Measuring the Performance of Evolutionary Multi-Objective Feature Selection for Prediction of Musical Genres and Styles

Igor Vatolkin

Chair of Algorithm Engineering
Department of Computer Science
Technische Universität Dortmund
44227 Dortmund
igor.vatolkin@tu-dortmund.de

Abstract: The prediction of high-level music categories, such as genres, styles, or personal preferences, helps to organise music collections. The relevance of single audio features for automatic classification depends on a certain category. Relevant feature subsets for each classification task can be identified by means of feature selection. Continuing our previous studies on multi-objective feature selection for music classification, in this work we measure an impact of evolutionary multi-objective feature selection on classification performance and compare it to the baseline application without feature selection. As confirmed by statistical tests, the integration of evolutionary multi-objective feature selection leads to a significant increase of performance according to both evaluation criteria as well as to classification error. This holds for all four tested classification methods and six music categories.

1 Introduction

Music classification is one of the most prominent applications in music information retrieval (MIR). Many different music classification tasks exist: recognition of tempo and rhythm, cover song detection, identification of harmony and melody characteristics, etc. However, probably the most user-centered application is the prediction of high-level music categories, such as genres and styles. This categorisation helps to manage large music collections but also to recommend new music to listeners. A number of possible high-level categories is hardly limited: genre taxonomies are not consistent [PC00] and may evolve over time. Music listeners may define new categories over again depending on their changing preferences. Also the listening content plays a role, consider a list with music listening strategies [Hur02].

Automatic classification systems try to predict categories from numerical characteristics, or features, which should represent music pieces. The first basic categorisation approach is to classify data in a *supervised* way, where classification models are trained from labelled feature vectors, or ground truth. Another possibility is to learn from unlabelled feature

vectors and to organise music collections in an *unsupervised* way, building, e.g., clusters of similar music. The particular difficulty in the last case is that predicted classes may not correspond to clearly interpretable genres and personal preferences. For both approaches, their success depends on the quality of music features. Too many irrelevant or noisy features often lead to a diminished classification performance: with an increasing number of features the probability increases that some of them would be identifed as relevant by chance. For example, the impact of noisy features for decision tree classifiers is discussed in [WF05], and for random forests in [HTF09].

Feature selection (FS) [GNGZ06] provides a solution for this problem, aiming at the removal of irrelevant and redundant features and an increase of classification performance. But there exist other reasons for FS: categorisation models created with less features provide faster classification and have less storage demands. Combined with a proper validation method, feature selection may reduce the danger of highly overfitted models, which classify classify well the training data but generalise poorly. Further, an initial large feature set can be extracted only once, and for each new possible classification category the corresponding relevant features may be identified by feature selection. However, it should not be forgotten that FS has its own runtime demands as any optimisation step.

In recent years, feature selection became a part of many MIR classification studies. To name several relevant contributions, one of the first works was [Fuj98], where feature selection with genetic algorithms (GA) was applied for classification of instrument tones. A sequential forward feature selection was applied for drum sound classification in [FF06]. In [GCK03], two sequential FS methods and principal component analysis (PCA) were compared for recognition of music genres. PCA performed at best–however this statistical method has several disadvantages: the extraction of all original feature dimensions is still required, and the interpretability of original features is not given anymore. The application of feature selection for mood recognition is analysed in [SEL11].

In almost all studies the success of feature selection is measured by a single evaluation criterion, such as accuracy or classification error. However, many conflicting evaluation criteria exist for algorithm evaluation. For example, an algorithm with a higher classification quality may be slower, the performance may vary on different subsets of songs, or high user efforts may be required for the definition of ground truth. In [VPR11], we discussed several groups of evaluation criteria, which make sense for music classification evaluation: (1) Quality-based metrics measure classification performance. This group contains many confusion matrix-based metrics, such as accuracy, precision, and recall, but also metrics developed for imbalanced sets. (2) Resource metrics describe runtime and storage demands. (3) Model complexity metrics validate the generalisation ability of classification models. (4) User interaction metrics measure user satisfaction, efforts for ground truth definition and interaction with an algorithm, etc. (5) Specific quality-based metrics are defined for a concrete classification task only.

