Forecast Verification and the Uncertain Truth

Silja Renooij

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

Abstract

Forecasting scores, such as the Brier score, are used to assess the quality of a forecaster. It was demonstrated previously that the Brier score can be successfully applied to establish the forecasting quality of a Bayesian network. Existing forecasting scores, including the Brier score, assume that the eventual outcome of a predicted event can be observed with certainty. In practice, however, there is often uncertainty as to what the truth is. We propose a new score which takes this additional uncertainty into account.

1 Introduction

A Bayesian network is a compact representation of a joint probability distribution over a set of statistical variables [2], which enables efficient computation of prior and posterior distributions for any variable of interest. In various domains, the results of such network computations can be regarded as *probabilistic forecasts* for the variable of interest. Bayesian networks can thus be considered probabilistic forecasters and *forecast verification methods* can be used to assess the quality of their forecasts. In fact, in the context of Bayesian networks the use of the *Brier score* was previously illustrated [1].

Measures for assessing forecasting quality exist for both deterministic and probabilistic forecasts, both for discrete (binary and multi-valued) and for continuous events. The quality of probabilistic forecasts for discrete events is usually measured by a *scoring rule*, such as the Brier score [3, 4]. Existing scoring rules assume that a forecast for an event is followed by a subsequent observation of the true outcome of the event. In practice, however, this true outcome generally results from observations which may be erroneous or an uncertain indication of the truth. Consider, for example, a prognosis of the progression of a patient's disease, which can only be monitored using tests and not truly verified until after the patient's death; the results from these tests are typically uncertain. Current scoring rules ignore such uncertainty, under the assumption that the uncertainty in the observation is much smaller than the expected uncertainty in the forecast. To the best of our knowledge, no forecast verification methods exist that take uncertainty about the truth of the observation into account.

In this paper we present a preliminary definition of a scoring rule for forecasters that probabilistically predict events that are not observable without some uncertainty. The new scoring rule is a generalisation of the Brier score. In Section 2 we review the Brier score, followed by the introduction of our new score in Section 3. In Section 4 we compare our score to the Brier score; we present some conclusions and directions for further research in Section 5.

2 The Brier score

In the remainder of this paper we will use the term *forecaster* to refer to a Bayesian network, a human forecaster, or any other system that produces a *discrete probability distribution* for some variable of interest.

Consider a forecaster which provides $m \geq 1$ probability distributions, or forecasts, for a variable C with $n \geq 2$ possible outcomes $c_1, \ldots c_n$. Let $f_k(C)$ represent the probability distribution of the kth forecast. In forecast verification, it is assumed that each forecast is eventually followed by an observation of the true outcome for the variable of interest. The true outcome of C following forecast k is captured by an *outcome indicator* $s_k(i)$, $i = 1, \ldots, n$, such that

$$s_k(i) = \begin{cases} 1 & \iff \quad c_i \text{ is the true outcome corresponding to forecast } k \\ 0 & \iff \quad c_i \text{ is \textit{not} the true outcome for forecast } k \end{cases}$$

The *Brier score* B now is the average value of the squared error between forecast and outcome for variable C over m sequential forecasts [3]

$$B = \frac{1}{m} \sum_{k=1}^{m} \sum_{i=1}^{n} \Big(f_k(c_i) - s_k(i) \Big)^2$$

The lowest score is obtained iff, for all i and k, $f_k(c_i) = s_k(i)$; that is, the best score $B^\top$ is obtained for a correct deterministic forecast $B^\top = (1-1)^2 + (n-1) \cdot (0-0)^2 = 0$. The worst score $B^\perp$ results from an incorrect deterministic forecast $B^\perp = (1-0)^2 + (0-1)^2 + (n-2) \cdot (0-0)^2 = 2$. The Brier score is a strictly proper scoring rule, awarding the forecaster for stating what it believes: a *(strictly) proper* scoring rule (uniquely) optimises the expected score if the forecaster states forecasts that match its subjective probabilities for the outcomes of the variable of interest. Not all scoring rules are proper [4].

The Brier score is most easily computed without explicit use of the outcome indicator; let c_t be the observed true outcome of variable C following forecast $f_k(C)$, that is, $s_k(t) = 1$, then the Brier score for the k-th forecast can be computed from the following equivalence

$$\sum_{i=1}^{n} \Big(f_k(c_i) - s_k(i) \Big)^2 = \Big(f_k(c_t) - 1 \Big)^2 + \sum_{i \neq t} \Big(f_k(c_i) - 0 \Big)^2 = 1 - 2 f_k(c_t) + \sum_{i=1}^{n} f_k(c_i)^2$$

Note that the last term can already be computed before knowing the outcome that follows the forecast. From here on we will often consider a only single forecast and drop the subscript k if no confusion is possible.

