Statistical Comparison of Classifiers for Multi-Objective Feature Selection in Instrument Recognition

Igor Vatolkin[1], Bernd Bischl[2], Günter Rudolph[1], and Claus Weihs[2]

[1] TU Dortmund, Chair of Algorithm Engineering
{igor.vatolkin;guenter.rudolph}@tu-dortmund.de
[2] TU Dortmund, Chair of Computational Statistics
{bischl;weihs}@statistik.tu-dortmund.de

Abstract. Many published articles in automatic music classification deal with the development and experimental comparison of algorithms - however the final statements are often based on figures and simple statistics in tables and only a few related studies apply proper statistical testing for a reliable discussion of results and measurements of the propositions' significance. Therefore we provide two simple examples for a reasonable application of statistical tests for our previous study recognizing instruments in polyphonic audio. This task is solved by multi-objective feature selection starting from a large number of up-to-date audio descriptors and optimization of classification error and number of selected features at the same time by an evolutionary algorithm. The performance of several classifiers and their impact on the pareto front are analyzed by means of statistical tests.

1 Introduction

A large share of interdisciplinary research as music information retrieval (MIR) (Downie (2003)) corresponds to experimental studies with comparison and evaluation of established and new algorithms. However, it can be observed that in many cases the suggestions or improvements of a novel technique are not properly evaluated: e.g. only one evaluation metric like accuracy is estimated, the holdout set is not completely independent, or the final assumptions are not underlined by any statistical tests which provide a solid estimation of the investigations reliability. Especially the lack of statistical testing holds also for the most of our own previous studies in music classification. Therefore we decided to check again the results of our study for instrument recognition in polyphonic recordings (Vatolkin et al. (2012)) and to apply exemplary tests on two experimental results. The target of this paper is not to provide a comprehensive introduction into statistical testing - but to

encourage the MIR community to use statistical tests for a better and more reliable evaluation of algorithms and results.

In the following subsections we introduce shortly the instrument recognition problem and refer to the relevant works about statistical tests for algorithm comparisons in MIR. Then we describe our study and discuss the application of two different tests for multi-objective classifier comparison. Finally, we conclude with recommendations for further research.

1.1 MIR and Instrument Recognition

Almost all MIR tasks deal directly or indirectly with classification: identification of music harmony and structure, genre recognition, music recommendation etc. One of these subtasks is instrument identification, allowing for further promising applications: music recommendation, organization of music collections or understanding of instrument role in a certain musical style. The largest challenge for successful instrument recognition in audio is that it is usually polyphonic: several simultaneously playing sources with different overtone distribution, noisy components and frequency progress over time make this task very complicated if the number of instruments is too large. Another problematic issue is that many different instrument playing possibilities (for example open or fretted strings) hinder the creation of well generalizable classification models which distinguish not only between different instruments but are also not influenced by these playing techniques. One of the recent comprehensive works related to instrument recognition in polyphonic audio is Fuhrmann (2012). An overview of the previous works mainly for recognition of singular instrument samples is provided by Eronen (2001).

1.2 Statistical Tests in Music Classification

Statistical hypothesis testing is a formal methodology for making judgments about stochastically generated data. In this article we will mainly consider two sample-location tests. In detail this means: We have observed numerical observations $v_1, \ldots, v_n$ and $w_1, \ldots, w_n$ and want to compare these two w.r.t. a specific "location parameter", e.g. their mean or median value. In a one-sample test we would compare the location parameter of only one population to a constant value, while "paired" means that we are actually interested in the location of $v_i - w_i$, because both observations have been measured at the same object and / or belong together. A common example of pairing in machine learning is that v_i and w_i are performance values of predictive models and have both been observed during resampling in iteration i on the same training and test sets. Depending on whether the statistics of interest is approximately normally distributed, the two most popular tests for this scenario are the paired t-test and the Wilcoxon signed-rank test (Hollander and Wolfe (1973)).

During the last years a vast number of conference and journal papers has been published in the MIR research area, but only a rather small fraction of them apply statistical tests. From a collection of 162 MIR-related publications we studied (from rather short conference papers to dissertations and master theses) only about 15 percent apply or directly mention statistical tests. Furthermore, in almost all of these works hypothesis tests were employed only in very few cases instead of a systematic analysis for algorithm comparison.