Therefore, we recommend to measure the success of feature selection in a multi-objective way. In our previous studies, we have successfully applied evolutionary multi-objective feature selection for different tasks: recognition of instruments [VPR$^+$12], genres and styles [VPR11], or personal preferences [VTR09]. However, we did not measure the impact of feature selection by means of statistical tests. This is addressed in this study. In

Sect. 2, we provide a short formal definition of the multi-objective feature selection task and briefly describe how evolutionary algorithms (EA) can be applied for feature selection. In the succeeding section, we describe the experiment setup. In Sect. 4, the impact of proposed feature selection method is evaluated by means of statistical tests. We conclude with summary and outlook.

2 Evolutionary multi-objective feature selection

For supervised music classification based on audio features, labelled classification instances usually describe music intervals of several seconds up to complete songs. Let F be the number of features, I the number of classification instances, $\mathbf{X} \in \mathbb{R}^{F \times I}$ the complete feature matrix, and let $\mathbf{y}_L \in [0;1]^I$ describe the relationships of instances to classification labels (for binary classification, $\mathbf{y}_L \in \{0;1\}^I$). Then, the target of feature selection is to find an optimal feature subset θ^*, so that a relevance function m (also referred to as an objective, or an evaluation metric) has to be minimised. Typical relevance functions are classification error, accuracy, precision, etc. We adapt the definition from [Tor06]:

$$\theta^* = \arg\min_{\theta} \left[m\left(\mathbf{y}_L; \Phi(\mathbf{X}, \theta)\right) \right] \tag{1}$$

where $\Phi(\mathbf{X}, \theta)$ corresponds to a feature matrix built only with selected features, denoted with their indices θ. For relevance functions, which have to be maximised, Eq. 1 can be simply rewritten, or a corresponding metric can be redefined.

If O different criteria play a role in feature selection, multi-objective feature selection is defined as follows:

$$\theta^* = \arg\min_{\theta} \left[m_1\left(\mathbf{y}_L; \Phi(\mathbf{X}, \theta)\right), ..., m_O\left(\mathbf{y}_L; \Phi(\mathbf{X}, \theta)\right) \right]. \tag{2}$$

The comparison of optimisation solutions (feature subsets) is not straightforward anymore, when two and more evaluation metrics are optimised. It is possible to compare solutions in terms of *Pareto dominance*. A feature subset $\Phi(\mathbf{X}, \theta_a)$ Pareto dominates the subset $\Phi(\mathbf{X}, \theta_b)$ (denoted by $\Phi(\mathbf{X}, \theta_a) \prec \Phi(\mathbf{X}, \theta_b)$), iff:

$$\begin{aligned} &\forall i \in \{1, ..., O\} : m_i\left(\mathbf{y}_L; \Phi(\mathbf{X}, \theta_a)\right) \leq m_i\left(\mathbf{y}_L; \Phi(\mathbf{X}, \theta_b)\right) \text{ and} \\ &\exists j \in \{1, ..., O\} : m_j\left(\mathbf{y}_L; \Phi(\mathbf{X}, \theta_a)\right) < m_j\left(\mathbf{y}_L; \Phi(\mathbf{X}, \theta_b)\right). \end{aligned} \tag{3}$$

The target of multi-objective feature selection can be then decribed as a search for the *Pareto front* of feature subsets, which are not dominated by any other subsets. Often a multi-objective optimisation algorithm outputs a set of solutions, which does not contain all Pareto front solutions. The *non-dominated* front contains all feature subsets, which are not dominated by any other feature subsets among the output solutions.

The quality of a non-dominated front with cardinality P_{ND} can be measured by its dominated hypervolume, defined as follows:

$$\mathcal{S}(\Phi(\mathbf{X},\theta_1),...,\Phi(\mathbf{X},\theta_{P_{ND}})) = vol\left(\bigcup_{i=1}^{P_{ND}}[\Phi(\mathbf{X},\theta_i),\mathbf{r}]\right) \tag{4}$$

where $\mathbf{r}$ is a reference point in the objective space, which position is set to the worst possible metric values for all dimensions ($\forall i : \mathbf{m}(\Phi(\mathbf{X},\theta_i)) \prec \mathbf{r}$, where $\mathbf{m} \in \mathbb{R}^O$ is the vector of all objectives). $[\Phi(\mathbf{X},\theta_i),\mathbf{r}]$ is a hypercube spanned between the subset $\Phi(\mathbf{X},\theta_i)$ and $\mathbf{r}$, and $vol(\cdot)$ is the overall volume of all those hypercubes.

