AUSTRIAN JOURNAL OF STATISTICS
Volume 35 (2006), Number 2&3, 325–330

Web-Bootstrap Estimate of Area Under ROC Curve

Hana Skalská and Václav Freylich
University of Hradec Kralove, Czech Republic

Abstract: The accuracy of binary discrimination models (discrimination between cases with and without any condition) is usually summarized by classification matrix (also called a confusion, assignment, or prediction matrix). Receiver operating characteristic (ROC) curve can visualize the association between probabilities of incorrect classification of cases from the group without condition (False Positives) versus the probabilities of correct classification of cases from the group with condition (True Positives) across all the possible cut-point values of discrimination score.

Area under ROC curve (AUC) is one of summary measures. This article describes the possibility of AUC estimate with the use of web based application of bootstrap (resampling). Bootstrap is useful mainly to data for which any distributional assumptions are not appropriate. The quality of the bootstrap application was evaluated with the use of a special programme written in $C\sharp.NET$ language that allows to automate the process of repeating different experiments. Estimates of AUC and confidence limits given by bootstrap method were compared with bi-normal and nonparametric estimates. Results indicate that usually bootstrap confidence intervals are narrower than nonparametric one, mainly for small data samples.

Keywords: Discrimination, AUC Estimate, Resampling.

1 Introduction

Classification solves the problem of classifying the individuals into one of several categories on the basis of a number of measurements made on each of individuals. An individual is a random observation given from any of finite number of categories or populations. Solutions of the problem of classification can be given as statistical decision problems (linear or nonlinear discrimination rule, logistic regression, Bayesian methods) as well as semi statistical or non-statistical (decision trees, decision rules, neural networks, logic programming approaches, multi-criteria classification methods, etc.). In the process of classification it is desirable to minimize on the average the bad effects of misclassification (usually the average costs of misclassification). Classification algorithms can be optimized with respect to costs of misclassification and probabilities of classes. When these external conditions changes then the classification model can lose their optimality. Therefore an evaluation of different models given within the full range of external conditions is necessary. We will assume the problem of binary classification when only two groups (classes) are admitted. One of the groups is assumed without any condition (negative) and the second one with any condition present (positive one). Then ROC curve can be used for description of accuracy the classification model. This article describes bootstrap estimate of area under ROC curve (AUC) and confidence limits for AUC. The web application was prepared with this possibility. Results are compared with ROCkit

and NCSS software. The advantage of the solution is in that bootstrap is distribution free and allows to reach narrower confidence limits no matter the size of the samples is small.

2 Accuracy of Binary Classification Models

Predictive accuracy of discrimination model can be estimated in proportions of correctly classified cases that is based on training or test sample. Apparent error rate (APER) is calculated from confusion matrix. APER is based on re-substitution of learning sample and usually underestimate the future proportion of misclassification. Unbiased estimator can be achieved through jackknife (leaving-one-out) or cross-validation on training sample. In constructing classification matrix, the optimum cutting score should be determined. Receiver operating characteristic (ROC) curve is a set of the points with coordinates given with a probability of incorrect classification of cases from the group without condition (often called FP rate of False Positives) versus the probability of correct classification of cases from the group with condition (TP rate of True Positives) across all the possible cut-point values of discrimination score. ROC describes and visualizes all possible confusion matrices. As the optimal cutting point depends on external conditions (probabilities of the groups and costs of misclassification) than ROC describes the quality of classification model under all possible external conditions.

Area under receiver operating characteristic curve (AUC) is one of the summary measures (Hanley and McNeil, 1982). The AUC can take values between 0.0 and 1.0 with practical lower bound value 0.5 (chance diagonal). The AUC can be interpreted as the probability that an object randomly selected from the group with the condition will have discrimination score indicating greater suspicion than that of a randomly selected (from the group) without condition.

In the literature, estimation of AUC may be based either on parametric model (very often bi-normal distribution is assumed) or on nonparametric approach (mostly with the use of Wilcoxon statistic). Statistical software packages in latest versions enable this type of analysis (NCSS, SPSS, Statistica, etc.) Mostly estimates of AUC are based on trapezoidal rule or assume bi-normal distribution. Software RocKit prepared by Metz (2003) is specialized in ROC and AUC, calculates maximum-likelihood estimates of the parameters of a bi-normal model and statistical significance of the difference between two ROC curves.