To name a couple of examples for further reading, Gillick and Cox (1989) argue about the importance of applying statistical tests to speech recognition, in particular they mention McNemar and matched-pairs test. The k-fold cross-validated t-test for comparison of temporal autoregressive feature aggregation techniques in music genre classification is applied in Meng et al. (2007) and demonstrates the significantly improved performance of the proposed methods. McKay (2010) uses the Wilcoxon signed-rank test for the comparison of features from different sources (symbolic, audio and cultural) for genre classification. Another evaluation of audio feature subsets for instrument recognition by statistical testing was performed by Bischl et al. (2010a). Noland and Sandler (2009) mention the application of the z-test for correlation measurements in key estimation based on chord progression.

2 Instrument Identification in Intervals and Chords

Here we provide a short description of our study, for details please refer to Vatolkin et al. (2012). The binary classification task was to detect piano, guitars, wind or strings in the mixtures of 2 up to 4 samples playing at the same time. The complete set included 3000 intervals (2 tone mixtures) and 3000 chords (3 and 4 tone mixtures). 2000 intervals and 2000 chords were used for model training and optimization based on cross-validation and the remaining mixtures were used as an independent holdout set for validation.

A 1148-dimensional audio feature vector was preprocessed and provided as input for 4 classifiers: decision tree C4.5, random forest (RF), naive Bayes (NB) and support vector machine (SVM). Since using a too large feature set comes with the additional costs of increased prediction time and storage space (both very relevant in MIR, e.g. see Blume et al. (2011)) and the trade-off between the size of the feature set and the prediction performance of the model is difficult to specify a priori we decided to perform multi-objective feature selection by means of an evolutionary algorithm (EA) w.r.t. to classification error E^2 and the proportion of selected features f_r. Three different initial feature rates $i_{FR} \in \{0.5; 0.2; 0.05\}$ (probabilities that each feature was selected for model building at the beginning of the optimization process) and three different crossover operators for EA were tested as optimization parameters. The results confirmed our suggestion that the feature selection is an important step providing successful and generalizable models.

The formal definition of multi-objective feature selection (MO-FS) can be described as:

$$\theta^* = \arg\min_{\theta} \left[m_1 \left(Y; \Phi(X, \theta) \right), ..., m_O \left(Y; \Phi(X, \theta) \right) \right], \tag{1}$$

where X is the full feature set, Y the corresponding labels, θ the indices of the selected features, $\Phi(X, \theta)$ the selected feature subset and $m_1, ..., m_O$ are O objectives to minimize.

The output of a MO-FS algorithm is a solution set, each of them corresponding to the subset of initial features. The *non-dominated front* of solutions consisted of the feature subsets with the best compromises between E^2 and f_r. Such front can be evaluated by a corresponding hypervolume:

$$\mathcal{S}(\mathbf{x}_1, ..., \mathbf{x}_N) = \bigcup_i vol(\mathbf{x}_i), \tag{2}$$

where $vol(\mathbf{x}_i)$ relates to the hypercube volume spanned between the solution $\mathbf{x}_i$ and the reference point which should be set to the worst possible solution responding to all metrics ([1;1] in our case).

3 Application of Tests

In Vatolkin et al. (2012) we provided 9 experimental results examining the overall performance of our method and comparing different classifiers and settings of EA parameters. These observations were in most cases clearly underlined by the corresponding experimental statistics and figures - however no significance measurements were done in a proper way. For the following subsections we selected two results and considered appropriate statistical tests for the reliability measurements (for instrument detection in chords).

3.1 All Classifiers are Important

The first result was that if all selected feature sets after FS were compared, it was hardly possible to claim that some of the classification methods were irrelevant: the non-dominated fronts of solutions contained solutions from all classification methods. This statement is illustrated by Fig. 1. Here we plotted all final solutions from 10 statistical repetitions and marked the non-dominated fronts by thick dashed lines. It can be stated that often certain classifiers occupy specific regions of the front: RF and SVM provide often the smallest E^2 values but require larger feature sets whereas C4.5 and NB perform worse corresponding to E^2 but may build models from extremely small feature sets.