A single feature subset $\Phi(\mathbf{X},\theta_i)$ can be evaluated by its individual contribution to a dominated hypervolume, i.e. an area of possible solutions, which are dominated only by $\Phi(\mathbf{X},\theta_i)$:

$$\begin{aligned}\Delta\mathcal{S}(\Phi(\mathbf{X},\theta_i)) = \mathcal{S}(\Phi(\mathbf{X},\theta_1),...,\Phi(\mathbf{X},\theta_{P_{ND}})) \; - \\ \mathcal{S}(\Phi(\mathbf{X},\theta_1),...,\Phi(\mathbf{X},\theta_{i-1}),\Phi(\mathbf{X},\theta_{i+1}),...,\Phi(\mathbf{X},\theta_{P_{ND}})).\end{aligned} \tag{5}$$

Figure 1 illustrates the difference between hypervolume $\mathcal{S}$ (left subfigure) and individual hypervolume contribution $\Delta\mathcal{S}$ (right subfigure).

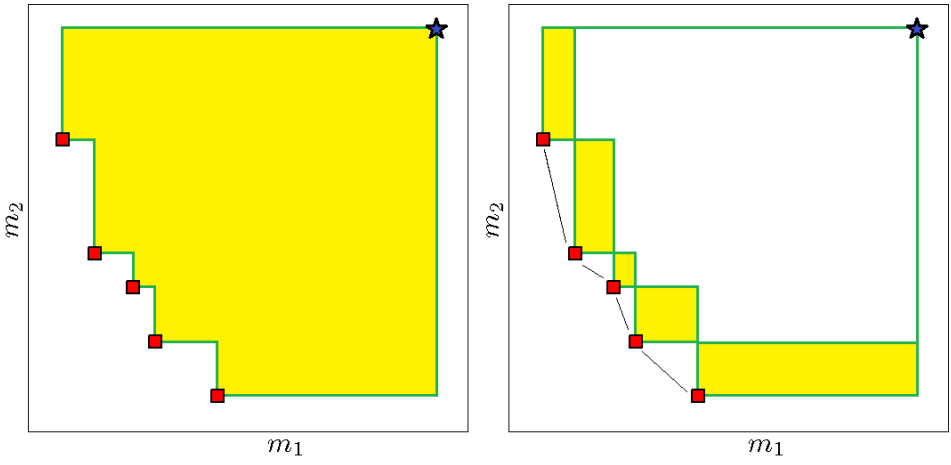

Figure 1: m_1 and m_2 are two objectives to minimise. Reference point is marked by an asterisk. Non-dominated solutions (feature subsets) are marked by small red squares. Left subfigure: hypervolume of the non-dominated solution front is marked with a filled yellow area. Right subfigure: individual hypervolume contributions of feature subsets are marked with filled yellow rectangles.

Evolutionary algorithms mimic the natural evolution process in their operating method [RBK12]. A problem solution is coded by its representation in the *search space*: in case of feature selection, a typical representation is a bit vector. A one at position i means that

the i-th feature ($i \in \{1, ..., F\}$) is selected, and a zero that this feature is not selected. A solution is evaluted in the *objective space*, e.g., by a corresponding classification error. A single solution is referred to as an *individual*, and a set of solutions as a *population*. During the evolutionary process, population alters by means of update rules with integrated random components. Two basic operators are crossover and mutation. The target of crossover is to keep positive characteristics of two or more parent solutions during the generation of one or more offspring solutions from the parents. The target of mutation is to overcome local optima in the objective space by a stochastic variation in the search space.

As discussed above, many solutions are not directly comparable in multi-objective optimisation, and the goal is to find a set of trade-off solutions. Population-based EA are very well suited to solve these tasks [CLV07]. The selection of optimal feature subsets is NP-hard, as it was shown for related problems [KJ97], and the stochastic component of EA may help to overcome local optima, compared to deterministic feature selection methods. Further, EA require a significantly lower amount of evaluations than sequential FS methods. Evolutionary algorithms were recommended for "large" feature selection tasks with more than 100 features [KS00]. Evolutionary multi-objective algorithms were proposed for feature selection for the first time in [EHM00]. Since that study, they have been applied in many applications, but are almost unexplored in MIR classification tasks.