3 A new score

The Brier score, like all forecasting scores, builds upon the idea that the quality of a forecaster can be established from some relation between prediction and true outcome. This true outcome generally results from observations, which may be erroneous or an uncertain indication of the truth. Current scoring rules ignore such uncertainty. In this section we propose a new score that also allows for taking uncertainty about the truth into account. We identify the benefits and drawbacks and compare results from both this new score and the Brier score.

3.1 Forming a definition

In this section, we work our way towards a definition for a new scoring rule. We take the Brier score as a point of departure, because the Brier score is a (strictly) proper scoring rule with a finite range of possible values; that is, unlike some scores, the Brier score never assigns an infinite penalty to an incorrect deterministic forecast.

As an example, consider a forecaster who presents a forecast $f(c_1), f(c_2)$ for a binary variable C with values c_1 and c_2. Assume that the true outcome of variable C can only be observed indirectly and that the possible values of these indirect observations are captured by variable O with w values $o_1, \ldots, o_w$. Let the set of conditional probability distributions $\Pr(C \mid O)$ describe the relation between observed and true values of the forecasted variable.

Now assume that the observed outcome following a given forecast is value o_t of variable O. If o_t is an indication of the true value c_1, then the prediction $f(c_1)$ corresponds to a prediction for the true outcome. For the Brier score this would mean that with a probability of $\Pr(c_1 \mid o_t)$ the term $(f(c_1) - 1)^2$ is included in the score. On the other hand, with probability $1 - \Pr(c_1 \mid o_t)$ the same prediction $f(c_1)$ corresponds to a prediction for an incorrect outcome, in which case the term $(f(c_1) - 0)^2$ is included in the Brier score. Similarly, for prediction $f(c_2)$: with probability $\Pr(c_2 \mid o_t)$ this is a prediction for the true outcome, and with probability $1 - \Pr(c_2 \mid o_t)$ it is a prediction of the incorrect outcome. To account for the uncertainty regarding the true outcome following a forecast, we take all four mentioned factors as ingredients of a new score S

$$\begin{aligned}
S &= p_{1t} \cdot \big(f(c_1) - 1\big)^2 + (1 - p_{1t}) \cdot \big(f(c_1) - 0\big)^2 + \\
&\quad + p_{2t} \cdot \big(f(c_2) - 1\big)^2 + (1 - p_{2t}) \cdot \big(f(c_2) - 0\big)^2 \\
&= \sum_{i=1,2} p_{it} \cdot \big(f(c_i) - 1\big)^2 + (1 - p_{it}) \cdot f(c_i)^2 \\
&= \sum_{i=1,2} \big(f(c_i) - p_{it}\big)^2 + p_{it} \cdot (1 - p_{it})
\end{aligned}$$

where p_{it} represents the probability $\Pr(c_i \mid o_t)$, $i = 1, 2$. The previous argument lies at the basis of our new score.

3.2 Definition for a single forecast

Again consider the problem of assessing the quality of a forecaster. We assume that a *single* probabilistic forecast is presented for a variable C with n possible *true* outcomes, $c_1, \ldots c_n$. These true outcomes can only be observed indirectly and the possible values of the indirect observations are captured by variable O. In the previous section, we assumed that variable O had w values; here we argue that most often we will have that $w = n$: if variable O has more than n values, then some of them should be grouped in order to be able to compare forecasts and outcomes; if variable O has less than n values, then some of the categories of C should be grouped in order to be able to compare a forecast and an outcome. Here we indeed assume that $w = n$ and that each o_i is an (uncertain) indication of c_i, $i = 1, \ldots, n$.

