Honest Evaluation of Classification Models

Jose A. Lozano*, Guzman Santafe†, Iñaki Inza‡
Intelligent Systems Group, University of the Basque Country (Spain)
http://www.sc.ehu.es/isg

1 Summary

Supervised classification is a part of machine learning that has grown in interest over the last years. In the literature, there are many proposals for classification paradigms and learning algorithms that can be applied to a specific classification task. Therefore, an honest classifier evaluation and a fair comparison among classification models are key points in order to obtain right conclusions about the results achieved as well as to choose the best model/paradigm to deal with a classification task. However, there are many researchers that focus their work on proposing new classification algorithms, leaving the honest evaluation of the results aside.

This tutorial aims to offer an overview of honest performance evaluation methodology for classifiers with detailed information about scores to measure the goodness of classification models, estimation methods and hypothesis tests to carry out as fair comparison as possible among different models. Thus, we provide researcher with sufficient information to choose the best alternative for each specific problem in order to obtain the fairest conclusions. Since, there is no single literature source covering all the aspects of the evaluation process, we think this tutorial may help the research community to have a better understanding of the evaluation methodologies. Additionally, the tutorial provides some useful guidelines to apply these evaluation methodologies to real problems.

The tutorial is organized in four parts. In the first part we introduce the classification problem and we motivate the relevance of a honest evaluation of classification models as well as the model comparison [Hatie et al., 2001, Michell, 1996]. The second part is devoted to the scores that can be used to measure the goodness of a classifier. Classification error is the most studied score [Bengio and Grandvalet, 2005, James, 2003, Rodríguez et al., 2010] and also the most commonly used to evaluate classification models. However there are other scores that may be of interest in certain application domains [Fawcett, 2006, Hand and Till, 2001, Makhoul et al., 1999, Saracevic, 1996, van Rijsbergen, 1979]. In this part of the tutorial we analyze the main characteristics and properties of different scores: classification error, recall, specificity, precision, balanced accuracy, balanced error, f-score and area under the ROC curve; and different application domains for these scores. The third part of the tutorial is related to the estimation methods. We present and motivate the problem of estimating the value of a score for a classifier given a (finite) data set and the bias and variance problem for the score estimation [Domingos, 2000, Friedman, 1997, Geman et al., 1992, James, 2003, Kohavi and Wolpert, 1996]. Then, we elaborate on different estimation methods such as resubstitution [L. P. Devroye, 1979], hold-out (and variants) [Larson, 1931, McLachlan, 1992], cross-validation (and variants) [Stone, 1974], bootstrap (and variants) [Efron and Tibshirani, 1993, Efron and Tibshirani, 1997], their properties and some application domains [Efron and Tibshirani, 1986, Burman, 1989, Michael R. Chernick, 2008, Kim, 2009, Rodríguez et al., 2010] and additionally, some problems that can arise when dealing with some special characteristics of the data set [Braga-Neto and Dougherty, 2004, Isaksson et al., 2008, Ojala and Garriga, 2010].

Finally, the fourth part of the tutorial is dedicated to classifier comparison. In this part we introduce statistical hypothesis testing and different types of statistical tests such as T-test, 5X2 CV test, Wilcoxon tests, Man-Whitney test, McNemar's test, etc. that can be used to compare two classification models using one or more data sets and Friedman test + post-hoc that can be used to compare multiple classification models using several data sets [Shaffer, 1995, Dietterich, 1998, Alpaydin, 1999, Nadeau and Bengio, 2003, Demsar, 2006, Garcia and Herrera, 2008, García et al., 2010]. Additionally, each part of the tutorial presents recommendations to perform honest classifier evaluation according to specific characteristics of the problem or the data set as well as general best practices in the use of the presented methodology.

References

[Alpaydin, 1999] Alpaydin, E. (1999). Combined 5 x 2 cv F test for comparing supervised classification learning algorithms. *Neural computation*, 11(8):1885–92.

*email: ja.lozano@ehu.es
†email: guzman.santafe@ehu.es
‡inaki.inza@ehu.es

[Bengio and Grandvalet, 2005] Bengio, Y. and Grandvalet, Y. (2005). *Bias in Estimating the Variance of k-fold Cross-Validation*, chapter 5, pages 75–95. Springer.