In our studies, we use an adapted $\mathcal{S}$-metric selection evolutionary multi-objective algorithm (SMS-EMOA) [EBN05] for feature selection. Since crossover did not lead to any significant performance increase in [VPR$^+$12], we use only an asymmetric mutation operator as proposed in [JPE02], which prefers to switch bits off. For each bit k of a feature subset representation vector $\mathbf{q}$, the probability of a bit flip $p_q(k)$ is estimated as follows:

$$p_q(k) = \frac{\gamma}{F} \cdot (|q_k - p_{01}|) \tag{6}$$

where γ is a general mutation probability of a bit flip, and p_{01} controls the probability of a zero-to-one flip. In this study, we use $\gamma = 32$ and $p_{01} = 0.01$, because this settings performed well in the past experiments.

The initial cardinality of a feature set depends on a further parameter if_r, which controls the probability of feature selection during the initialisation of a bit vector. Each of F features is selected with a probability if_r. In this study $if_r \in \{0.2; 0.5\}$. The population size is set to 50 individuals, and in each iteration step a new offspring feature subset is generated with the help of mutation operator after a random selection of the parent individual.

3 Experiment setup

Three music genres (Classic, Pop, Rap) and three music styles (ClubDance, HeavyMetal, ProgressiveRock) are predicted. The songs belong to our music database[1], and the category labels are from the AllMusicGuide[2] web site.

[1] http://ls11-www.cs.tu-dortmund.de/rudolph/mi#music_test_database
[2] http://www.allmusic.com

As features, a large set of 636 audio signal characteristics is used (for multidimensional features, each dimension is treated separately, so that, e.g., a chroma vector contributes to 12 feature dimensions). The advantage of these features is that they can be extracted from any available digital signal, whereas, e.g., score-related and metadata descriptors are not always available for less popular songs. The left subfigure of Fig. 2 shows the feature distribution. The largest share of features describes timbre and energy: spectral characteristics, time signal energy, mel frequency cepstral coefficients, phase domain characteristics [MM05], cepstral modulation ratio regression [MN09], etc. The right subfigure of Fig. 2 illustrates the extraction domains of timbre and energy features (the right subfigure corresponds to the blue part (404) of the left subfigure). Other features comprise several chroma implementations, including bass chroma [MD10] and chroma DCT-reduced log pitch [ME10]. The remaining features describe fluctuation patterns [Lar12] and correlation statistics. All features were extracted with Advanced MUSic Exporer (AMUSE) [VTB10].

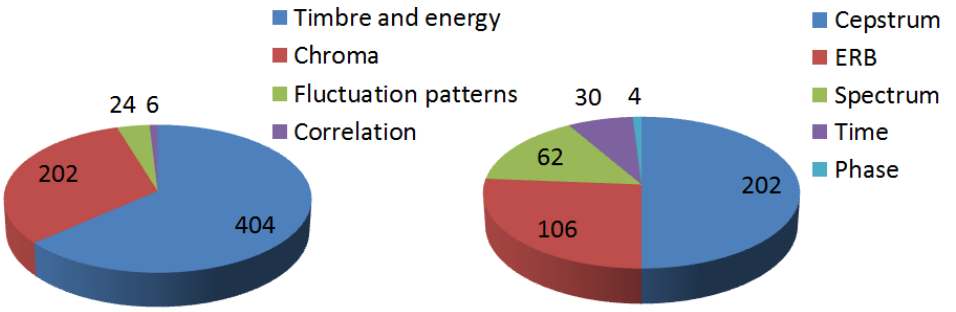

Figure 2: Left subfigure: groups of audio signal features used for genre and style prediction. Right subfigure: extraction domains for features, which belong to timbre and energy characteristics.

For all music pieces, classification instances are built from time windows of 4 second with 2 second overlap. Only features from the short extraction frames between previously extracted onset events are taken into account, since this method performed well in our study on comparison of different feature processing methods [VTB12]. Classification is done with four algorithms: decision tree C4.5, random forest (RF), naive Bayes (NB) and support vector machine (SVM) with linear kernel.