To establish the quality of a forecaster which forecasts a distribution for variable C, where only values of variable O can be observed, we require a relationship between observation and truth. This relationship is modelled by a set of conditional probability distributions $\Pr(C \mid O)$, where we write p_{ij} to denote the probability $\Pr(c_i \mid o_j)$. We assume that we have estimates for the $n \cdot (n-1)$ required probabilities available, or are able to compute them from estimates such as $\Pr(O \mid C)$, and $\Pr(C)$ or $\Pr(O)$. We again use an outcome indicator $s(i)$, $i = 1, \ldots, n$, which now captures the *observed* outcome rather than the true outcome:

$$s(i) = \begin{cases} 1 & \iff \quad o_i \text{ is the observed outcome corresponding to the forecast} \\ 0 & \iff \quad o_i \text{ is \textit{not} the observed outcome for the forecast} \end{cases}$$

The new score S for a *single* forecast is now given by the following formula

$$S = \sum_{j=1}^{n} s(j) \cdot \sum_{i=1}^{n} \Big(\big(f(c_i) - p_{ij}\big)^2 + p_{ij} \cdot (1 - p_{ij}) \Big)$$

This score can be also be computed without explicit use of the outcome indicator: let o_t be the observed value of variable O corresponding to the forecast under consideration, that is, $s(t) = 1$, then the new score can be computed using the equivalence

$$\begin{aligned} S &= \sum_{j=1}^{n} s(j) \cdot \sum_{i=1}^{n} \Big(\big(f(c_i) - p_{ij}\big)^2 + p_{ij} \cdot (1 - p_{ij}) \Big) \\ &= s(t) \cdot \sum_{i=1}^{n} \Big(\big(f(c_i) - p_{it}\big)^2 + p_{it} \cdot (1 - p_{it}) \Big) + \\ &\quad + \sum_{j \neq t} s(j) \cdot \sum_{i=1}^{n} \Big(\big(f(c_i) - p_{ij}\big)^2 + p_{ij} \cdot (1 - p_{ij}) \Big) \\ &= 1 \cdot \sum_{i=1}^{n} \Big(\big(f(c_i) - p_{it}\big)^2 + p_{it} \cdot (1 - p_{it}) \Big) + 0 \end{aligned}$$

Note that for a binary variable C ($n = 2$), this is exactly the formula established in the previous section.

3.3 Properties of the single-forecast score

We argued earlier that the Brier score ranges from a best value of zero to a worst value of two. The range of the new score depends on the values of $\Pr(C \mid O)$ and should be normalised in order to facilitate the comparison of scores for different forecasts and forecasters. Recall that for observed outcome o_t of variable O, the new score for a single forecast can be written as

$$S = \sum_{i=1}^{n} \big(f(c_i) - p_{it}\big)^2 + \sum_{i=1}^{n} p_{it} \cdot (1 - p_{it})$$

where the last summation represents a constant not influenced by the forecast. To receive a good score, a forecaster should therefore try to minimise, for each i, the term $\big(f(c_i) - p_{it}\big)^2$ in such a way that it is ensured that $\sum_{i=1}^{n} f(c_i) = 1$. It is obvious that these terms are minimised whenever $f(c_i) = p_{it}$ (note that $\sum_{i=1}^{n} p_{it} = 1$). The best value $S^{\top}$ of the new score is therefore

$$S^{\top} = \sum_{i=1}^{n} p_{it} \cdot (1 - p_{it}) = 1 - \sum_{i=1}^{n} {p_{it}}^2$$

From the fact that $\sum_{i=1}^{n} {p_{it}}^2 \leq \big(\sum_{i=1}^{n} p_{it}\big)^2 = 1$ (all p_{it}s are non-negative), we have that $S^{\top}$ is always in the interval $[0, 1\rangle$. Note that if the forecaster predicts a probability of 1 for the value c_t of C that is implied by the observed outcome o_t of O, then we get the following (higher = worse) score

$$S = \big(1 - p_{tt}\big)^2 + \sum_{i \neq t} \big(0 - p_{it}\big)^2 + \sum_{i=1}^{n} p_{it} \cdot (1 - p_{it}) = 2 \cdot (1 - p_{tt})$$

A bad score is obtained for a forecaster who maximises

$$\sum_{i}^{n} \big(f(c_i) - p_{it}\big)^2 = \sum_{i}^{n} f(c_i)^2 - 2 \cdot \sum_{i}^{n} f(c_i) \cdot p_{it} + \sum_{i} {p_{it}}^2$$

the last term of which the forecaster has no control over. The first term is maximised if $\sum_i f(c_i)^2 = 1$, which is only achieved by assigning a probability of 1 to a single c_i and zero to all other c_j, $j \neq i$. The second term should be minimised, which is achieved by assigning a probability of 1 to that value c_r of C for which $\Pr(c_r \mid o_t)$ has the smallest value in the $\Pr(C \mid o_t)$ distribution, and zero to all other c_j, $j \neq r$. The worst possible score $S^{\perp}$ then corresponds to a deterministic forecast and becomes

$$S^{\perp} \; = \; \big(1 - p_{rt}\big)^2 + \sum_{i \neq r} \big(0 - p_{it}\big)^2 + \sum_{i=1}^{n} p_{it} \cdot (1 - p_{it}) \; = \; 2 \cdot (1 - p_{rt}) \leq 2$$