2.1 Bootstrap Estimate of AUC

There are different methods of rearranging a given data set. The basic idea of resampling (bootstrap) is that observed sample is considered to be the population. Resampling uses many samples taken from a single sample given from the population of interest. The probability distribution of statistic is simulated by random samples from original sample. This way bootstrap allows estimation of variance of a statistic. We assume two groups of objects G_0 (without condition), and G_1 (with any condition) and ordinal or quantitative discriminating variable (or discrimination score) X. We will assume that the smaller value of variable X is in association with largest probability that the object belongs to group G_0.

Suppose that we have the random samples of sizes n_0 and n_1 from the groups G_0 and G_1 respectively: $x = (x_1^{(0)}, \ldots, x_{n_0}^{(0)}, x_1^{(1)}, \ldots, x_{n_1}^{(1)}) = (x^{(0)}, x^{(1)})$. The unknown parameter to be estimated is θ. The statistic AUC is considered to be an estimate of θ. Properties of AUC will be estimated by the use of bootstrap samples $\{x_1^*, \ldots, x_k^*\}$ where $x_i^* = (x_{1*}^{(0)}, \ldots, x_{n_0*}^{(0)}, x_{1*}^{(1)}, \ldots, x_{n_1*}^{(1)})$ for $i = 1, \ldots, k$. Each bootstrap sample of size n consists of two bootstrap samples of sizes n_0, n_1, $n_0 + n_1 = n$ and is chosen from the original samples $x^{(0)}$ and $x^{(1)}$ with replacement at each of them. Estimates $\{AUC_1^*, \ldots, AUC_k^*\}$ are of the same functional form as the original estimator (here calculated on each bootstrap sample via trapezoidal rule).

Then specifically nonparametric Monte Carlo (Gentle, 2002; Prášková, 2004) estimate $\overline{AUC}$ was used as an unbiased estimator of θ. Its distribution is related to the distribution of AUC and the estimate is given with $\overline{AUC}^* = k^{-1} \sum_{i=1}^{k} AUC_i^*$.

Bootstrap confidence limits were given as the bootstrap percentile confidence intervals. If $F_{AUC^*}(t)$ is distribution function of AUC^*, then the upper $(1 - \alpha)$ confidence limit for θ is the value $AUC_{(1-\alpha)}^*$ such that $F_{AUC^*}(AUC_{(1-\alpha)}^*) = 1 - \alpha$. The lower limit AUC_α^* is such a value that $F_{AUC^*}(AUC_\alpha^*) = \alpha$. Then a $(1-\alpha)\cdot 100\%$ confidence interval is given as $(AUC_{[\alpha/2]}^*, AUC_{[1-\alpha/2]}^*)$, where $AUC_{[\alpha/2]}^*$ is the $k \cdot (\alpha/2)$th and $AUC_{[1-(\alpha/2])}^*$ is the $k \cdot (1 - (\alpha/2)$th order statistic of the sample of size k of AUC^*.

In this way confidence limits can be non-symmetrical. Also they will retain the range of possible values for AUC (going from 0 to 1).

2.2 An Application of Bootstrap Estimate of AUC

A bootstrap application of the method was developed by the authors. This application is accessible at `http://www.freccom.cz/stomo/input.php`. User can input his (her) own data and can change some bootstrap settings (e.g. the number of bootstrap samples).

Tests of the quality of this bootstrap application required the use of large number of sets of samples. Therefore the programme written in $C\sharp.NET$ language working under .NET framework version 2.0 was prepared. This programme runs on personal PC and allows to repeat automatically the different batches of bootstrap AUC estimates. It offers calculation of empirical and bootstrap estimates based on data values that are stored in database table. This database uses three attributes: value of discriminating variable, group ID, and sample set ID.

