ADVANCES IN ATMOSPHERIC SCIENCES, VOL. 25, NO. 1, 2008, 83–88

Guidance on the Choice of Threshold for Binary Forecast Modeling

Keon Tae SOHN* and Sun Min PARK

Department of Statistics, Pusan National University, Busan 609-735, *Korea*

(Received 14 June 2006; revised 16 March 2007)

ABSTRACT

This paper proposes useful guidance on the choice of threshold for binary forecasts. In weather forecast systems, the probabilistic forecast cannot be used directly when estimated too smoothly. In this case, the binary forecast, whether a meteorological event will occur or not, is preferable to the probabilistic forecast. A threshold is needed to generate a binary forecast, and the guidance in this paper encompasses the use of skill scores for the choice of threshold according to the forecast pattern. The forecast pattern consists of distribution modes of estimated probabilities, occurrence rates of observations, and variation modes.

This study is performed via Monte-Carlo simulation, with 48 forecast patterns considered. Estimated probabilities are generated by random variate sampling from five distributions separately. Varying the threshold from 0 to 1, binary forecasts are generated by threshold. For the assessment of binary forecast models, a 2×2 contingency table is used and four skill scores (Heidke skill score, hit rate, true skill statistic, and threat score) are compared for each forecast pattern. As a result, guidance on the choice of skill score to find the optimal threshold is proposed.

Key words: binary forecast, Monte-Carlo simulation, threshold, skill score

DOI: 10.1007/s00376-008-0083-8

1. Introduction

There are many binary observations and forecasts in meteorological data. For example, the occurrence of heavy rain, heavy snow, forest fire, hail, drought, strike of a typhoon, weather patterns etc. In weather forecast systems, one possibility is to generate a probabilistic forecast, such as "the probability of precipitation tomorrow is 60%", according to Murphy (1993). However, the probabilistic forecast cannot be used directly when estimated too smoothly. In such cases, the binary (dichotomous) forecast is preferable to the probabilistic forecast. For the binary forecast, a threshold is needed. That is, the forecaster announces that a meteorological event will occur if the estimated probability is greater than a selected threshold. In order to assess the predictability, the results of a binary forecast can be summarized as a 2×2 contingency table, such as Table 1, which consists of observations and forecasts where values are either 0 (not occur) or 1 (occur).

In order to improve the predictability, it is important to choose the optimal threshold, a task for which skill scores may be used. For the objective forecast quality evaluation of forecast models, skill scores have been proposed and applied by many authors (e.g., Heidke, 1926; Woodcock, 1976; Burrows, 1991; Barnston, 1992; Zhang and Casey, 2000; Mcbride and Ebert, 2000; Sohn and Han, 2004; Sohn et al., 2005a,b). For binary forecasts, some examples of approaches used include hit rate, bias score, probability of detection, false alarm rate, true skill statistic, threat score, probability of false detection, equitable threat score, and Heidke skill score. The Heidke skill score is mainly used because it eliminates the effect of the reference forecast [see Hans and Francis (1999) and Sohn and Han (2004) for more detail of skill scores].

Knowing that a larger skill score indicates a better forecast model, the aim is to find the optimal threshold which has the maximal skill score. However, the problem is that the optimal threshold using one skill score is different from the optimal threshold using another. In binary forecasting, the hit case (Table 1) is more serious than the correct negative case for some meteorological events (for instance, binary forecasts of

*Corresponding author: Keon Tae SOHN, ktsohn@pusan.ac.kr

Table 1. 2 × 2 contingency table of a binary forecast.

	Forecast		Total
	Not occur (0)	Occur (1)	
Observation			
Not occur (0)	d (correct negative)	c (false alarm)	$c + d$
Occur (1)	b (miss)	a (hit)	$a + b$
Total	$b + d$	$a + c$	$a + b + c + d$

heavy rain or snow).

