are indicative of the effects and complications to be expected from the different available therapeutic alternatives. To establish these factors in a patient, typically a number of diagnostic tests are performed.

The oesophagus network allows for computing any probability of interest over its variables. More specifically, the posterior probabilities for the various possible stages of a patient's cancer can be established by entering his or her symptoms and test results into the network and computing the effect of these data on the probability distribution for the variable that models the cancer's stage.

2.2 The patient data

For studying the ability of the oesophagus network to correctly predict the stage of a patient's cancer, the medical records of 156 patients diagnosed with oesophageal cancer are available from the Antoni van Leeuwenhoekhuis in the Netherlands. We would like to note that these data have not been used in the construction of the oesophagus network. For each patient, various data are available, such as the patient's ability to swallow food and the results from an endosonographic examination of his or her oesophagus. For all patients, fewer than the maximum number of 25 symptoms and test results are available. The number of data ranges between 6 and 21 per patient, with an average of 14.8. For each patient, also the stage of his or her cancer, as established by the attending physician, is recorded. This stage can be either I, IIA, IIB, III, IVA, or IVB, in the order of advanced disease.

3 The percentage correct and its shortcomings

To establish its practical value, a probabilistic network is typically subjected to an evaluation study using data from real-life problem cases in the domain of application. For each case, the outcome with highest posterior probability is determined from the network and compared against a given standard of validity. The results of the study are summarised in the percentage of cases for which the network yields the correct outcome as the most likely one. Using the data described in the previous section, we have conducted such a standard evaluation study of the oesophagus network. We have entered, for each patient, all diagnostic symptoms and test results available and computed the most likely stage for the patient's cancer. We have then compared the computed stage with the stage recorded in the data. The results from this comparison are shown in the matrix of Figure 2. The numbers on the diagonal of the matrix are the numbers of patients per stage for whom the network yields the same stage as the one recorded in the data. Taking the stages from the medical records for our standard of validity, we find that the network establishes the correct stage for 133 of the 156 patients, that is, the network has a percentage correct of 85%.

As probabilistic networks in general, the oesophagus network in essence does not produce a deterministic outcome. Rather, it yields a probability distribution for its outcome variable. More specifically, it yields, for each patient, a posterior probability distribution over the six possible stages of his or her cancer. As an example, Figure 3 shows the distributions that are computed for three different real patients. Now, for some patients the computed posterior distribution clearly points to a single most likely stage. The medical record of patient 1, for example,

		network						
		I	IIA	IIB	III	IVA	IVB	*total*
data	I	**2**	0	0	0	0	0	2
	IIA	0	**37**	0	1	0	0	38
	IIB	0	1	**0**	3	0	0	4
	III	1	10	0	**36**	0	0	47
	IVA	0	0	0	4	**35**	0	39
	IVB	0	0	0	3	0	**23**	26
	total	3	48	0	47	35	23	156

Figure 2: The results from the evaluation study, expressed in terms of the numbers of correctly and incorrectly staged patients.

mentions IVA for the stage of this patient's cancer. Stage IVA is indeed yielded by the network as the most likely one. Moreover, the stage has associated a high probability, indicating that there is not much doubt as to the true stage of this patient's cancer. For other patients, however, the posterior distribution can reveal considerable uncertainty. The medical record of patient 2, for example, mentions stage III. Although the network yields this stage with highest probability, it computes relatively high probabilities for the stages IVA and IVB as well: the stage yielded by the network as the most likely one apparently is not unequivocal. A similar observation pertains to the posterior distribution computed for patient 3, for whose cancer the medical record also states stage III. Unfortunately, the network finds stage IIA for the most likely stage, but not without considerable doubt: the probability computed for stage III is almost the same as the probability of stage IIA. It is not unlikely, therefore, that the incorrect conclusion of the network can be attributed to the effect of random variation, rather than to, for example, a modelling error. In the percentage correct reported for the network, however, the distribution of uncertainty over the various different stages is not taken into account. For the patients shown in Figure 3, the network's outcome is classified simply as correct for the first two patients and as incorrect for patient 3.

patient 1, stage IVA	
stage I	0
stage IIA	0
stage IIB	0.0159
stage III	0.0882
stage IVA	0.8245
stage IVB	0.0714

(a)

patient 2, stage III	
stage I	0
stage IIA	0
stage IIB	0.0002
stage III	0.3616
stage IVA	0.3498
stage IVB	0.2884

(b)

patient 3, stage III	
stage I	0.0222
stage IIA	0.3753
stage IIB	0.0459
stage III	0.3714
stage IVA	0.0916
stage IVB	0.0936

(c)

Figure 3: The posterior probabilities of the six stages for three different patients.

4 The evaluation score

As mentioned in the previous section, probabilistic networks typically yield a probability distribution for their outcome variable. While for some cases from the domain of application the computed posterior distribution will point to a single most likely outcome, it may reveal considerable uncertainty for other cases. The percentage correct as a summary of evaluation results does not take these uncertainties into account. We feel that for assessing the practical value of a network, however, not just the most likely outcome but also the posterior distribution over the various possible outcomes should be taken into consideration.