2.3 Examples

Example 1. Different classification models were compared in the article Skalská (2003) for prediction of good and bad loans from financial data set. This data is accessible at `http://lisp.vse.cz/pkdd99/` and it is described in Berka (2001). Here we compare estimates given with ROCkit software (nonparametric and bi-normal AUC) with estimates from our web based bootstrap application. Two discrimination models are compared here, linear discrimination function (LDF) and logistic regression (LR).

Estimates of AUC (samples of sizes $n_0 = 203$ and $n_1 = 31$) are summarized in the Table 1. Confidence interval for nonparametric (Wilcoxon) AUC should to be calculated

extra via Fisher transformation of AUC and then under assumption of normality of transformed AUC (it is not given here as this is not included in the software). As an example of empirical distribution of discriminating variable X, Figure 1 visualizes the distribution of discriminating score variable of LR model conditioned G_0 and G_1. The ROC curves on Figure 1 indicate the dominance of LR model under LDF model. Asymmetry in distribution of X in G_0 and non-normality in G_1 are characteristics of these distributions. Skewness equals 2.5 (2.9) in G_0 and 0.7 in G_1 for models LR (LDF respectively).

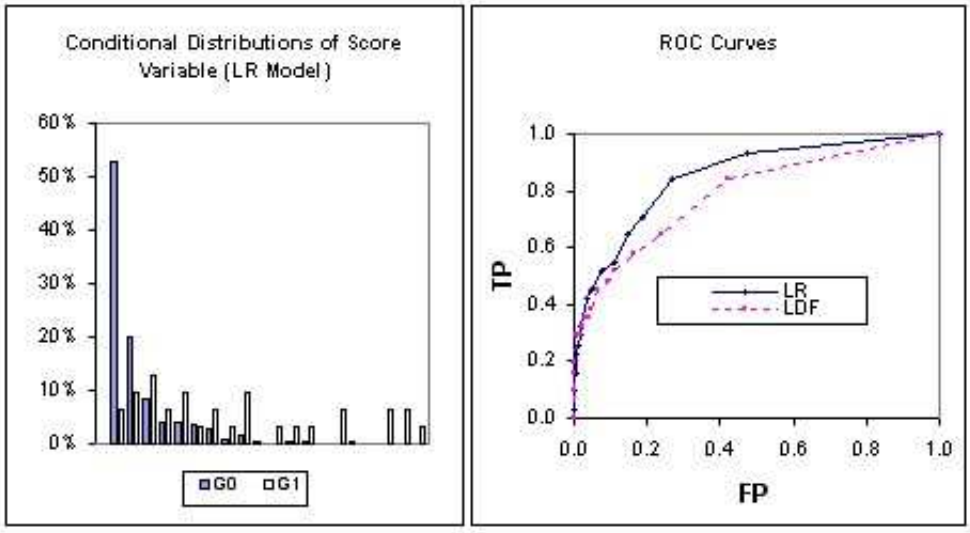

Figure 1: Distribution of X for LR model (conditioned G_0 and G_1) and ROC

Small differences can be seen in Table 1 among different AUC estimates. Confidence interval limits for nonparametric (Wilcoxon) AUC are not included in ROCkit. This is disadvantage mainly for small sample sizes. Bootstrap confidence limits are narrower than that given under bi-normal assumption.

Table 1: AUC estimates with 95% confidence limits

Bootstrap		ROCkit				
	$k = 3000$		Bi-normal assumption		Wilcoxon	
Model	AUC	95% CI	AUC	95% CI	AUC	SE
LR	0.8607	$0.7928 - 0.9184$	0.8595	$0.7714 - 0.9211$	0.8625	0.0430
LDF	0.8221	$0.7409 - 0.8946$	0.8238	$0.7246 - 0.8968$	0.8198	0.0476

Example 2. This example compares estimates from our bootstrap application with NCSS estimates Hintze (2005). The samples from Gamma, respectively Beta distributions with different parameters (Table 2) were prepared. The comparison of estimates is based on averages from 100 sample statistics resulting from the samples of sizes $n_0 = 30$, or $n_0 = 100$ respectively, and $n_1 = 30$. Each set consists of 100 samples from G_0 and 100 samples from G_1, ID of the group and ID of the sample. The sets of the samples used for bootstrap and NCSS were the same within each experiment, but the sets were different

Table 2: Description of distributions used in experiments.