In section 2, the motivation behind this study is introduced, with a case study of a binary forecast of heavy snow. The Monte-Carlo simulation scheme is presented in section 3, which is a statistical simulation that uses artificial data generated by random variate sampling from statistical distributions. Forecast patterns which consist of distribution modes of estimated probabilities, occurrence rates of observations, and variation modes are considered. In section 4, some results of the Monte-Carlo simulation are presented. For the assessment of binary forecast models, using a 2 × 2 contingency table, four skill scores (Heidke skill score, hit rate, true skill statistic, and threat score) are calculated and compared for each forecast pattern. Guidance on the choice of skill score to find the optimal threshold is then proposed.

Table 2. 2 × 2 contingency table for threshold=0.5.

	Forecast		Total
	0	1	
Observation			
0	4733 (99.75%)	12 (0.25%)	4745
1	61 (81.33%)	14 (18.67%)	75
Total	4794	26	4820

Table 3. 2 × 2 contingency table for threshold=0.23.

	Forecast		Total
	0	1	
Observation			
0	4700 (99.05%)	45 (0.95%)	4745
1	34 (45.33%)	14 (54.67%)	75
Total	4734	86	4820

Table 4. 2 × 2 contingency table for threshold=0.02.

	Forecast		Total
	0	1	
Observation			
0	4324 (91.13%)	421 (8.87%)	4745
1	4 (5.33%)	71 (94.67%)	75
Total	4328	492	4820

2. Case study: heavy snow forecast

Sohn (2006) applied the logistic regression model to forecast the occurrence of heavy snow in the Honam area, Korea. Observations of daily snow cover and the numerical model outputs for synoptic factors during the cold season from 2002–2005 were used for the statistical modeling. As a result, the distribution of estimated probabilities, which were generated from the estimated logistic regression model, was too smooth because of underestimation. Therefore, the estimated probability of heavy rain could not be used directly for the weather forecast, and the need to use a threshold was apparent.

It seems reasonable to take 0.5 as the threshold. That is, heavy snow will occur if the estimated probability is greater than 0.5, and vice versa. In this case, the hit rate (Table 2) is only 18.67%.

Considering the Heidke skill score, the threshold 0.23 produces the maximum (Table 3), with a hit rate of 54.67%. Though this is much improved, it is still worth considering another forecast strategy. Varying the threshold 0 to 1, 2 × 2 tables were produced for all thresholds and compared. Following the forecasters' opinions, 0.02 was selected as an optimal threshold. The result is summarized in Table 4 and the hit rate is 94.67%.

Does the maximal Heidke skill score indicate the best binary forecast model? This case study suggests not. Therefore, what is the most appropriate skill score for a given forecast pattern? How can a skill score be chosen objectively? These questions are the motivation behind this study, with an aim to propose useful guidance on the choice of an appropriate skill score for finding the optimal threshold.

3. Monte-Carlo simulation scheme

This study is performed using Monte-Carlo simulation, which uses artificial data generated by random variate sampling from some statistical distributions.

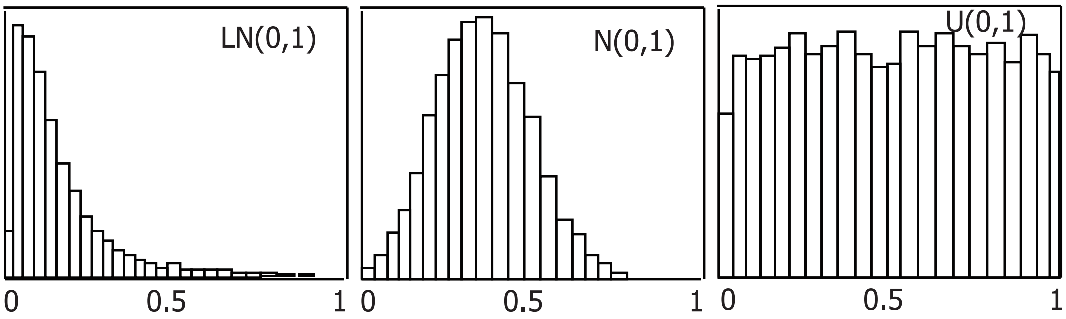

Fig. 1. Histograms of 10,000 probabilities from LN(0,1), N(0,1) and U(0,1).

