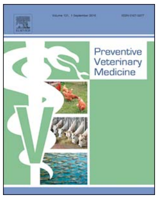
Preventive Veterinary Medicine 149 (2018) 107–114

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

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



A discussion of calibration techniques for evaluating binary and categorical predictive models

Caroline Fenlon[a,*], Luke O’Grady[b], Michael L. Doherty[b], John Dunnion[a]

[a] *School of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland, Ireland*
[b] *School of Veterinary Medicine, University College Dublin, Belfield, Dublin 4, Ireland*

ARTICLE INFO

Keywords:
Predictive modelling
Data mining
Evaluation
Calibration
Deviance
Discrimination

ABSTRACT

Modelling of binary and categorical events is a commonly used tool to simulate epidemiological processes in veterinary research. Logistic and multinomial regression, naïve Bayes, decision trees and support vector machines are popular data mining techniques used to predict the probabilities of events with two or more outcomes. Thorough evaluation of a predictive model is important to validate its ability for use in decision-support or broader simulation modelling. Measures of discrimination, such as sensitivity, specificity and receiver operating characteristics, are commonly used to evaluate how well the model can distinguish between the possible outcomes. However, these discrimination tests cannot confirm that the predicted probabilities are accurate and without bias.

This paper describes a range of calibration tests, which typically measure the accuracy of predicted probabilities by comparing them to mean event occurrence rates within groups of similar test records. These include overall goodness-of-fit statistics in the form of the Hosmer-Lemeshow and Brier tests. Visual assessment of prediction accuracy is carried out using plots of calibration and deviance (the difference between the outcome and its predicted probability). The slope and intercept of the calibration plot are compared to the perfect diagonal using the unreliability test. Mean absolute calibration error provides an estimate of the level of predictive error. This paper uses sample predictions from a binary logistic regression model to illustrate the use of calibration techniques. Code is provided to perform the tests in the R statistical programming language. The benefits and disadvantages of each test are described.

Discrimination tests are useful for establishing a model’s diagnostic abilities, but may not suitably assess the model’s usefulness for other predictive applications, such as stochastic simulation. Calibration tests may be more informative than discrimination tests for evaluating models with a narrow range of predicted probabilities or overall prevalence close to 50%, which are common in epidemiological applications. Using a suite of calibration tests alongside discrimination tests allows model builders to thoroughly measure their model’s predictive capabilities.

1. Introduction

Predictive statistical models are powerful tools for the mathematical representation of biological systems. Accurate predictive models can be useful for standalone decision-support or as part of larger simulation models. The ability of models to highlight the differences in scenarios with different predictor values allows for informed management and planning. Predictive models are often built using data mining techniques to identify the factors which have the most impact on the system under scrutiny and the magnitude of their impact. The resultant tool is used to predict a probability of the outcome event’s occurrence. Simulation or decision-support applications may use these raw probabilities stochastically or convert them to a binary or categorical outcome using threshold probabilities (typically 50% for binary outcomes). Probabilistic modelling is becoming popular for the simulation of binary and categorical outcomes in animal epidemiology (Petrie and Watson, 2013), with applications such as the prediction of reproductive and health events. Many supervised data mining techniques can predict the probabilities of one or more outcomes.

Verifying the predictive ability of a model is a critical step in the model-building process. The inclusion of variables in the model is typically determined by statistical significance or other measures such as

Abbreviations: ROC, receiver operating characteristic; MACE, mean absolute calibration error

[*] Corresponding author.

E-mail address: caroline.fenlon@ucdconnect.ie (C. Fenlon).

https://doi.org/10.1016/j.prevetmed.2017.11.018
Received 19 May 2017; Received in revised form 31 October 2017; Accepted 22 November 2017

information gain. When choosing the model that best fits the training data, internal evaluation such as information criteria (e.g. Akaike information criterion (Akaike, 1974), analysis of variance and analysis of fitted residuals can be used. Model-building procedures may also focus on optimising measures of predictive ability.

Evaluation of model predictions is an important task to verify the validity of the model, using either the model training data, or, for more generalisation, an external dataset. This consists of comparing true and predicted values. The evaluation of continuous estimations is well established, with several comparison techniques, e.g. scatter plots, R^2, and error calculations. This task is not so straightforward for models of discrete outcomes, where the observed outcomes are binary or categorical, but the predictions are probabilistic. The predictive ability of probabilistic models is typically evaluated using discrimination tests. Discrimination measures a model's ability to correctly classify cases, i.e. the separation between the successful and unsuccessful outcomes. Rather than evaluating the accuracy of the raw model predictions, these probabilities are transformed into the most likely binary or categorical outcomes. The predicted outcomes are then compared to the true outcome.

