Position papers of the Federated Conference on Computer Science and Information Systems, pp. 37–40
DOI: 10.15439/2017F530
ISSN 2300-5963 ACSIS, Vol. 12

Evaluation of classifiers: current methods and future research directions

Katarzyna Stąpor
Silesian University of Technology
ul. Akademicka 16, 44-100 Gliwice, Poland
Email: Katarzyna.Stapor@polsl.pl

***Abstract*—This paper aims to review the most important aspects of the classifier evaluation process including the choice of evaluating metrics (scores) as well as the statistical comparison of classifiers. Some recommendations, limitations of the described methods as well as the future, promising directions are presented. This article provides a quick guide to understand the complexity of the classifier evaluation process and tries to warn the reader about the wrong habits.**

I. Problem description

LEARNING A CLASSIFIER from a dataset of labeled data instances taken from unknown distribution where each instance is characterized by a feature vector and a class to which it belongs is a central task of supervised classification. A learned classifier is a function mapping whole feature space into a label space. Then, the learned classifier, after evaluation its quality, can be used to classify new samples with unknown class label. There are many classification paradigms/models: the detailed description of the supervised classification problem can be found in books on machine learning (see for example [1]). The following question usually arises: "which is the best classification paradigm for a given problem?". Answering this question requires the evaluation as well as the comparison of the many candidate models. Usually, the problem of classifier evaluation is performed by using the scores that try to summarize the specific conditions of classifier behavior. The examples of such scores are classification error or accuracy. It is now generally agreed that the whole evaluation process of a classifier should include the following steps ([4], [5], [6], [10], [11], [12]):

1) choosing the score(s) according to the properties of the classifier as well as the domain objectives,
2) choosing the score estimation method,
3) choosing the statistical test,
4) choosing the datasets,
5) running the evaluation.

The main purpose of this paper is to provide the reader with a better understanding about the overall classifier evaluation process. As there is no fixed, concrete recipe for the classifier evaluation procedure, we believe that this paper will facilitate the researcher in the machine learning area to decide which alternative to choose for each specific case.

This paper is focused only on a supervised classification problem as defined in the beginning. Other types of classification such as classification from data streams or multi-label classification are not addressed here, since they may impose specific conditions to the calculation of the score.

The paper is set up as follows. In section 2 till 4 we shortly present the mentioned steps of classifier evaluation. In section 5 we conclude giving some recommendations and propose new, future directions for classifier evaluation methodology.

TABLE I
Confusion matrix for a two-class problem

	Predicted positive	Predicted negative
Positive class	True Positive (TP)	False Negative (FN)
Negative class	False Positive (FP)	True Negative (TN)

II. Choosing classifier scores

Typical scores for measuring the performance of a classifier are accuracy and classification error, which for a two-class problem can be easily derived from a 2x2 confusion matrix as that given in table reftable1. These scores can be computed as:

$$Acc = (TP + TN)/(TP + FN + TN + FP)$$

$$Err = (FP + FN)/(TP + FN + TN + FP)$$

Empirical evidence shows that accuracy and error rate are biased with respect to data imbalance: the use of these scores might produce misleading conclusions since they are strongly biased to favor the majority class, and are sensitive to class skews.

In some application domains, we may be interested in how our classifier classifies only a part of the data, i.e. positive or negative data samples. Examples of such measures are: True positive rate (Recall or Sensitivity): $TPrate = TP/(TP + FN)$, True negative rate (Specificity): $TNrate = TN/(TN + FP)$, Precision $= TP/(TP + FP)$.

Each entry in the confusion matrix may be misleading by two confounding issues: asymmetric misclassification costs and asymmetric class distributions. Shortcomings of the accuracy or error rate have motivated search for new balanced measures which aim to obtain a trade-off between the evaluation of the classification ability on both positive and negative data samples. Some straightforward examples of such alternative scores are: the arithmetic, geometric or harmonic means between Recall and Specificity. They give the same relevance to both components. There are other proposals that

try to enhance one of the two components of the mean. For instance, Index of Balanced Accuracy [7]:

$$IBA_\alpha = (1 + \alpha(TPrate - TNrate)) \times TPrate \times TNrate$$

and *F-score* [14]:

$$\textit{F-score}_\beta = \frac{(\beta^2 + 1)Precision \times Recall}{\beta^2 \times Precision + Recall}$$

The parameters α, β can be tuned to obtain different trade-offs between both components.

