Folia Oeconomica
Acta Universitatis Lodziensis
ISSN 0208-6018 e-ISSN 2353-7663

www.czasopisma.uni.lodz.pl/foe/
6(339) 2018

DOI: http://dx.doi.org/10.18778/0208-6018.339.06

Michał Trzęsiok
University of Economics in Katowice, Faculty of Finance and Insurance, Department of Economic and Financial Analysis, michal.trzesiok@ue.katowice.pl

Measuring the Quality of Multivariate Statistical Models

Abstract: Assessing the quality of a statistical model is very important, since it is crucial for the utility of the modelling process' outcome. There are many different ways of measuring statistical models' quality. Some of the measures represent a "goodness of fit" approach, some are "prediction ability" orientated. Among them there are absolute and relative measures. It is a researcher's decision, which model quality measure is the most adequate for the given task. In the paper we present an overview of statistical models' quality measures and a suggestion of using different ones during the model type selection stage and the stage of assessing the quality of the final model.

Keywords: model quality, goodness of fit, prediction error

JEL: C150, C180, C300, C310, C380

1. Introduction

The contemporary multivariate statistical modelling methods are very powerful tools used in many different fields because of their explanatory power and overall good predictive abilities. In order to build a model with high predictive accuracy, it is equally important: to choose adequate modelling method and to provide data of good quality. This paper focuses on the process of measuring the quality of the model used in the analysis. The scope of the paper is limited to the methods for cross-sectional data and more specifically to classification and multiple regression methods.

The implication of the *No Free Lunch Theorem* (Wolpert, Macready, 1997) is that, when averaged over all possible problems, no given method will perform better than any other. In other words, there is no such thing as the best method for all possible problems. Nevertheless, among the classification and regression methods there is a group of machine learning methods that proved in different benchmarking studies to have a very strong position in terms of predictive accuracy (Meyer, Leisch, Hornik, 2003; Trzęsiok, 2006; 2007). Two ensemble methods based on classification and regression trees – *Random Forests*, *Bagging*, and *Support Vector Machines* are very often in the top five in the rankings. These three methods also have the ability to be used, and perform well both in classification and regression tasks. Thus, we used these methods for illustration in the paper. In order to build a model with good predictive power you need to tune some internal parameters which the selected methods depend on. This choice of parameters' values is usually based on simulation study (e.g. *b*-fold cross validation) with cross-validated prediction error used as a measure for assessing the model quality. The same measure is then used for assessing the quality of the final model. This common practice (using the same measure when choosing the model variant and then for assessing final model's predictive abilities) is very controversial. The goal of the paper is to present an overview of different model quality measures and then select two distinct approaches and use one of them to support model variant selection and another one to assess the final model quality.

2. Model quality measures – a short overview

Let us assume that we are given the training set $D = \left\{ \left(\mathbf{x^1}, y^1\right), \ldots, \left(\mathbf{x^N}, y^N\right) \right\}$, where $\mathbf{x^i} \in \mathbf{R}^d$ is the vector of predictors' values and $y^i \in \{-1,1\}$ defines the class the *i*-th observation belongs to, $i \in \{1, \ldots, N\}$ (we will consider only classifications problems with two classes). Then the goal of supervised learning is to find a "good" predictive classification function $y = f(\mathbf{x})$, based on the available training set.

For the classification task, the most common measure of model quality is the *classification error*, which is defined as:

$$\varepsilon(Q) = \frac{1}{|Q|}\sum_{i=1}^{|A|} I\left((\mathbf{x}^i, y^i) \in Q\right) \cdot I\left(y^i = f(\mathbf{x}^i)\right), \tag{1}$$

where A is the set of all available observations and Q is a subset of A ($Q \subseteq A$) containing the observations the classification error is measured on. If $Q = D$, then (1) is a *goodness-of-fit* measure (resubstitution error) denoted by ε_{TRAIN}. If Q is a test set or validation set, the measure is referred to as *predictive ability* of the model (ε_{TEST}).