Discrimination methods are undoubtedly a valuable technique for evaluating a model's ability to separate the different possible outcomes. However, they cannot confirm that the model's probability predictions are free from bias. Even if the model scores highly in discrimination tests, there may be probability ranges or covariate combinations that it does not handle well. Identifying these issues is a key area of evaluation, particularly in models designed for use in probability-focused applications such as decision-support or simulation programs. Calibration tests can be used to measure the reliability of predicted probabilities. These tests include overall goodness-of-fit measures, absolute error calculation and in-depth visual assessment of predictions. As the modelled outcomes are typically binary or categorical values, most calibration tests work by grouping the records and comparing mean occurrence rates to the mean predicted probabilities within each group. Calibration tests have been used to evaluate only a few applications related to animal health or agriculture: a model predicting the probability that a herd's somatic cell count exceeded the acceptable limit (Fauteux et al., 2015); studies related to the presence or absence of a species in a region (Pearce and Ferrier, 2000); and models of dairy cow conception (Fenlon et al., 2017a) and calving difficulty (Fenlon et al., 2017b).

This paper describes the usage, advantages, and disadvantages of a range of calibration measures. Examples of code and results are presented for each.

2. Modelling techniques

Any model capable of predicting probabilities is suitable for use with calibration evaluation methods. Commonly used modelling methods include logistic regression, naïve Bayes, classification and regression trees, support vector machines, neural networks and many others (Olson and Delen, 2008). All of the methods are suitable for both binary and categorical outcomes; evaluating the accuracy of probabilities for categorical outcomes is done individually for each possible outcome, while binary outcomes are evaluated in terms of the event's probability of occurrence. Ensemble techniques such as bagging and boosting are all equally compatible with calibration methods.

For thorough assessment of the predictive ability of a model and to prevent the under-reporting of its error, an external dataset should be used for testing. If this is not possible, random partitioning of the data should be used to generate training and testing datasets.

3. Calibration techniques

As described in the introduction, calibration techniques compare the predicted probabilities to the true proportions of events occurring, i.e. determining if the observed frequency of actual events is similar to the predicted probability, within groups of records in the test dataset (Hosmer et al., 2013).

All code samples below are in the R statistical programming language (R Core Team, 2015) and packages available for use with it. Similar functions are available in most other similar languages or statistical tools.

3.1. Brier test

The Brier score is an overall goodness-of-fit check for binary and categorical values (Brier, 1950). It is calculated as the average squared difference between each binary outcome and its predicted probability:

$$B = \frac{\sum_{i=1}^{N} (Y_i - p_i)^2}{N}$$

where N = the number of records in the test set, Y_i = the true event outcome of record i (1 or 0), p_i = the predicted probability of record i.

The score will be 0 for a perfect model. Its maximum value will depend on the incidence of the outcome; for a binary outcome with 50% overall incidence, the maximum Brier score will be 0.25 (Steyerberg et al., 2010).

To instead set the result to range between 0 and 1, it can be scaled by the maximum score, such that the scaled Brier score is the mean squared error of prediction, with a perfect model having 0% error. The maximum score is calculated from the mean predicted probability p_{mean} and equals $p_{mean} * (1 - p_{mean})$.

The Brier test is a general evaluation of performance similar to the R^2 for linear predictions. It is appropriate for binary and unordered categorical variables, but not ordered variables with more than two outcomes, as the calculation assumes an equal distance between the values. The (unscaled) Brier test can be calculated using the "val.prob" function from the rms package (Harrell, 2015a). The transformation described above can then be used to produce the scaled Brier score.

3.2. Hosmer-Lemeshow test

The Hosmer-Lemeshow test (Hosmer and Lemeshow, 1980) also evaluates the overall goodness-of-fit of model predictions. The test splits the observations (sorted by predicted probability) into *g* (user-defined) equal-sized groups of risk and compares the observed proportion of event outcomes ($\text{observed}_{\text{true}}$ and $\text{observed}_{\text{false}}$) to the mean predicted proportion of events ($\text{expected}_{\text{true}}$ and $\text{expected}_{\text{false}}$) within each group. This was calculated as:

$$\hat{C} = \sum_{i=1}^{g} \frac{(observed_{true\,i} - expected_{true\,i})^2}{expected_{true\,i}} + \frac{(observed_{false\,i} - expected_{\,false\,i})^2}{expected_{false\,i}}$$

$\hat{C}$ is distributed as χ^2 with g–2 degrees of freedom. The null hypothesis of the test is that the model fits the data in question correctly; the resultant Pearson χ^2 p-value should be lower than the chosen α (typically set to 0.05) to reject the null hypothesis and find a statistically significant difference between the true and predicted outcomes.

The value of g is chosen arbitrarily by the user, but is typically chosen to be 10 (with each of the groups called "deciles of risk") by Hosmer et al. (2013). Hosmer and Lemeshow do not believe there is a necessity for a minimum frequency of event occurrence within the groups, but suggest aggregating adjacent groups for higher frequencies and lower degrees of freedom if desired (Hosmer et al., 2013).

In R, the Hosmer-Lemeshow test can be performed for binary, categorical or ordinal outcomes using the "logitgof" function of the generalhoslem package (Jay, 2017). The test may also be performed for binary outcomes using the ResourceSelection package (Johnson et al., 2006). Alternative goodness-of-fit tests for ordinal outcomes include the Lipsitz and Pulkstenis-Robinson tests, which are also available in the generalhoslem R package.