The cost matrix can be used if the severity of misclassifications can be quantified in terms of costs and then, to weight the entries in the confusion matrix. When the classification costs cannot be accessed, the above mentioned balanced scores may be used to set more relevance to the costliest misclassification. Another most widely-used technique in this case is the ROC curve [3]. However, recent studies have shown that AUC (Area under the ROC curve) is a fundamentally incoherent measure since it treats the costs of misclassification differently for each classifier. This is undesirable because the cost must be a property of the problem, not of the classification method. In [8], the H measure has been proposed as an alternative to AUC.

Ground truth assumption states that the true class labels of data samples are deterministically known even though they are the result of an arbitrary unknown distribution that a classifier aims to approximate. This make it impossible to take into account that correct classification could be a result of coincidental concordance between classifier's output and label-generation process. Cohen's kappa statistics corrects for this problem:

$$\kappa = \frac{P_o - P_o^c}{1 - P_o^c}$$

where P_o represents the probability of overall agreement over the label assignments between the classifier and the true process, and P_o^c represents the chance agreement over the labels as is defined as the sum of the proportion of examples assigned to a class times the proportion of true labels of that class in the dataset.

Performance measures for multi-class classification are still an open research topic. Generally, the two approaches are commonly used. Macroaveraging (per category) takes the average of measures on separate classes:

$$B_{macro} = \frac{1}{n}\sum_{i=1}^{n} B(TP_i, FP_i, FN_i, TN_i)$$

where B is a binary score. Microaveraging (per case) sums up individual TP, FP, FN, TN for different classes and then apply to get a measure:

$$B_{micro} = B(\sum_{i=1}^{n} TP_i, \sum_{i=1}^{n} FP_i, \sum_{i=1}^{n} FN_i, \sum_{i=1}^{n} TN_i)$$

There is no complete agreement among the authors on which is better. In this paper, we focus on the scores since they are popular way to measure classification quality. But these measures do not capture all the information about the quality of classification methods some graphical methods may do. The presented list of scores is by no means exhaustive. There are other important aspects of classification such as robustness to noise, scalability, stability under data shifts, etc. which are not addressed here.

III. Choosing score estimation method

Various re-sampling methods are commonly used to estimate the classifier scores (the review of re-sampling methods can also be found in the mentioned literature on machine learning). The most commonly used k-fold cross-validation (CV) creates a k-fold partition of the entire dataset once. Then, for each of k experiments, it uses (k-1) folds for training and a different fold for testing. The classification error is estimated as the average of separate errors obtained from k experiments. In order to obtain more stable estimates, it is useful to perform multiple runs of simple re-sampling schemes. Two specific schemes has been suggested: 5x2CV and 10x10CV.

The danger of re-sampling is that it is usually followed by statistical testing which relies on the fundamental assumption that the data used to obtain the sample must be independent. In re-using the data, this important assumption is broken and the results of the statistical test are invalid.

IV. Choosing statistical test

In most situations, the statistical assessment of the observed classifier scores such as hypothesis testing is required. For the comparison of two classifiers on one dataset, the corrected resampled t test has been suggested in the literature [2]. This test is associated with a repeated estimation method: in i-th of the m iterations, a random data partition is conducted and the values for the scores $A_{k1}^{(i)}$ and $A_{k2}^{(i)}$ of compared classifiers $k1$ and $k2$, are obtained. The statistic is:

$$t = \frac{\overline{A}}{\sqrt{\left(\frac{1}{m} + \frac{N_{test}}{N_{train}}\right) \cdot \sum_{i=1}^{m} \frac{\left(A^{(i)} - \overline{A}\right)^2}{m-1}}}$$

where $\overline{A} = \frac{1}{m}\sum_{i=1}^{m} A^{(i)}$, $A^{(i)} = \left(A_{k1}^{(i)} - A_{k2}^{(i)}\right)$, N_{test}, N_{train} are the number of samples in the test and train partitions. A non-parametric alternative for comparing two classifiers that is suggested in the literature is McNemar's test [9].

For the comparison of two classifiers on multiple datasets the Wilcoxon signed-ranks test [9] is widely recommended. It ranks the differences $d_i = A_{k1}^{(i)} - A_{k2}^{(i)}$ between scores of two classifiers $k1$ and $k2$ obtained on i-th of N datasets, ignoring the signs. The test statistic of this test is:

$$T = \min(R^+, R^-)$$

where:

$$R^+ = \sum_{d_i>0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i),$$

$$R^- = \sum_{d_i<0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i)$$

are the sums of ranks on which the $k2$ classifier outperforms $k1$, respectively. Ranks $d_i = 0$ are split evenly among the sums.