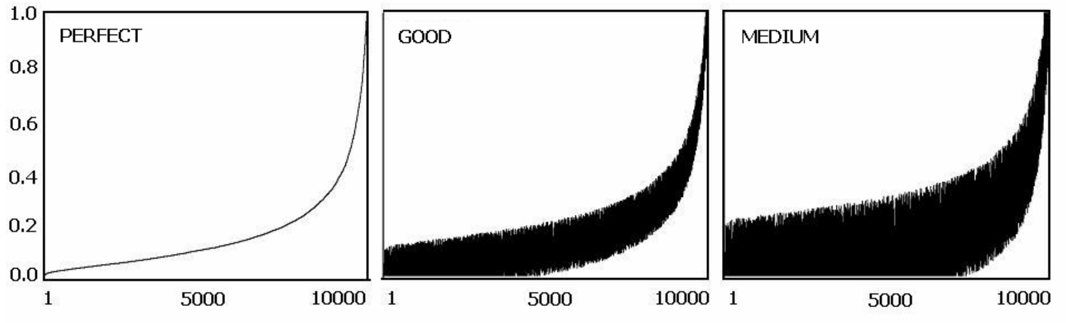

Fig. 2. Scatter diagrams of (sorted estimated probabilities + variation): perfect forecast case (left); good forecast case (center); and medium forecast case (right).

3.1 *Skill scores for the binary forecast*

Four skill scores (the Heidke skill score, hit rate, true skill statistic, and the threat score) are considered. These scores are defined as follows:

$$\text{Heidke skill score (HSS)} = (a+d-r)/(a+b+c+d-r)\ , \quad (1)$$

where $r = [(a+b)(a+c)+(b+d)(c+d)]/(a+b+c+d)$. The range of HSS is [0, 1].

$$\text{Hit rate (HR)} = (a+d)/(a+b+c+d)\ , \quad (2)$$

where the range of HR is [0, 1].

$$\text{True skill statistic (TSS)} = a/(a+c) - b/(b+d)\ , \quad (3)$$

where the range of TSS is $[-1, 1]$.

$$\text{Threat score (TS)} = a/(a+b+c)\ , \quad (4)$$

where the range of TS is [0, 1]. All of the above skill scores have the value 1 for the exact forecast case.

3.2 *Forecast pattern*

Forty-eight forecast patterns consisting of five distribution modes of estimated probabilities, four occurrence rates of observations, and two variation modes are considered. The forecast pattern is written by (occurrence rate, distribution mode, variation mode).

3.2.1 *Distribution modes*

The five distributions considered are: two lognormal distributions [LN(0, 1) and LN(0, 0.5)]; a normal distribution [N(0, 1)]; and two uniform distributions [U(0, 1) and U(0, 0.5)]. The above notation LN(a, b) means that its domain is the interval (a, b). For the simulation, 10000 estimated probabilities were generated from each distribution separately. For example, three histograms of 10000 estimated probabilities generated from LN(0, 1), N(0, 1) and U(0, 1) are shown in Fig. 1.

3.2.2 *Occurrence rate of observations*

The four cases of occurrence rate considered in observations are 1%, 25%, 50%, and 75%. For example, the occurrence rate of heavy snow in observations is about 1%.

3.2.3 *Variation modes*

Perfect forecast cases are easily made by sorting estimated probabilities. If the occurrence rate is 0.99, the 99th percentile is taken as the threshold, and this makes the perfect forecast. This, however, is not practical. Therefore, two variation modes (good forecast

Table 5. 2 × 2 tables of maximal skill scores: 1%, LN(0, 1), good forecast case.