The new score is not a proper scoring rule, since the best expected score is obtained using a strategy in which the forecaster presents forecasts that match $\Pr(C \mid o_t)$ rather than its (subjective) beliefs concerning the outcomes of the variable of interest. This can be considered a drawback of the score, but a number of arguments can be given to support the score. First of all, the use of a non-proper scoring rule does not imply that a forecaster can easily obtain a good score. For one thing, the forecaster obviously does not know in advance what the observed outcome will be, and therefore will not know which $\Pr(C \mid o_i)$ to predict. In addition, the forecaster may not be aware that its forecasts are verified, and if it is, may not know which scoring rule is used; without this knowledge the forecaster cannot choose a clever strategy with the sole purpose of getting a good score. Especially non-human forecasters will be unaware of verification and the methods used. Secondly, although the new score punishes forecasters for predictions more extreme than the $\Pr(C \mid o_t)$ distribution, the uncertainty inherent in the observations allows us no way to verify if the forecaster is right in being so sure of itself.

3.4 Definition for multiple forecasts

In the previous section we demonstrated that the range of the new score for a single forecast depends on the distribution $\Pr(C \mid o_t)$ and thus on the outcome corresponding to the forecast. Therefore, the range of the score may differ from forecast to forecast. In order to compare different forecasts and forecasters and, especially, to compute an average score over m forecasts, we normalise for each forecast k the single-forecast score S_k to the score $\hat{S}_k$. Let o_t^k denote the observed value of variable O corresponding to forecast k, that is, $s_k(t) = 1$; let $p_{it}^k = \Pr(c_i \mid o_t^k)$, and let $S_k^\top$ and $S_k^\perp$ denote the best and worst values, respectively, of the score S_k for forecast k, then

$$\hat{S}_k = 2 \cdot \frac{S_k - S_k^\top}{S_k^\perp - S_k^\top} = 2 \cdot \frac{\sum_i \left(f_k(c_i) - p_{it}^k\right)^2}{1 - 2p_{rt}^k + \sum_i \left(p_{it}^k\right)^2}$$

where $p_{rt}^k = \min_i\{\Pr(c_i \mid o_t^k)\}$. Note that we assume all forecasted events to be independent; for the k-th forecast, $1 < k \leq m$, however, we can also consider $f_k(C \mid o_t^1 \ldots o_t^{k-1})$ and $\Pr(C \mid o_t^1 \ldots o_t^k)$ instead. The new score for $m \geq 1$ forecasts now ranges from zero (best value) to two and is given by

$$S = \frac{1}{m} \sum_{k=1}^{m} \hat{S}_k$$

4 The new score versus the Brier score

The new score is inspired by the Brier score and is in fact a generalisation of the Brier score for uncertain observations concerning the truth. This implies that if there is no uncertainty in our observations, that is, if $\Pr(c_i \mid o_i) = 1$ for all i and $\Pr(c_i \mid o_j) = 0$ for all $i \neq j$, then our new score is equivalent to the Brier score. To demonstrate this, consider again a single forecast upon which value o_t is observed for variable O. Then,

$$\begin{aligned}
S &= 1 \cdot \sum_{i=1}^{n} \Big(\big(f(c_i) - p_{it}\big)^2 + p_{it} \cdot (1 - p_{it}) \Big) + 0 \\
&= \big(f(c_t) - p_{tt}\big)^2 + p_{tt} \cdot (1 - p_{tt}) + \sum_{i \neq t} \Big(\big(f(c_i) - p_{it}\big)^2 + p_{it} \cdot (1 - p_{it}) \Big) \\
&= \big(f(c_t) - 1\big)^2 + \sum_{i \neq t} \big(f(c_i) - 0\big)^2 = \sum_{i=1}^{n} \big(f(c_i) - s(i)\big)^2
\end{aligned}$$

As the best and worst possible scores are now independent of the forecast under consideration, normalisation is not required and the score for m forecasts reduces to the Brier score over m forecasts as well.