Requiring separate training set and test set usually means wasting the information that is enclosed in the test set, which is available and could be used in the training process. In order to incorporate this information the b-fold cross validation technique can be applied, where the original sample is randomly partitioned into b subsamples and one is left out in each iteration as validation set (on which the classification error is computed) and the remaining part is used for training. Then the average of the obtained b classification errors (ε_{b-CV}) is used and it is an unbiased estimator of the true classification error over all possible observations (Kohavi, 1995; Rozmus, 2008: 40–41). There is also a possibility of using a different sampling technique, namely *bootstrapping* and computing the classification error on the set of observations that were not included in the given bootstrap sample (OOB – Out of Boost observations). As a result we obtain another measure of predictive ability of the classifier – ε_{OOB}. Although all presented measures are computed in a similar way, they must be seen as distinct model quality measures.

With imbalanced data sets, an algorithm does not get the necessary information about the minority class to make an accurate prediction (especially for observations from the minority class). None of the presented classification errors take into account the consequences of dealing with imbalanced data. One of the possible solution is to use a different performance measure based on *sensitivity* and *specificity*. These two measures are defined for the situation of two class classification problem, where one of the classes is labelled as "positive" and the other one as "negative". After building the classification model we get a contingency table presented as Table 1.

Table 1. Contingency table for two class classification

		Observed (true) class	
		positive	**negative**
Predicted class	positive	*TP* (True Positives)	*FP* (False Positives)
	negative	*FN* (False Negatives)	*TN* (True Negatives)

Source: own results

The *sensitivity* (or True Positives Rate, *TPR*) is defined as:

$$TPR = \frac{TP}{TP + FN}\,. \tag{2}$$

The *specificity* (or True Negatives Rate, *TNR*) is defined as:

$$TNR = \frac{TN}{FP + TN}\,. \tag{3}$$

The ROC (Receiver Operating Characteristics) curve is the base for measuring the accuracy of prediction. It is a widely used evaluation metric. ROC curve is formed by plotting *TPR* (sensitivity) vs *FPR* = 1 – *TNR* (one minus specificity) for different possible cut-points of a classifier. Any point on ROC graph, corresponds to the performance of a single classifier on a given distribution. The optimal point on the ROC curve is (FPR, TPR) = (0, 1) – no false positives and all true positives. So the closer we get there the better (Figure 1). The larger the *area under ROC curve* (*AUC*), the higher the accuracy (Altman, Bland, 1994; Misztal, 2014). The measure *AUC* is equal to 0.5 for a random classifier and *AUC* = 1 for a perfectly classifying model. *AUC* of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance, higher than a randomly chosen negative instance.

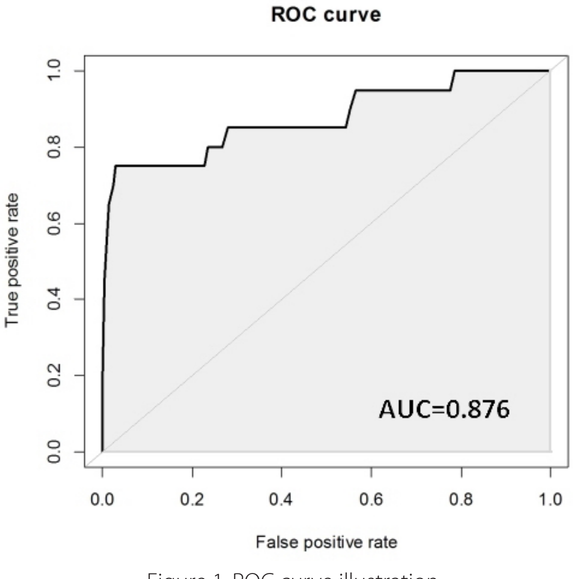

Figure 1. ROC curve illustration

Source: own results

In the regression case (i.e. $y^i \in R$), the most often used model quality measure is the *mean squared error* (*MSE*):

$$MSE\left(Q\right)=\frac{1}{|Q|}\sum_{i=1}^{|A|}I\left((\mathbf{x}^i,y^i)\in Q\right)\cdot(y^i-f(\mathbf{x}^i))^2. \tag{4}$$