For the measurement of statistical significance of the observation, that all classifiers are reasonable for non-dominated solution fronts, we need at first a null hypothesis. Therefore, H0 can be formulated as follows: given a classifier $\mathcal{A}$, the hypervolumes for all solutions fronts $\mathcal{S}_{all}$ and fronts built

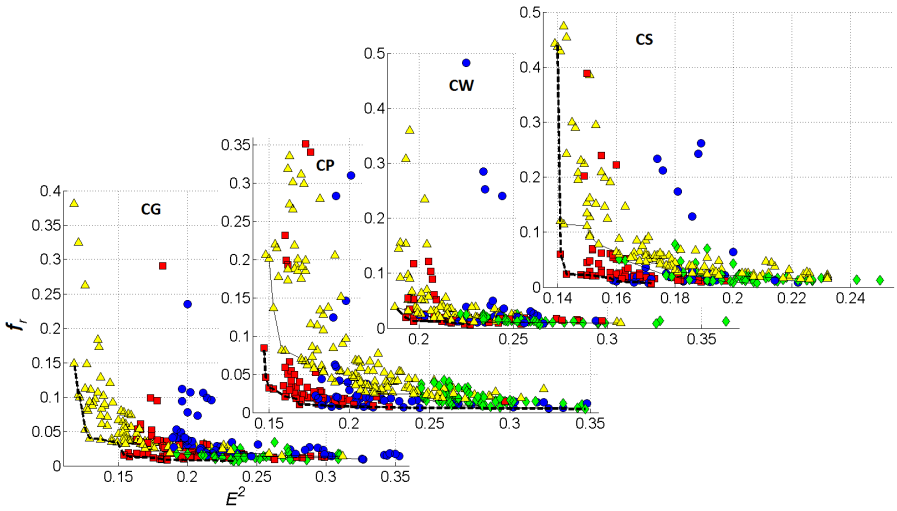

Fig. 1. Solutions after optimization from 10 statistical runs for each classification method optimizing mean classification error E^2 and feature rate f_r ($i_{FR} = 0.5$). Circles: C4.5; rectangles: RF; diamonds: NB; triangles: SVM. From left to right: CG: identification of guitar in chords; CP: piano; CW: wind; CS: strings.

of solutions without this classifier $\mathcal{S}_{all/\mathcal{A}}$ have the same distribution across r statistical repetitions of the experiment. Since: (1) the number of statistical repetitions was rather low ($r = 10$ because of large computing time and overall experiment number); (2) no assumption of the normal distribution and (3) the clear relationship between $\mathcal{S}_{all}$ and $\mathcal{S}_{all/\mathcal{A}}$, we selected the Wilcoxon signed rank test for paired observations. We run the test for 9 optimizer parameter settings (3 i_{FR} values $\times$ 3 crossover operators) separately. The frequency of H0

Table 1. Frequencies for H0 rejection for the test of classifier importance.

	C4.5	**RF**	**NB**	**SVM**
How often H0 rejected?	72.2%	100.0%	38.9%	55.6%
How often H0 rejected for $i_{FR} = 0.5$?	41.7%	100.0%	25.0%	83.3%
How often H0 rejected for $i_{FR} = 0.2$?	75.0%	100.0%	50.0%	50.0%
How often H0 rejected for $i_{FR} = 0.05$?	100.0%	100.0%	41.7%	33.3%

rejections for each classifier averaged across all combinations of optimization parameters is given in the first row of Table 1. It means, that the removal of RF solutions from the non-dominated front leads to decrease of hypervolume in all cases. The 'least important' NB still contributes to the hypervolumes in 38.9% of all experiments. Another interesting observation is the dependency of the classifier performance on the feature set size. We observed already in

Vatolkin et al. (2012), that SVM performs better starting with large feature sets whereas C4.5 suffers from too large amount of features despite of an integrated pruning technique. Now we can underline this by statistical test results: for experiments started with initial feature rate of 0.5 the removal of SVM solutions leads in 83.3% of the cases to hypervolume decrease. For $i_{FR} = 0.05$ this holds only for 33.3% of the runs. For C4.5 the situation is exactly opposite: C4.5 solutions were required even in all runs with $i_{FR} = 0.05$ for the fronts with largest hypervolumes. For NB no such clear behavior can be observed, but it seems to perform worse with larger feature sets.