We present a few examples to investigate when the new score and the Brier score differ and in what way. We consider five (different) forecasters F_i, $i = 1 \ldots, 5$, which all provide a forecast for a variable C with two possible true outcomes c_1 and c_2. We assume that outcomes c_1 and c_2 can only be observed with some uncertainty through outcomes o_1 and o_2, respectively. Consider the forecasts and probability distributions shown in Table 1. Note that each of the forecasters predicts c_1 as the more likely value of C and hence expects to observe o_1.

We now compare the Brier score B and the new score S for $m = 1$ forecasts for all forecasters and both outcomes; these scores are shown in Table 1 (right). The Brier score is computed under the assumption that o_1 corresponds to c_1 and o_2 to c_2; the uncertainty in the observations is thereby disregarded. From the table, we can make a number of observations. Firstly, a uniform forecast (F_1) results in equal Brier scores for all outcomes, but not in equal S scores. The scores with the new score are in fact better, because less extreme scores are rewarded given the uncertainty in the observations. For a deterministic forecast (F_5) we observe the opposite: a forecast more extreme than $\Pr(c_1 \mid o_1)$ is considered perfect using the Brier score, but is punished for possible overconfidence using the new score.

The Brier scores given different outcomes almost display the behaviour of communicating vessels: if one goes up, the other goes down. The S scores show a very different pattern: scores given outcome o_1 decrease until the prediction corresponds to $\Pr(c_1 \mid o_1)$ and then increase again; for outcome o_2 the decreasing predictions

$\Pr(C \mid O)$	o_1	o_2
c_1	0.80	0.10
c_2	0.20	0.90
$\sum_i {p_{ij}}^2$	0.68	0.82

	forecast		o_1		o_2	
	$f(c_1)$	$f(c_2)$	B	S	B	S
F_1:	0.50	0.50	0.50	0.28	0.50	0.40
F_2:	0.75	0.25	0.13	0.18	1.13	1.04
F_3:	0.80	0.20	0.08	0.00	1.28	1.21
F_4:	0.90	0.10	0.02	0.03	1.62	1.58
F_5:	1.00	0.00	0.00	0.13	2.00	2.00

Table 1: (left) Distribution $\Pr(C \mid O)$. (right) Forecasts with their Brier score B and new score S, per outcome.

result in increasing scores. In essence, however, both scores basically display the same behaviour, in that the closer the assessment $f(c_i)$ is to $\Pr(c_i \mid o_t)$ (either 1 or 0 for the Brier score) the better the score for outcome o_t is.

5 Conclusions and further research

In this paper we proposed a new scoring rule for assessing the quality of probabilistic forecasters in situations in which there is uncertainty concerning the subsequent observations. The new score requires knowledge of the actual uncertainty in the observations, which is information that is often available. In medical domains, for example, test-characteristics and incidence rates provide the ingredients necessary for establishing the required probability distributions.

The new score is based on the Brier score, but is no longer a proper scoring rule. We argued that this is not necessarily a problem and that, for example for extreme forecasters, it may well be unwished for to create a proper scoring rule that takes the uncertainty in observations into account. Instead of basing a new score on the Brier score, it is also possible to consider other existing scoring rules such as, for example, the logarithmic score $-\log f(c_t)$. Drawbacks of this latter score are, however, that only the prediction of the outcome that subsequently occurs is considered and not the entire forecast, and that a deterministic forecast for the wrong outcome is punished with an infinitely large score.

We do not claim that our new score is the answer to the problem posed, but it is at least a step in the right direction. It is possibly the combination of both Brier score and new score that gives the most insight in the quality of a forecaster. Also, decompositions of the new score, similar to those existing for the Brier score [5], may be useful to this end. Further research is required to establish if the current score can be improved upon.

Acknowledgements This research has been (partly) supported by the Netherlands Organisation for Scientific Research (NWO).

References

[1] L.C. van der Gaag, S. Renooij (2003). Probabilistic networks as probabilistic forecasters. *Proceedings of the Ninth Conference on Artificial Intelligence in Medicine in Europe*, Springer Verlag, Berlin, LNAI 2780, pp. 294 – 298.

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

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

[4] J.Q. Smith (1988). *Decision Analysis, a Bayesian Approach.* Chapman & Hall, New York.

[5] J.F. Yates (1994). Subjective probability accuracy analysis. In: G. Wright, P. Ayton (eds), *Subjective Probability*, John Wiley & Sons.