Comparison among multiple classifiers on multiple datasets, the general recommended methodology is as follows. First, we apply an omnibus test to detect if at least one of the classifiers performs different than the others. Friedman nonparametric test [9] with Iman-Davenport extension is probably the most popular omnibus test. It is a good choice when comparing more than five different classifiers. Let R_{ij} be the rank of the j-th of K classifiers on the i-th of N data sets and

$$R_j = \frac{1}{N}\sum_{i=1}^{N} R_{ij}$$

is the mean rank of the j-th classifier. The test compares the mean ranks of the classifiers and is based on the test statistic:

$$F_F = \frac{(N-1)\chi_F^2}{N(K-1)-\chi_F^2}$$

$$\chi_F^2 = \frac{12N}{K(K+1)}\left[\sum_{j=1}^{K} R_j^2 - \frac{K(K+1)^2}{4}\right]$$

which follows an F distribution with $(K-1)$ and $(K-1)(N-1)$ degrees of freedom.

For the comparison of five or less different classifiers, Friedman aligned ranks [9] is a more powerful alternative.

Second, if we find such a significant difference, then we apply a pairwise test with the corresponding post-hoc correction for multiple comparisons to control the family-wise error [13]. For the described above Friedman test, comparing the r-th and s-th classifiers is based on the mean ranks and has the form:

$$z = \frac{R_r - R_s}{\sqrt{\frac{K(K+1)}{6N}}}$$

The z value is used to find the corresponding probability from the table of normal distribution, which is then compared with an appropriate significance level α. There are multiple proposals in the literature to adjust the significance level α: for example, Holm, Hochberg, Finner [9].

V. Dataset selection

The commonly accepted approach in classifier evaluation methodology is to use benchmark datasets as a representation of all the classification problems that can arise in reality and then, to demonstrate that one classifier is, on average, better than the others. However, such representation assumption is questionable as well as the following conclusions (i.e. the generalization to unseen problems).

No free lunch theorem [15] states that for any two classifiers, there are as many classification problems for which the first classifier performs better than the second as vice versa. Thus, it does not make sense to demonstrate that one classifier is, on average, better than the others. Instead, we should focus our attention on exploring the conditions of the classification problems which make our classifier to perform better or worse than others. Additionally, artificially generated datasets may easily reproduce the specific conditions of interest.

VI. Recommentations and future directions

The evaluation of classification performance is very important to the construction and selection of classifiers. Below, we give some recommendations and limitations of the presented methods for classifier evaluation. We also try to define new promising research directions.

- There are many scores for evaluating classifiers: generally, you shouldn't take any of the existing scores in an isolated way. No single metric is capable of encapsulated all the aspects of interest. Multiple metrics need to be reported to detail classifier's performance even for a single aspect of interest. There is not a best way to evaluate any system, but different scores give us different and valuable insights into how a classification model performs. *Many research efforts should be undertaken to investigate principles of combining these scores to yield a summary measure*.
- The vast majority of the published articles use the accuracy (or classification error) as the score in the classifier evaluation process. But these two scores may be appropriate only when the datasets are balanced and the misclassification costs are the same for false positives and false negatives. In the case of skew datasets, which is rather typical situation, the accuracy/error rate is questionable and other scores, especially balanced scores such as Index of Balanced Accuracy, F-score, geometric or harmonic means, H measure are more appropriate. *New methods that aim to obtain a trade-off between the evaluation of the classification ability on both positive and negative classes are need to be developed*.
- Ground truth assumption make it impossible to take into account that correct classification could be a result of coincidental concordance between classifier's output and label-generation process. Cohen's kappa is the simplest measure that corrects for this problem. *New, better chance-corrected measures of the validity of classifiers are needed.*
- In the case of multi-class classification, generally, macroaveraging can be bad practice in cases that there is a considerable difference in number of examples of each class label. Actually, the majority believe that class examples should indeed count proportionally to their frequency, and thus lean towards microaveraging. But, there is no complete agreement among authors on which is better. *Performance measures for multi-class classification are still an open research topic and many empirical investigations are needed*.
- k-fold cross-validation is the best known resampling technique which is commonly used in score estimation. Through high overlapping in the training folds, main independence assumption of many statistical tests used further for statistical comparison is not fulfilled. *This can*

affect the bias of the classifier score and requires new, corrected versions of classical statistical tests which still should be developed.