Similarly to the classification task, there are different versions of *MSE* depending on what is the dataset *Q* on which the measure is computed. Thus we get goodness-of-fit measure MSE_{TRAIN} and a few prediction ability measures: MSE_{TEST}, MSE_{b-CV}, MSE_{OOB}. For interpretational purposes *root mean squared error* $RMSE\left(Q\right)=\sqrt{MSE\left(Q\right)}$ is often used. Another measure is *mean absolute error*:

$$MAE\left(Q\right)=\frac{1}{|Q|}\sum_{i=1}^{|A|}I\left((\mathbf{x}^i,y^i)\in Q\right)\cdot\left|y^i-f(\mathbf{x}^i)\right|, \tag{5}$$

or *mean absolute percentage error*:

$$MAPE\left(Q\right)=\frac{1}{|Q|}\sum_{i=1}^{|A|}I\left((\mathbf{x}^i,y^i)\in Q\right)\cdot\left|\frac{y^i-f(\mathbf{x}^i)}{y^i}\right|, \tag{6}$$

or widely used as a *goodness-of-fit* measure (for *Q* = *D*) – determination coefficient:

$$R^2\left(Q\right)=1-\frac{\sum_{i=1}^{|A|}I\left((\mathbf{x}^i,y^i)\in Q\right)\cdot(y^i-f(\mathbf{x}^i))^2}{\sum_{i=1}^{|A|}I\left((\mathbf{x}^i,y^i)\in Q\right)\cdot(y^i-\overline{y^i})^2}. \tag{7}$$

Naturally there is also an adjusted version of determination coefficient $\hat{R}^2\left(Q\right)$ which is a modified version of $R^2(Q)$ and it penalizes you for adding independent variables that do not affect the dependent variable.

These approaches are similar in methodological sense, since they are based on residuals and the only difference is mainly whether we use absolute or squared values. A slightly different approach is applied when using *prediction quality indicator* for proportion *m* ($m\in(0,1)$):

$$pred\left(m,Q\right)=\frac{\sum_{i=1}^{|A|}I\left((\mathbf{x}^i,y^i)\in Q\right)\cdot I\left(f(\mathbf{x}^i)\in\left[(1-m)y^i,(1+m)y^i\right]\right)}{|Q|}. \tag{8}$$

pred(*m*, *Q*) is simply the percentage of estimates that are within $m \cdot 100\%$ of the actual value (Kitchenham et al., 2001). Typically *m* is set to 0.25, so the indicator reveals what proportion of estimates are within a tolerance of 25%. Clearly, *pred*(*m*, *Q*) is insensitive to the degree of inaccuracy of estimates outside the specified tolerance level.

The presented list of model quality measures is not complete. There are also other measures, e.g.: *Akaike or Bayesian information criterion* (*AIC*, *BIC*), *mean absolute scaled error* (*MASE*) or *Mallows's* C_p, but all these measures can be used in a specific context only, i.e. *information criteria* for parametric models only, *MASE* for time series only, and *Mallows's* C_p for models of hierarchical structure. In this paper we focus on universal measures of model quality.

3. Procedure for model selection and evaluation using different quality measures

As mentioned in the introduction, building a model with good predictive power requires some internal parameters tuning. This choice of parameters' values is usually based on simulation study (e.g. *b*-fold cross validation) with cross-validated prediction error (for classification problem) and mean squared error (for regression). The same measure is then used for assessing the quality of the final model. In this section we suggest to use different measures in the stage of model (parameters) selection and the stage of model quality evaluation.

We consider three machine learning methods, that can be used both – for classification and regression tasks: *bagging* (for detailed description see Hastie, Tibshirani, Friedman, 2001: 246–247; Gatnar, 2008), *random forest* (Breiman, 2001; Rozmus, 2004) and *Support Vector Machines* (SVMs), (Vapnik, 1998; Trzęsiok, 2006). These methods have the following parameters (crucial for the method performance) that need to be carefully chosen by the user [in the parenthesis we present the range of values searched for in the cross-validation]:

1) for *bagging*: `nbagg` [in: 10, 20, 50, 80, 100] – the number of bootstrap replications (i.e. the number of models in the ensemble) and `minsplit` [in: 2, 3,…, 10] – the minimum number of observations that must exist in a node in order for a split to be attempted;
2) for *random forest*: `mtry` [in: $\frac{\sqrt{d}}{2}$, $\sqrt{d}$, *and* $2\sqrt{d}$] – number of variables randomly sampled as candidates at each split, `nodesize` [in: 1, 2,…, 10] – minimum size of terminal nodes, and `ntree` [in: 1, 2,…, 10] – number of trees to grow (i.e. the number of models in the ensemble);

3) for *Support Vector Machines*: `kernel` [`'polynomial'` or `'radial'`] – the kernel used in training and predicting; `degree` [in: 2, 3, 4] – parameter needed for kernel of type polynomial, `gamma` [in: 0.1, 0.5] – parameter needed for all kernels except linear, `epsilon` [in 0.01, 0.1] – epsilon in the insensitive-loss function (regression case), `cost` [in 0.01, 0.1] – cost of constraints violation (regularization parameter).

In the procedure, the model selection stage was performed using *b*-fold cross validation and for the final model we chose one which had:

1) the maximum value of the *AUC* measure in case of classification problems,
2) the maximum value of the *pred*(0.25, *b* – *CV*) measure in case of regression problems.

After choosing the suboptimal configuration of the parameters and building the final model, we evaluated the model quality using the standard measures:

1) the cross-validated classification error ε_{b-CV} in case of classification problems,
2) the cross-validated mean squared error MSE_{b-CV} in case of regression problems.

We used the *b*-fold cross validation technique with $b = 10$.

4. Examples illustrating the procedure

We present two empirical examples illustrating the procedure of using different measures for model selection and evaluation – one example for classification problem and one for regression.

4.1. An example of the model selection and evaluation procedure applied to classification problem

To illustrate how the procedure described in Section 3 works in the classification problem we used a real-world dataset `german credit` shared by prof. dr hab. Hans Hofmann from the Institute of Statistics and Econometrics, University of Hamburg. This dataset set is available in the UCI Repository of Machine Learning Databases[1] (University of California, Irvine). The dataset includes information about short term loans. The task is the classical *credit scoring* problem – given a dataset representing the credit history of 1000 bank customers, find the classification function that would classify a new client into one of two groups: "good clients" who represent low credit risk and "bad clients" with high credit risk. This function should be an automatic support in the decision making process whether

[1] ftp://ftp.ics.uci.edu/pub/machine-learning-databases.

or not to accept an application form for granting a loan. The general information about the analysed dataset is presented in Table 2.

Table 2. General information about the `german credit` dataset

No. of observations	No. of input variables	
	interval	nominal
1000	7	13

Source: own results

The set of input variables consist of: status of the checking account, loan duration in number of months, credit history (no credits taken, all credits paid back duly, delay in paying off in the past, other credits existing – not at this bank), purpose of the loan, credit amount, savings account/bonds, present employment since, instalment rate in percentage of disposable income, personal status and sex, other debtors/guarantors, present residence since, property, age in years, other instalment plans, housing, number of existing credits at this bank, job, no. of people being liable to provide maintenance for, telephone, foreign worker. The dependent variable is a categorical one and has two levels: "good" and "bad".

Because the dataset included some categorical inputs, these variables were transformed and represented by dummy variables. This option was required for SVMs only, since the tree based methods (bagging and random forest) can deal with categorical explanatory variables. Thus the objects in the analysed training set for SVMs were described by 7 interval input variables and 54 categorical predictors (some of them – dummy variables).

Results for *bagging*

The procedure pointed out `nbagg`=100, and `minsplit`=4, as the best configuration (with the highest AUC_{10-CV} = 0.6824) for *bagging*. This configuration is exactly the same when compared with parameters' values obtained using ε_{b-CV} as a model selection criterion. In both cases the final model has $\varepsilon_{b-CV} = 0.231$ and $\varepsilon_{TRAIN} = 0.064$.

Results for *random forest*

We obtained the following parameters' values as the outcome of the procedure for *random forest* (AUC_{10-CV} = 0.6825): `ntree`=50, `mtry`=8, `nodesize`=5. The classification errors for the final model are: $\varepsilon_{b-CV} = 0.227$ and $\varepsilon_{TRAIN} = 0.019$. The values of the parameters are different using ε_{b-CV} as a selection criterion: `ntree`=200, `mtry`=6, `nodesize`=2, but the cross-validated classification error of the final model is very similar $\varepsilon_{b-CV} = 0.224$.