To arrive at an alternative way of expressing their practical value, we observe that probabilistic networks basically are probabilistic *forecasters*, as they repeatedly present predictions for an outcome variable in terms of probabilities. For the oesophagus network, for example, the posterior probability distribution that is computed for a specific patient can be looked upon as a forecast for the stage of this patient's cancer. Establishing the practical value of a probabilistic network now amounts to assessing its quality as a forecaster. The quality of a probabilistic forecaster is often expressed in terms of its *calibration*, that is, the degree to which its forecasts match the true distribution of outcomes. For the oesophagus network, more specifically, we say that it is (empirically) *well calibrated* if, among the patients for whom the network predicts a specific stage S with probability x_S, the proportion of patients who in fact have stage S, denoted x'_S, equals x_S. The smaller the difference between x_S and x'_S, that is, the closer the network's distribution matches the true distribution, the better calibrated the network is [4, 5]. Building upon this concept of calibration, various scores for expressing the quality of a forecaster have been developed in the field of statistics.

Among the best-known evaluation scores is the *Brier score* [6]. We illustrate the basic idea of this score for our oesophagus network. For each patient i, the network yields a forecast of posterior probabilities p_{ij} over the stages $j = \mathrm{I}, \ldots,$ IVB. The Brier score B_i of this forecast is defined as

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

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, that is, if the network would yield a correct deterministic forecast for the patient, then the associated Brier score would be equal to 0. If the network would yield an incorrect deterministic forecast, the score would be 2. For the forecast for a single patient, therefore, the Brier score ranges between 0 and 2, and the better the forecast, the lower the score. The Brier scores for the network's forecasts for the three patients from Figure 3 now are:

$$B_1 = 0.04 \qquad B_2 = 0.61 \qquad B_3 = 0.56$$

These scores reveal that the quality of the forecast for the first patient is very good, as expected. The other scores show that the forecasts for the patients 2 and

3 are of less quality. We recall that for patient 3 the forecast is unequivocal as a result of two stages being almost equally likely, among which is the correct stage. 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 B_3 for patient 3 indicates higher quality than the score B_2 for the second patient. 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 Brier score results in a more balanced and, hence, a more insightful quality assessment. Figure 4 summarises the Brier scores averaged over all, correctly and incorrectly staged, patients.

		network					
		I	IIA	IIB	III	IVA	IVB
data	I	**0.21**	–	–	–	–	–
	IIA	–	**0.28**	–	1.52	–	–
	IIB	–	1.17	–	0.98	–	–
	III	1.40	0.89	–	**0.26**	–	–
	IVA	–	–	–	0.75	**0.08**	–
	IVB	–	–	–	0.87	–	**0.06**

Figure 4: The results from the evaluation study, expressed in terms of average Brier scores.

The quality of the oesophagus network as a forecaster can now be expressed in an overall score that is computed from the scores of the separate forecasts for our collection of patients. For n patients, the overall Brier score B is defined as

$$B = \frac{1}{n} \sum_{i=1,\ldots,n} B_i$$

It is readily seen that the overall Brier score again ranges between 0 and 2, and the better the forecaster, the lower the score. For the oesophagus network, an overall Brier score of 0.29 is found. To interpret this number, we compare the score with the overall scores obtained for two *uninformed* forecasters. The first forecaster gives a uniform probability distribution for each patient; this forecaster has an overall Brier score of 0.83. The second forecaster gives for each patient the prior distribution over the stages recorded in the data; this forecaster has an overall Brier score of 0.76 and is therefore slightly more informed than the uniform forecaster. The much lower Brier score of the oesophagus network now conveys the information that the network is quite informed and indeed builds upon its knowledge of oesophageal cancer to arrive at relatively good forecasts.

5 Conclusions

To establish the practical value of a probabilistic network, it is typically subjected to an evaluation study that amounts to entering data from real-life cases from the domain of application into the network, computing the most likely outcome, and comparing it against a given standard of validity. The evaluation results are often summarised in a percentage correct, which is then taken to convey the network's practical value. We have argued that such a percentage hides the distribution of uncertainties over the values of the outcome variable and consequently hides the network's doubt as to the most likely outcome. We have suggested the use of an evaluation score to yield a more balanced value assessment for a network. Such a score takes not just the most likely outcome but all possible outcomes with their associated uncertainties into consideration. We feel that an evaluation score provides useful information about the practical value of a probabilistic network in addition to a percentage correct.

Acknowledgements. This research has been (partly) supported by the Netherlands Computer Science Research Foundation with financial support from the Netherlands Organisation for Scientific Research (NWO). We are most grateful to Babs Taal and Berthe Aleman from the Netherlands Cancer Institute, Antoni van Leeuwenhoekhuis, who spent much time and effort in the construction of the oesophagus network.

References

[1] F.V. Jensen (1996). *An Introduction to Bayesian Networks*. UCL Press, London.

[2] R.H. Fletcher, S.W. Fletcher, and E.H. Wagner (1996), *Clinical Epidemiology. The Essentials*, 3rd edition. Baltimore: Williams & Wilkins.

[3] L.C. van der Gaag, S. Renooij, C.L.M. Witteman, B.M.P. Aleman, and B.G. Taal (2001). Probabilities for a probabilistic network: A case-study in oesophageal cancer. *Artificial Intelligence in Medicine*, to appear.

[4] A.P. Dawid (1985). Calibration-based empirical probability. *Annals of Statistics*, vol. 13, pp. 1251 – 1274.

[5] M.H. DeGroot and S.E. Fienberg (1983). The comparison and evaluation of forecasters. *The Statistician*, vol. 32, pp. 12 – 22.

[6] H.A. Panofsky and G.W. Brier (1968). *Some Applications of Statistics to Meteorology*. The Pennsylvania State University, University Park, Pennsylvania.