- In order to obtain more stable estimates of classifier performance, it is useful to perform multiple runs of simple re-sampling schemes. Two such schemes are recommended: 5x2CV and 10x10CV. *More experiments on different schemes are needed to investigate replicability of the results*.
- The comparison of two classifiers on a single dataset is generally unsafe due to the lack of independence between the obtained score values. *Thus, the new corrected versions of the resampled t test or t test for repeated cross-validation are more appropriate*. McNemar's test, being non-parametric, does not make the assumption about distribution of the scores but it does not directly measure the variability due to the choice of the training set nor the internal randomness of the learning algorithm.
- When comparing two classifiers on multiple datasets (especially from different sources), the measured scores are hardly commensurable. Therefore, the *Wilcoxon signed-rank test* is more appropriate.
- Regarding the comparison of multiple classifiers on multiple datasets, if the number of classifiers involved is higher than five, the use of the Friedman test with Iman and Davenport extension is recommended. When this number is low, four or five, Friedman aligned ranks and the Quade test are more useful. If the null hypothesis has been rejected, we should proceed with a post-hoc test to check the statistical differences between pairs of classifiers. The multiple comparisons are usually performed using the mean-ranks test. *Because of fundamental inconsistencies of this test we discourage its use in machine learning. To overcome these issues, we suggest instead to perform the multiple comparison using a test whose outcome only depends on the two algorithms being compared, such as the sign-test or the Wilcoxon signed-rank test*.
- Regarding dataset selection, we must carefully choose the datasets to be included in the evaluation process to reflect the specific conditions, for example class imbalance, classification cost, dataset size, application domain, etc. *The choice of the datasets should be guided in order to identify specific conditions that make a classifier to perform better than others*.

Summarizing, this review tries to provide the reader with a better understanding about the overall process of classifier evaluation. We believe, that this review can improve the way in which researchers and practitioners in machine learning contrast the results achieved in their experimental studies using statistical methods. The propositions mentioned above (in italic) can direct researchers in their work on the new, better solutions for classifier evaluation procedures.

REFERENCES

[1] Bishop Ch. "Pattern recognition and machine learning," Springer, New York, 2006.

[2] Bouckaert R., "Estimating replicability of classifier learning experiments," Proc. 21st Conf. ICML, AAAI Press, 2004, http://dx.doi.org/10.1145/1015330.1015338.

[3] Bradley P., "The use of the area under the ROC curve in the evaluation of machine learning algorithms," Pattern recognition, 30, 1997, pp. 1145–1159, http://dx.doi.org/10.1016/S0031-3203(96)00142-2.

[4] Dietterich T., "Approximate statistical tests for comparing supervised classification learning algorithms," Neural Computation, 10, 1998, pp. 1895–1924, http://dx.doi.org/10.1162/089976698300017197.

[5] Demsar J., "Statistical comparison of classifiers over multiple data sets," Journal of Machine Learning Research, 7, 2006, pp. 1–30.

[6] Garcia S. Fernandez A., Lutengo J. and Herrera F., "Advanced non-parametric tests for multiple comparisons in the design of experiments in the computational intelligence and data mining: experimental analysis of power," Inf. Sci., 180(10), 2010, pp. 2044–2064, http://dx.doi.org/10.1016/j.ins.2009.12.010.

[7] Garcia V. et. al., "Index of balanced accuracy: a performance measure for skewed class distributions," 4th IbPRIA, 2009, pp. 441–448, http://dx.doi.org/10.1007/978–3–642–02172–5_57.

[8] Hand D., "Measuring classifier performance: a coherent alternative to the area under the ROC curve," Machine Learning, 77, 2009, pp. 103–123, http://dx.doi.org/10.1007/s10994–009–5119–5.

[9] Hollander M. and Wolfe D., "Nonparametric statistical methods," John Wiley & Sons, 2013, http://dx.doi.org/10.1002/9781119196037.

[10] Japkowicz N. and Shah M., "Evaluating learning algorithms: a classification perspective," Cambridge University Press, Cambridge, 2011.

[11] Salzberg S., "On comparing classifiers: pitfalls to avoid and recommended approach," Data Mining and Knowledge Discovery, 1, 1997, pp. 317–328, http://dx.doi.org/10.1023/A:1009752403260.

[12] Santafe G. et. al., "Dealing with the evaluation of supervised classification algorithms," Artif. Intell. Rev. 44, 2015, pp. 467–508, http://dx.doi.org/10.1007/s10462–015–9433–y.

[13] Shaffer J. P., "Multiple hypothesis testing," Annual Review of Psychology, 46, 1995, pp. 561–584.

[14] Sokolova M. and Lapalme G., "A systematic analysis of performance measures for classification tasks," Inf. Proc. and Manag., 45, 2009, pp. 427–437, http://dx.doi.org/10.1016/j.ipm.2009.03.002.

[15] Wolpert D., "The lack of a priori distinctions between learning algorithms," Neural Comput. 8(7), 1996, pp. 1341–1390, http://dx.doi.org/10.1162/neco.1996.8.7.1341.