Results for *Support Vector Machines*

We obtained the following parameters' values as the outcome of the procedure for *SVMs* (AUC_{10-CV} = 0.6886): kernel=polynomial, degree=2, gamma=0.1, cost=0.1. The classification errors for the final model are: ε_{b-CV} = 0.243 and ε_{TRAIN} = 0.05. This configuration is exactly the same when compared with parameters' values obtained using ε_{b-CV} as a model selection criterion.

4.2. An example of the model selection and evaluation procedure applied to regression problem

To illustrate how the procedure described in Section 3 works in the regression problem we used a real-world dataset flats which was created on the basis of the information published by the portal oferty.net. The data represent a sales transactions from about 16 different real estate agencies in Warsaw. The dataset consists of 990 observations. The general information about the analysed dataset is presented in Table 3.

Table 3. General information about the flats dataset

No. of observations	No. of input variables		
	ratio	ordinal	nominal
990	4	1	2

Source: own results

The set of input variables consist of: distance to the central point of the city, number of rooms, year the property was built in, location (name of the city district), type of the ownership, condition of the apartment. The dependent variable is the price per 1 square meter the estate was sold for. Because of the missing values problem, the dataset used in the analysis was reduced to 747 complete observations. In the analysis with *SVM*, 22 dummy variables were introduced for the nominal ones.

Results for *bagging*

The procedure pointed out nbagg=80, and minsplit=3, as the best configuration (with the highest $pred(0.25)_{b-CV}$ = 0.0755) for *bagging*. In this case the mean squared errors of the final model are: MSE_{b-CV} = 2.1245 and R^2 = 0.7689. The values of the parameters differ from the ones that resulted from using MSE_{b-CV} as a model selection criterion: nbagg=50, and minsplit=2, but the cross-validated mean squared error of the final model is very similar MSE_{b-CV} = 2.1005.

Results for *random forest*

We obtained the following parameters' values as the outcome of the procedure for *random forest* ($pred(0.25)_{b-CV} = 0.0642$): `ntree`=200, `mtry`=4, `nodesize`=2. The mean squared errors for the final model are: $MSE_{b-CV} = 2.0016$ and $R^2 = 0.9278$. The values of the parameters are different using MSE_{b-CV} as a model selection criterion: `ntree`=200, `mtry`=2, `nodesize`=2, but the cross-validated classification error of the final model is again very similar $MSE_{b-CV} = 1.974$.

Results for *Support Vector Machines*

We obtained the following parameters' values as the outcome of the procedure for *SVMs* ($pred(0.25)_{b-CV} = 0.0817$): `kernel=polynomial`, `degree`=2, `gamma`=0.1, `cost`=0.1, `epsilon`=0.1. The classification errors for the final model are: $MSE_{b-CV} = 2.4358$ and $R^2 = 0.5344$. This configuration is exactly the same when compared to parameters' values obtained using MSE_{b-CV} as a model selection criterion.

5. Conclusions

As a consequence of the *No Free Lunch* theorem, the search for the best classification or regression method is pointless (for all possible problems), because such method does not exist. Thus, the choice of modelling method and its parameters must be performed with due care. However, it seems reasonable to use different criterion when tuning the parameters and during the stage of evaluating the final (selected) model. In the paper we present an approach of using area under the ROC curve and prediction quality indicator as a model selection criterion in the first stage, for classification and regression problems respectively, and the standard cross-validated classification error and mean squared error in the latter stage (for classification and regression, respectively). As illustrated by the two examples, this approach can lead to different configuration of model parameters (different models), but the overall predictive ability of the final model does not differ much from the standard and widely used approach of using the same measure for model selection and model evaluation. Both approaches give very similar results and the superiority of any of them cannot be proved (*No Free Lunch* theorem), but the proposed procedure has the methodological advantage, since we use *independent* criteria in the two crucial stages of modelling (model selection and model evaluation phase). If we agree that model evaluation should be performed independently from the stage of building the model (using observations that were not used in the modelling phase and also evaluation criteria that were not used when building the model), then the advantages of the presented procedure become clear.