3.2 Generalization Ability

The second important observation is that the classifiers provided models with different generalization ability, i.e. performance on an independent data set. Figure 2 lists hypervolumes of the last populations on the holdout set (1000 chords) divided by hypervolumes on the optimization set (2000 chords). A value above 1 means that the models perform better for holdout set than for optimization set. From the figure it can be clearly seen that SVM models are almost all less generalizable than RF models; in general C4.5 and RF provide the most robust models. For statistical analysis of model generalization ability

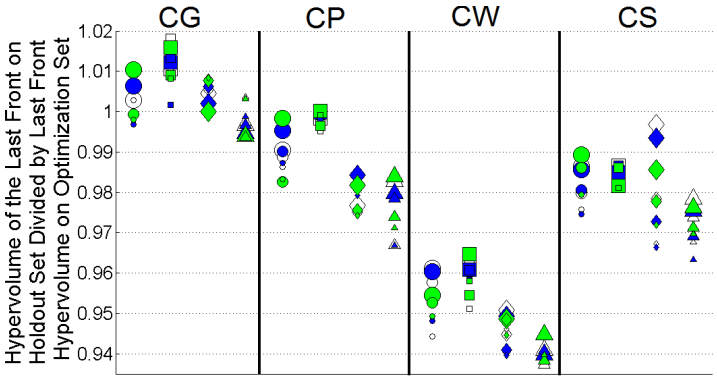

Fig. 2. Mean hypervolumes of the holdout set divided by mean hypervolumes of the optimization set. CG: recognition of guitar in chords; CP: piano; CS: strings; CW: wind. Circles: C4.5; rectangles: RF; diamonds: NB; triangles: SVM. Large signs: $i_{FR} = 0.5$; medium signs: $i_{FR} = 0.2$; small signs: $i_{FR} = 0.05$. Different shades correspond to different crossover settings.

between two classifiers $\mathcal{A}, \mathcal{B}$ we compare the distributions of $\widetilde{h_A}$ and $\widetilde{h_B}$, where $h_C(r_i) = \mathcal{S}_{holdout}(C, r_i)/\mathcal{S}_{opt}(C, r_i)$ is the rate of holdout hypervolume divided by optimization hypervolume for classifier C and run r_i and $\widetilde{h_C}$ is the mean value across 10 statistical repetitions. The H0 hypothesis is that the $\widetilde{h_\mathcal{A}}$ and $\widetilde{h_\mathcal{B}}$

distributions are equal, meaning that there is no significant difference between the model generalization abilities for classifiers $\mathcal{A}$ and $\mathcal{B}$. In Table 2, the first

Table 2. Frequencies for H0 rejection for the test of model generalization ability.

Classifier $\mathcal{A}$	**Classifier $\mathcal{B}$**	**How often H0 rejected?**	**How often $\widetilde{h_\mathcal{A}} > \widetilde{h_\mathcal{B}}$?**
RF	**SVM**	88.9%	100.0%
RF	**NB**	66.7%	86.1%
RF	**C4.5**	22.2%	91.7%
C4.5	**SVM**	61.1%	88.9%
C4.5	**NB**	27.8%	69.4%
NB	**SVM**	22.2%	88.9%

table row can be interpreted as follows: the mean $\widetilde{h_C}$ across all optimizer parameters and statistical repetitions for RF was in 100% cases larger than for SVM (last column). H0 was rejected in 88.9% cases - although this is below 100%, we can indeed state that RF tends to create significantly more generalizable models than SVM. The further lines provide less clear results, however we can state, that RF provides rather more generalizable models than NB and C4.5 than SVM. This behaviour can be also observed from Fig. 2 - but it does not illustrate all concrete values from the statistical repetitions and provides no statistical significance testing.