[Braga-Neto and Dougherty, 2004] Braga-Neto, U. M. and Dougherty, E. R. (2004). Is cross-validation valid for small-sample microarray classification? *Bioinformatics (Oxford, England)*, 20(3):374–80.

[Burman, 1989] Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3):503–514.

[Demsar, 2006] Demsar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7:1–30.

[Dietterich, 1998] Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, 10(7):1895–1923.

[Domingos, 2000] Domingos, P. (2000). A unified bias-variance decomposition and its applications. In *17th International Conference on Machine Learning*, number x, pages 231–238.

[Efron and Tibshirani, 1986] Efron, B. and Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistics*, 1(1):54–77.

[Efron and Tibshirani, 1993] Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

[Efron and Tibshirani, 1997] Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: The. 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548– 560.

[Fawcett, 2006] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

[Friedman, 1997] Friedman, J. (1997). On Bias, Variance, 0/1—Loss, and the Curse-of-Dimensionality. *Data Mining and Knowledge Discovery*, 1:55–77.

[García et al., 2010] García, S., Fernández, A., Luengo, J., and Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences*, 180(10):2044–2064.

[Garcia and Herrera, 2008] Garcia, S. and Herrera, F. (2008). An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons. *Journal of Machine Learning Research*, 9:2677–2694.

[Geman et al., 1992] Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural Networks and the Bias/Variance Dilema. *Neural Computation*, 4:1–58.

[Hand and Till, 2001] Hand, D. A. and Till, R. J. (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, 45:171–186.

[Hatie et al., 2001] Hatie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer.

[Isaksson et al., 2008] Isaksson, A., Wallman, M., Goransson, H., and Gustafsson, M. (2008). Cross-validation and bootstrapping are unreliable in small sample classification. *Pattern Recognition Letters*, 29(14):1960–1965.

[James, 2003] James, G. (2003). Variance and bias for general loss functions. *Machine Learning*, 51 (2)(1998):115–135.

[Kim, 2009] Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, 53(11):3735–3745.

[Kohavi and Wolpert, 1996] Kohavi, R. and Wolpert, D. H. (1996). Bias Plus Variance Decomposition for Zero-One Loss Functions. In Saitta, L., editor, *International Conference on Machine Learning*, pages 275–283. Morgan Kaumann.

[L. P. Devroye, 1979] L. P. Devroye, T. J. W. (1979). Distribution-free performance bounds with the resubstitution error estimate. *IEEE Transactions on Information Theory*, 25(2):208–210.

[Larson, 1931] Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology*, 22:45–55.

[Makhoul et al., 1999] Makhoul, J., Kubala, F., Schwartz, R., and Weischedel, R. (1999). Performance measures for information extraction. In *In Proceedings of DARPA Broadcast News Workshop*, pages 249–252.

[McLachlan, 1992] McLachlan, G. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons.

[Michael R. Chernick, 2008] Michael R. Chernick (2008). *Bootstrap Methods. A Guide for Practitioners and Researchers*. Wiley & Sons.

[Michell, 1996] Michell, T. M. (1996). *Machine Learning*. McGraw Hill.

[Nadeau and Bengio, 2003] Nadeau, C. and Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, (1995):239–281.

[Ojala and Garriga, 2010] Ojala, M. and Garriga, G. C. (2010). Permutation Tests for Studying Classifier Performance. *Journal of Machine Learning Research*, 11:1833–1863.

[Rodríguez et al., 2010] Rodríguez, J. D., Pérez, A., and Lozano, J. A. (2010). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):569–75.

[Saracevic, 1996] Saracevic, T. (1996). Evaluation of Evaluation in Information Retrieval. In *18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 138–146.

[Shaffer, 1995] Shaffer, J. (1995). Multiple Hypothesis Testing. *Annual Review of Psychology*, 46:551–584.

[Stone, 1974] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society Series B*, 36:111–147.

[van Rijsbergen, 1979] van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworth-Heinemann.