	Forecast							
	HSS		HR		TSS		TS	
	0	1	0	1	0	1	0	1
OBS.								
0	9987	13	9887	13	9817	83	9987	13
	(99.87%)	(0.13%)	(99.87%)	(0.13%)	(99.16%)	(0.84%)	(99.87%)	(0.13%)
1	14	86	14	86	0	100	14	86
	(14.00%)	(89.00%)	(14.00%)	(89.00%)	(0.0%)	(100%)	(14.00%)	(89.00%)
Threshold	0.810		0.810		0.705		0.810	
Max SS	0.86296		0.9973		0.99162		0.76106	

Table 6. 2 × 2 tables of maximal skill scores: 1%, LN(0, 1), medium forecast case.

	Forecast							
	HSS		HR		TSS		TS	
	0	1	0	1	0	1	0	1
OBS.								
0	9841	59	9894	6	9686	214	9941	59
	(99.4%)	(0.6%)	(99.94%)	(0.06%)	(99.16%)	(0.84%)	(99.4%)	(0.6%)
1	18	82	51	49	0	100	18	82
	(18.00%)	(82.00%)	(51.00%)	(49.00%)	(0.0%)	(100%)	(18.00%)	(82.00%)
Threshold	0.765		0.910		0.610		0.765	
Max SS	0.67672		0.9943		0.97838		0.51572	

case and medium forecast case) are considered. The good (medium) cases are produced by first generating the variation terms from a normal distribution with mean 0 and small (large) variance, and then adding them to the estimated probabilities. Figure 2 shows plots for the cases of [1%, LN(0, 1)].

3.3 *Monte-Carlo simulation procedure*

For each occurrence rate, distribution and variation mode, the following steps should be followed:

Step 1: generate 10000 observations which consist of 0 and 1 and sort them.

Step 2: generate 10000 probabilities from each distribution and sort them.

Step 3: generate 10000 variations from a normal distribution.

Step 4: add the values of Step 3 to the probabilities of Step 2, then modify the data so that they belong to [0, 1].

Step 5: merge the data in Steps 1–4.

Step 6: with a varying threshold from 0 to 1, carry out the following:

Step 6.1: generate binary forecasts using the threshold.

Step 6.2: generate 2 × 2 tables for each case.

Step 6.3: compute the skill scores (HSS, HR, TSS, TS).

Step 7: find the threshold that has the maximal score for each skill score separately.

Step 8: compare the 2 × 2 tables for each forecast pattern.

Step 9: produce guidance on the choice of threshold for each forecast pattern.

4. Monte-Carlo simulation results

4.1 *Some results*

According to the procedure described in section 3.3, four skill scores were computed and compared for 48 cases (a combination of five distributions, four occurrence rates and two variation modes).

In the case of [1%, LN(0, 1); good forecast case], four skill scores (HSS, HR, TSS, TS) were computed separately with a varying threshold from 0 to 1. The forecast results of the maximal skill score case for each skill score are summarized in Table 5. As a result, TSS is preferable to the others.

The forecast results of maximal skill score the case [1%, LN(0, 1), Medium forecast case] are summarized in Table 6. As shown, TSS is the most preferable skill score and HR is the worst.

Table 7. Guidance on the use of skill scores.

