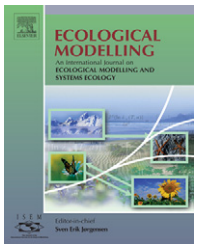
available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/ecolmodel



A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa

Elizabeth A. Freeman, Gretchen G. Moisen*

USDA Forest Service, Rocky Mountain Research Station, 507 25th Street, Ogden, UT 84401, USA

ARTICLE INFO

Article history:
Received 24 July 2007
Received in revised form
30 April 2008
Accepted 14 May 2008
Published on line 4 July 2008

Keywords:
Binary classification
ROC
AUC
Sensitivity
Specificity
Threshold

ABSTRACT

Modelling techniques used in binary classification problems often result in a predicted probability surface, which is then translated into a presence–absence classification map. However, this translation requires a (possibly subjective) choice of threshold above which the variable of interest is predicted to be present. The selection of this threshold value can have dramatic effects on model accuracy as well as the predicted prevalence for the variable (the overall proportion of locations where the variable is predicted to be present). The traditional default is to simply use a threshold of 0.5 as the cut-off, but this does not necessarily preserve the observed prevalence or result in the highest prediction accuracy, especially for data sets with very high or very low observed prevalence. Alternatively, the thresholds can be chosen to optimize map accuracy, as judged by various criteria. Here we examine the effect of 11 of these potential criteria on predicted prevalence, prediction accuracy, and the resulting map output. Comparisons are made using output from presence–absence models developed for 13 tree species in the northern mountains of Utah. We found that species with poor model quality or low prevalence were most sensitive to the choice of threshold. For these species, a 0.5 cut-off was unreliable, sometimes resulting in substantially lower kappa and underestimated prevalence, with possible detrimental effects on a management decision. If a management objective requires a map to portray unbiased estimates of species prevalence, then the best results were obtained from thresholds deliberately chosen so that the predicted prevalence equaled the observed prevalence, followed closely by thresholds chosen to maximize kappa. These were also the two criteria with the highest mean kappa from our independent test data. For particular management applications the special cases of user specified required accuracy may be most appropriate. Ultimately, maps will typically have multiple and somewhat conflicting management applications. Therefore, providing users with a continuous probability surface may be the most versatile and powerful method, allowing threshold choice to be matched with each maps intended use.

Published by Elsevier B.V.

1. Introduction

Binary classification mapping is a technique crucial to multiple areas of study. Applications include mapping species distribution, disturbance, wildlife habitat, insect and disease outbreaks, fire risk, and climate change. Modelling techniques often generate predictions that are analogous to a probability of presence. A common practice is to translate this surface

* *Corresponding author*. Tel.: +1 801 510 3765.
E-mail address: eafreeman@fs.fed.us (E.A. Freeman).
0304-3800/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.ecolmodel.2008.05.015

into a simple 0/1 classification map by a choice of threshold, or cut-off probability, beyond which something is classified as present. The selection of this threshold value can have dramatic effects on model accuracy as well as the predicted prevalence (the overall proportion of locations where the variable is predicted to be present). The traditional default is to simply use a threshold of 0.5 as the cut-off, but this does not necessarily preserve the observed prevalence or result in the highest prediction accuracy, especially for data sets with very high or very low observed prevalence.

Alternatively, the thresholds can be chosen to optimize map accuracy, as judged by one of several criteria. Because the utility of maps for different management applications cannot be captured in a single map accuracy number, several global measures are commonly used to assess the predictive performance of the models; these include percent correctly classified (PCC), sensitivity, specificity, kappa, and receiver operating curves (ROC plots), with their associated area under the curve (AUC). In addition, in many applications, it is important that the predicted prevalence reflects the observed prevalence, and agreement between these two may also be used as a measure of map accuracy. All of these numerous accuracy measures have been used as in various ways to create criteria for threshold optimization, as described below.

Beginning with the simplest of these measures, PCC is the proportion of test observations that are correctly classified. However this can be deceptive when prevalence is very low or very high. For example, species with very low prevalence, it is possible to maximize PCC simply by declaring the species absent at all locations, resulting in a map with little usefulness. Although sometimes used to optimize threshold values, the accuracy measure itself has little value in practice.