Trials	Model (G_0)	Skewness	Model (G_1)	Skewness	AUC trapezoidal
A, D	Gamma(2.0, 1.0)	1.5	Gamma(3.0, 1.0)	+1.3	0.6893
B, E	Beta(1.5, 3.0)	0.5	Beta(3.0, 1.5)	-0.6	0.8725
C, F	Beta(1.5, 3.0)	0.5	Beta(5.0, 1.5)	-0.9	0.9450

Table 3: Comparison between bootstrap and NCSS (averages from 100 samples)

	Bootstrap AUC		NCSS 2004			
	$k = 3000$		Bi-normal		AUC	Nonparametric
$n_0 = 30$					Empi-	
$n_1 = 30$	AUC	95% CI	AUC	95% CI	rical	95% CI
A	0.7032	$0.5940 - 0.8059$	0.6814	$0.5286 - 0.7906$	0.6949	$0.5362 - 0.8055$
B	0.8893	$0.8192 - 0.9478$	0.8845	$0.7747 - 0.9417$	0.8774	$0.7589 - 0.9388$
C	0.9595	$0.9215 - 0.9879$	0.9577	$0.8886 - 0.9837$	0.9484	$0.8648 - 0.9800$
$n_0 = 100$						
$n_1 = 30$						
D	0.6874	0.5975-0.7726	0.6721	0.5447-0.7685	0.6892	0.5698-0.7798
E	0.8753	0.8136-0.9294	0.8803	0.7945-0.9313	0.8721	0.7803-0.9267
F	0.9497	0.9156-0.9777	0.9554	0.9095-0.9782	0.9451	0.8882-0.9732

for each of experiments A to F. The computing time varied from 100 to 200 seconds for each of experiments. Description of populations that is based on samples of sizes 10000 (G_0) and 3000 (G_1) is given in Table 2. Results of experiments are summarized in Table 3. Each value is an average given from 100 samples. On average bootstrap estimates have narrower confidence intervals.

3 Conclusion

Bootstrap estimate of AUC, area under receiver operating characteristic curve was described here and compared with other commonly used methods of estimate. Web based application of the bootstrap method was prepared and the results were compared with other methods from NCSS and ROCkit software. Also another application for PC was developed that uses bootstrap routine and allows running in a batch mode and repeat experiments. Results of different experiments are described here. Examples present the data for which bi-normal assumption is not appropriate.

It can be concluded that bootstrap estimates give very similar results with bi-normal estimates. On average bootstrap provides narrower confidence limits than bi-normal and nonparametric estimates do. Bootstrap can probably be more useful mainly when distributions are strongly skewed and sizes of samples are small. Our experiments do not indicate substantial differences of bootstrap estimates from estimates based on bi-normal

assumption. This may indicate the robustness of bi-normal estimate to a wide variety of frequency distributions.

Acknowledgement

We acknowledge the support by the Grant Agency of Czech Republic under grant No 402/04/1308.

References

Berka, P. (2001). Discovery challenge: modelová data dobývání znalostí z realistických dat. In *Proceedings Znalosti 2001* (p. 547-553). Prague: VšE.

Gentle, J. E. (2002). *Elements of Computational Statistics*. New York: Springer Verlag.

Hanley, J. A., and McNeil, B. J. (1982). The meaning and use of the areas under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29-36.

Hintze, J. (2005). NCSS. In *NCSS and PASS. Number Cruncher Statistical Systems.* Kayswille, Utah: www.ncss.com.

Metz, C. E. (2003). ROCkit. In *http://xray.bsd.uchicago.edu.* University of Chicago: Kurt Rossmann Laboratories.

Prášková, Z. (2004). Metoda bootstrap. In *Robust 2004* (p. 299-314). Prague: JČMF.

Skalská, H. (2003). Software tools for ROC and AUC estimates. In *Proceedings of the 21th international conference MME 2003* (p. 238-243). Prague: VŠE.

Author's address:

Hana Skalská
Department of Informatics and Quantitative Methods
University of Hradec Králové
Rokitanského 63
CZ-500 03 Hradec Králové
Czech Republic

E-mail: hana.skalska@uhk.cz