Forecast pattern	Order	Forecast pattern	Order
1%-LN(0, 1)-Good	TSS>HSS=TS=HR	25%-LN(0, 1)-Good	TSS>HSS=TS=HR
1%-LN(0, 0.5)-Good	TSS>HSS=TS>HR	25%-LN(0, 0.5)-Good	TSS>TS>HSS>HR
1%-N(0, 1)-Good	TSS>HSS=TS>HR	25%-N(0, 1)-Good	TS>TS>HSS>HR
1%-U(0, 1)-Good	TSS>HSS=TS=HR	25%-U(0, 1)-Good	TSS>HSS=TS>HR
1%-U(0, 0.5)-Good	TSS>HSS=TS>HR	25%-U(0, 0.5)-Good	HSS=TS>HR>TSS
1%-LN(0, 1)-Medium	TSS>HSS=TS>HR	25%-LN(0, 1)-Medium	TSS>HSS=TS>HR
1%-LN(0, 0.5)-Medium	TSS> HSS=TS>HR	25%-LN(0, 0.5)-Medium	TSS>TS>HSS>HR
1%-N(0, 1)-Medium	TSS>HSS=TS>HR	25%-N(0, 1)-Medium	HSS>TS>TSS>HR
1%-U(0, 1)-Medium	TSS>HSS=TS=HR	25%-U(0, 1)-Medium	HSS=TS=HR>TSS
1%-U(0, 0.5)-Medium	TS>TSS>HSS>HR	25%-U(0, 0.5)-Medium	HR>HSS>TS>TSS
50%-LN(0, 1)-Good	HSS=HR=TSS>TS	75%-LN(0, 1)-Good	TSS>HSS>HR=TS
50%-LN(0, 0.5)-Good	HSS=HR=TSS>TS	75%-LN(0, 0.5)-Good	TSS>HSS=HR=TS
50%-N(0, 1)-Good	HSS=HR=TSS>TS	75%-N(0, 1)-Good	TSS>HSS>HR>TS
50%-U(0, 1)-Good	HSS=HR=TSS=TS	75%-U(0, 1)-Good	TSS>HSS=HR=TS
50%-U(0, 0.5)-Good	HSS=HR=TSS>TS	75%-U(0, 0.5)-Good	TSS>HSS>HR=TS
50%-LN(0, 1)-Medium	HSS=HR=TS>TS	75%-LN(0, 1)-Medium	TSS>HSS>HR=TS
50%-LN(0, 0.5)-Medium	HSS=HR=TSS>TS	75%-LN(0, 0.5)-Medium	TSS>HSS=HR=TS
50%-N(0, 1)-Medium	HSS=HR=TSS>TS	75%-N(0, 1)-Medium	TSS>HSS>HR>TS
50%-U(0, 1)-Medium	HSS=HR=TSS=TS	75%-U(0, 1)-Medium	HSS>TSS>HR=TS
50%-U(0, 0.5)-Medium	HSS=HR=TSS>TS	75%-U(0, 0.5)-Medium	TSS>HSS>HR>TS

4.2 *Guidance on the choice of skill score*

The results of comparing the four skill scores are summarized in Table 7 for all 48 forecast patterns. This table can be used to select the optimal threshold using the skill scores for a given binary forecast pattern. Looking at Table 7, it seems that the main factor is the occurrence rate. In the case of 1%, TSS can be recommended, except for [N(0.3, 0.7), good), [N(0, 1), medium) and [U(0, 0.5), medium]. In the case of 50%, HSS is preferable; and in the case of 75%, TSS is preferable, except for [U(0, 1), Medium]. There are, however, various options for the case of 25%.

4.3 *How to use the guidance*

When developing a binary forecast model based on a probability forecast model, it is necessary to find the optimal threshold by use of an appropriate skill score. The proposed guidance can be used for this purpose.

First of all, one should find the binary forecast pattern which consists of the occurrence rate of observations, the distribution of estimated probability from the probabilistic forecast model, and the variation mode using past data. If the binary forecast pattern is recognized, the appropriate skill score can then be chosen using the proposed guidance (Table 7), and the optimal threshold which maximized the selected skill score can then be found.

Using past data, one can find the binary forecast pattern, the appropriate skill score and the optimal threshold, which generates the binary forecast as follows:

Step 1 (occurrence rate): calculate the occurrence rate of the meteorological event of interest in binary observations.

Step 2 (distribution mode): generate the estimated probabilities using the applied probability forecast model, and then plot the histogram of the estimated probabilities.

Step 3 (variation mode): sort the binary observations and then draw a scatter diagram between the sorted observations and the corresponding estimated probabilities. The scatter diagram will show the variation mode.

Step 4 (choice of skill score): select the best skill score for the forecast pattern from Table 7.