As a result, classification accuracy is commonly broken down into two measures. Sensitivity, or proportion of correctly predicted positive observations, reflects a model's ability to detect a presence, given a species actually occurs at a location. Specificity, or proportion of correctly predicted negative observations, reflects a model's ability to predict an absence where a species does not exist. Sensitivity and specificity can be combined in various ways to assess model quality and optimize thresholds. Fielding and Bell (1997) suggest choosing the threshold where sensitivity equals sensitivity, in other words, where positive and negative observations have equal chance of being correctly predicted. Alternatively, Manel et al. (2001) and Hernandez et al. (2006) maximize the sum of sensitivity and specificity for threshold selection.

Allouche et al. (2006) subtract a constant of 1 from the sum of sensitivity and specificity. This is equivalent to the true positive rate (the proportion of observed presences correctly predicted) minus the false positive rate (the proportion of observed absences incorrectly predicted). They refer to this as the true skill statistic (TSS), and recommend it for model evaluation and comparison, especially when comparing across populations with differing prevalence. In the medical literature, TSS is referred to as Youden's index, and is used in the evaluation of diagnostic tests (Biggerstaff, 2000). However, there is a difference between assessing model performance, and selecting an optimal threshold, and while the true skill statistic itself is independent of prevalence, Manel et al. (2001) found that selecting a threshold to maximize the sum of sensitivity and specificity affects the predicted prevalence of the map, causing the distribution of rare species to be overestimated.

Another way to utilize sensitivity and specificity in threshold selection is to deliberately pick a threshold that will meet a given management requirement. Fielding and Bell (1997) discuss the possibility that a user may have a predetermined required sensitivity or specificity. Perhaps to meet management goals, it is determined that 15% is the minimum acceptable error in the observed presences, and thus a map is required that has a sensitivity of at least 0.85. In a similar vein, Wilson et al. (2005) studied the effects of the various methods of utilizing a probability surface with the goal of defining reserve networks to protect biodiversity. In this work, they contrasted three methods of threshold selection, one of which involved trading off sensitivity to meet a predetermined specificity requirement. They also looked at two methods of working directly from the probability surface, without first creating classification maps.

Another accuracy measure, the kappa statistic, measures the proportion of correctly classified locations after accounting for the probability of chance agreement. While still requiring a choice of threshold, kappa is more resistant to prevalence than PCC, sensitivity and specificity, and was found by Manel et al. (2001) to be well correlated with the area under the curve of ROC plots. Caution is required when using the kappa statistic to compare models across multiple populations. A particular value of kappa from one population is not necessarily comparable to the same kappa value from a different species or location, if the prevalence differs between the two populations (McPherson et al., 2004; Vaughan and Ormerod, 2005; Allouche et al., 2006). Kappa has been used extensively in map accuracy work (Congalton, 1991), and in presence–absence mapping, a threshold can be deliberately selected to maximize kappa (Guisan and Hofer, 2003; Hirzel et al., 2006; Moisen et al., 2006).

While threshold-dependent accuracy measures such as PCC, sensitivity, and specificity have a long history of use in ecology, ROC plots are a technique that has recently been introduced into ecology that provides a threshold-independent method of evaluating the performance of presence–absence models. In a ROC plot the true positive rate (sensitivity) is plotted against the false positive rate (1 – specificity) as the threshold varies from 0 to 1. A good model will achieve a high true positive rate while the false positive rate is still relatively small; thus the ROC plot will rise steeply at the origin, and then level off at a value near the maximum of 1. The ROC plot for a poor model (whose predictive ability is the equivalent of random assignment) will lie near the diagonal, where the true positive rate equals the false positive rate for all thresholds. Thus the area under the ROC curve is a good measure of overall model performance, with good models having an AUC near 1, while poor models have an AUC near 0.5

ROC plots can also be used to select thresholds. As the upper left corner of the ROC plot can be considered the 'ideal' model (sensitivity and specificity both equal 1.0), the threshold which minimizes the distance between the ROC plot and this 'ideal' point can be used as an optimization criteria. In the medical literature, Cantor et al. (1999) performed a review