References

Altman D.G., Bland J.M. (1994), *Statistics Notes: Diagnostic tests 1: sensitivity and specificity*, "British Medical Journal", vol. 308(6943), p. 1552.

Breiman L. (2001), *Random forests*, "Machine Learning", vol. 45(1), pp. 5–32.

Gatnar E. (2008), *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa.

Hastie T., Tibshirani R., Friedman J. (2001), *The Elements of Statistical Learning*, Springer Verlag, New York.

Kitchenham B.A., Pickard L.M., MacDonell S.G., Shepperd M.J. (2001), *What accuracy statistics really measure*, "IEE Proceedings-Software", vol. 148(3), pp. 81–85.

Kohavi R. (1995), *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*, "International Joint Conference on Artificial Intelligence", pp. 1137–1145.

Meyer D., Leisch F., Hornik K. (2003), *The support vector machine under test*, "Neurocomputting", vol. 55(1), pp. 169–186.

Misztal M. (2014), *Wybrane metody oceny jakości klasyfikatorów – przegląd i przykłady zastosowań*, [in:] K. Jajuga, M. Walesiak (eds.), „Taksonomia 23: Klasyfikacja i analiza danych – teoria i zastosowania. Prace Naukowe Akademii Ekonomicznej we Wrocławiu", no. 328, pp. 156–166.

Rozmus D. (2004), *Random forest jako metoda agregacji modeli dyskryminacyjnych*, [in:] K. Jajuga, M. Walesiak (eds.), „Taksonomia 11: Klasyfikacja i analiza danych – teoria i zastosowania. Prace Naukowe Akademii Ekonomicznej we Wrocławiu", no. 1022, pp. 441–448.

Rozmus D. (2008), *Agregacja modeli klasyfikacyjnych i regresyjnych*, Fundacja Promocji i Akredytacji Kierunków Ekonomicznych, Warszawa.

Trzęsiok M. (2006), *Metoda wektorów nośnych na tle innych metod wielowymiarowej analizy danych*, [in:] K. Jajuga, M. Walesiak (eds), „Taksonomia 13. Klasyfikacja i analiza danych – teoria i zastosowania. Prace Naukowe Akademii Ekonomicznej we Wrocławiu", no. 1126, pp. 536–542.

Trzęsiok M. (2007), *Symulacyjne porównanie jakości modeli otrzymanych metodą wektorów nośnych z innymi modelami regresji*, [in:] J. Dziechciarz (ed.), *Zastosowanie metod ilościowych,* „Prace Naukowe Akademii Ekonomicznej we Wrocławiu", no. 1189, Wrocław, pp. 234–241.

Vapnik V. (1998), *Statistical Learning Theory*, John Wiley & Sons, New York.

Wolpert D.H., Macready W.G. (1997), *No Free Lunch Theorems for Optimization*, "IEEE Transactions on Evolutionary Computation", vol. 1, pp. 67–82, doi: 10.1109/4235.585893.

Wybrane metody pomiaru jakości modeli statystycznych

Streszczenie: Bardzo ważnym elementem procesu modelowania statystycznego jest etap oceny jakości zbudowanego modelu. W zależności od wykorzystanej metody istnieje wiele różnych podejść do pomiaru jakości modelu. Pomiar ten może skupiać się na dopasowaniu do danych empirycznych albo może przede wszystkim uwzględniać zdolności prognostyczne modelu. Mierniki mogą być absolutne albo względne. Zestaw mierników jakości modelu obejmuje liczną grupę propozycji, z których analityk musi wybrać najodpowiedniejszy do danej sytuacji. W artykule przedstawiono zestawienie mierników jakości modelu oraz sugestię używania innych mierników jakości na etapie wyboru wariantu modelu oraz na etapie oceny jakości modelu końcowego.

Słowa kluczowe: jakość modelu, dopasowanie, błąd predykcji

JEL: C150, C180, C300, C310, C380

Received: 2017-01-14; verified: 2018-02-14. Accepted: 2019-01-28