Step 5 (optimal threshold): find the optimal threshold which maximizes the selected skill score.

Step 6 (generate forecasts): using the determined optimal threshold, generate the binary forecasts based on the probabilities from the probabilistic forecast model.

For instance, TS is recommended when the occurrence rate is about 25%, the shape of estimated probabilities from a forecast model is similar to N(0,1), and the variation mode seems to be good.

4.4 *Case study (revisited)*

Considering again the case study outlined in section 2, the shape of the distribution of estimated probabilities is similar to LN(0, 0.5) and the occurrence rate is about 1%. For the forecast pattern [1%, LN(0, 0.5), good] or [1%, LN(0, 0.5), medium], the guidance from Table 7 recommends using the TSS to find an optimal threshold. The TSS is maximized at threshold 0.02; that is, forecasters will announce that heavy snow will occur when the probability from the probabilistic forecast model (in this case, the estimated logistic regression model) is greater than 0.02.

5. Concluding remarks

This paper has proposed guidance on the choice of optimal thresholds using skill scores for binary forecast models based on forecast patterns that consist of a distribution of estimated probabilities from a probabilistic forecast model, occurrence rates in observations, and variation modes. The guidance was produced through Monte-Carlo simulation. Forty-eight forecast patterns were considered and four skill scores (HSS, HR, TSS, and TS) were compared for each.

As a result, the guidance set out in Table 7 is proposed for each pattern. Of course, this guidance cannot cover all kinds of binary forecast models because only 40 patterns were considered. However, the guidance is useful for the development of binary forecasts because many meteorological events have similar patterns to the 40 considered in this paper. In future work, guidance for multi-category forecasts and more types of forecast patterns will be developed.

Acknowledgements. This work was funded by the Korea Meteorological Administration Research and Development Program under Grant CATER 2006–2203.

REFERENCES

Barnston, A. G., 1992: Correspondence among the correlation, RMSE, and Heidke forecast verification measures; Refinement of the Heidke score. *Wea. Forecasting*, **7**, 699–709.

Burrows, W. R., 1991: Objective guidance for 0–24 hour and 24–48 hour mesoscale forecasts of lake-effect snow using CART. *Wea. Forecasting*, **6**, 357–378.

Hans, V. S., and W. Z. Francis, 1999: *Statistical Analysis in Climate Research*, Cambridge University Press, 391–406.

Heidke, P., 1926: Berechnung des Erfolges und der Güte der Windstärkevorhersagen im Sturmwarnungsdienst. *Geografiska Annaler*, **8**, 301–349.

Mcbride, J. L., and E. E. Ebert, 2000: Notes and correspondence, verification of quantitative precipitation forecasts from operational numerical weather prediction models over Australia. *Wea. Forecasting*, **15**, 103–121.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293.

Sohn, K. T., 2006: Binary forecast of heavy snow using statistical models. *Korean Communications in Statistics*, **13**(2), 369–378. (In Korean with English abstract)

Sohn, K. T., and J. I. Han, 2004: New skill score of forecast model for probability of ordinary predictands. *Journal of the Korean Data Analysis Society*, **6**(1), 267–278. (in Korean with English abstract)

Sohn, K. T., J. H. Lee, and C. S. Ryu, 2005a: Statistical prediction of heavy rain in South Korea. *Adv. Atmos. Sci.*, **22**(5), 703–710.

Sohn, K. T., J. H. Lee, S. H. Lee, and C. S. Ryu, 2005b: Statistical models for prediction of heavy rain in Honam area. *Journal of the Korean Meteorological Society*, **41**(6), 897–907. (in Korean with English abstract)

Woodcock, F., 1976: The evaluation of yes/no forecasts for scientific and administrative purpose. *Mon. Wea. Rev.*, **104**, 1209–1214.

Zhang, H., and T. Casey, 2000: Verification of categorical probability forecasts. *Wea. Forecasting*, **15**, 80–89.