4 Final Remarks

Another important issue for statistical test design is that that the hypotheses must be created before the data analysis - otherwise they may hold only for the concrete data set. The first hypothesis (all classifiers are important) was already influenced by our multi-objective feature selection study in Vatolkin et al. (2011) - and the second one (different model generalization performances) was considered after the creation of Fig. 2. The final and only accurate way to underline the significance of this statement - which was here not possible because of the large optimization times for all experiments - is to rerun the complete study for another 3000 chords and to apply the test again.

Concluding our short excursion with two examples of statistical test application in music instrument recognition, we strongly recommend the following three steps to be carefully planned for design of any new study comparing performance of classification algorithms (in MIR as well as in other domains): a) Design of an independent holdout set neither involved in training of classification models nor any optimization and parameter tuning (see Fiebrink and Fujinaga (2006) especially for feature selection in MIR and our previous publications from the reference list). b) Consideration of multi-objective optimization or at least evaluation comparing the methods: if an algorithm

performs better than another one with response to e.g. accuracy, it may be on the other side slower, fail on highly inbalanced sets or provide less generalizable models (see Vatolkin et al. (2011) for different evaluation scenarios). c) Application of statistical tests for reliable comparison of methods and significance measurements as discussed in this work.

References

BISCHL, B., EICHHOFF, M. and WEIHS, C. (2010): Selecting Groups of Audio Features by Statistical Tests and the Group Lasso. In: *Proc. of the 9th ITG Fachtagung Sprachkommunikation, VDE Verlag.*

BISCHL, B., VATOLKIN, I. and PREUß, M. (2010): Selecting Small Audio Feature Sets in Music Classification by Means of Asymmetric Mutation. In: *Proc. of the 11th Int'l Conf. on Parallel Problem Solving From Nature (PPSN), 314–323.*

BLUME, H., BISCHL, B., BOTTECK, M., IGEL, C., MARTIN, R., RÖTTER, G., RUDOLPH, G., THEIMER, W., VATOLKIN, I. and WEIHS, C. (2011). Huge Music Archives on Mobile Devices. *IEEE Signal Processing Magazine, 28(4), 24–39.*

DOWNIE, S. (2003): Music Information Retrieval. *Annual Review of Information Science and Technology, 37(1), 295–340.*

ERONEN, A. (2001): *Automatic Musical Instrument Recognition.* Master's thesis, Department of Information Technology, Tampere University of Technology.

FIEBRINK, R. and FUJINAGA, I. (2006): Feature Selection Pitfalls and Music Classification. In: *Proc. of the 7th Int'l Conf. on Music Information Retrieval (ISMIR), 340–341.*

FUHRMANN, S. (2012): *Automatic Musical Instrument Recognition from Polyphonic Music Audio Signals.* PhD thesis, Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona.

GILLICK, L. and COX, S. (1989): Some Statistical Issues in the Comparison of Speech Recognition Algorithms. In: *Proc. of the IEEE Conf. on Acoustics, Speech and Signal Processing (ICASSP), 532–535.*

HOLLANDER, M. and WOLFE, D. A. (1973): *Nonparametric statistical methods.* Wiley, New York.

MCKAY, C. (2010): *Automatic Music Classification with jMIR.* PhD thesis, Department of Music Research, Schulich School of Music, McGill University, Montreal.

MENG, A., AHRENDT, P., LARSEN, J. and HANSEN, L. K. (2007): Temporal Feature Integration for Music Genre Classification. *IEEE Transactions on Audio, Speech and Language Processing, 15(5), 1654–1664.*

NOLAND, K. and SANDLER, M. (2009): Influences of Signal Processing, Tone Profiles, and Chord Progressions on a Model for Estimating the Musical Key from Audio. *Computer Music Journal, 33(1), 42-56.*

VATOLKIN, I., PREUß and RUDOLPH, G. (2011): Multi-Objective Feature Selection in Music Genre and Style Recognition Tasks. In: *Proc. of the 2011 Genetic and Evolutionary Computation Conf. (GECCO), ACM Press, 411–418.*

VATOLKIN, I., PREUß, RUDOLPH, G., EICHHOFF, M. and WEIHS, C. (2012): Multi-Objective Evolutionary Feature Selection for Instrument Recognition in Polyphonic Audio Mixtures. *Soft Computing, 16(12), 2027–2047.*