The validation of algorithms requires the design of several steps: the choice of evaluation metric(s), the organisation of data sets, and the measurement of significance of study results by means of statistical tests.

Two evaluation metrics are used as optimisation objectives. The first one is the balanced relative error, defined as follows:

$$m_{BRE} = \frac{1}{2}\left(\frac{FN}{TP+FN} + \frac{FP}{TN+FP}\right) \tag{7}$$

where TP is a number of true positives (songs belonging to a category and predicted as belonging to it), TN is a number of true negatives (songs not belonging to a category and

predicted as not belonging to it), FP is a number of false positives (songs not belonging to a category and predicted as belonging to it), and FN is a number of false negatives (songs belonging to a category and predicted as not belonging to it). $TP + FN$ is the number of ground truth positives, and $TN + FP$ the number of ground truth negatives.

The predicted song binary label $y_P(i, j) \in \{0; 1\}$ is estimated by a majority voting across all T classification instances, which belong to i-th song:

$$y_P(i) = \left\lceil \frac{\sum_{j=1}^{T} y_P(i, j))}{T} - 0.5 \right\rceil \tag{8}$$

where $y_P(i, j)$ is a binary predicted label for classification instance j from song i.

The second metric, the selected feature rate m_{SFR}, relates to both resource and model complexity metrics. Classification models built with less features require less computing time and storage demands, and may have lesser tendency to be overfitted. m_{SFR} is defined as:

$$m_{SFR} = \frac{F^*}{F} \tag{9}$$

where F^* is the number of selected features, and F the overall number of features.

To avoid overfitting during the evaluation of feature subsets, it is essential to use an independent holdout set for data evaluation (see also [FF06, SEL11] for further recommendations). In our study, we use three disjoint song data sets for each classification task. A *training* set contains 20 songs: 10 positive and 10 negative examples for each category. These songs are used for training of classification models. An *optimisation* set contains 120 songs, which are used for evaluation of feature subsets generated during evolutionary multi-objective feature selection. A *holdout* set contains 120 songs, which were neither involved into training nor evaluation of classification models.

For the comparison of classification performance of the baseline method without feature selection and classification with previously applied FS we use the Wilcoxon signed rank test set [HW99]. In the null hypothesis it is assumed that two observation vectors $\mathbf{u}, \mathbf{v}$ belong to the same unknown distribution. The null hypothesis is rejected, if for two sample vectors $\mathbf{u}, \mathbf{v}$ of the same length A it holds:

$$T_S^W \geq \tau_{\alpha/2} \text{ or } T_S^W \leq \frac{A(A+1)}{2} - \tau_{\alpha/2} \tag{10}$$

where α is the level of significance (usually set to 0.05), τ_α is the critical value for the Wilcoxon signed rank test set, and T_S^W is the test statistic:

$$T_S^W = \sum_{i=1}^{A} R^W(|u_i - v_i|) \tag{11}$$

where $R^W(\cdot)$ is the rank from all sorted values $\{|u_1 - v_1|, ..., |u_A - v_A|\}$.

4 Discussion of results

Figure 3 plots non-dominated fronts, which are built from all statistical repetitions of experiments. Thin lines mark the fronts for each classifier. Thick lines mark general non-dominated fronts created from all classification methods. Upper left regions of fronts correspond to larger feature sets with smaller classification errors. Bottom right regions of fronts correspond to smaller feature sets with higher errors. Even if extreme subsets with too few features and too large error are less meaningful, the solutions from middle regions make sense: smaller feature sets lead to faster classification, require less storage space, and may reduce the tendency for overfitting. Furthermore, runtime and storage demands play an important role for mobile devices with limited hardware resources and also for very large personal music collections. As it can be seen in Fig. 3, it is possible, e.g., to reduce m_{SFR} from 0.1242 (79 features, classification by NB) to 0.042 (27 features, classification by RF) for Rap category, where the classification error increases rather slightly from 0.0473 to 0.0513.

The classification tasks are different according to their complexity: for Classic, the smallest m_{BRE} values for four classification methods are between 0.0113 (RF) and 0.0598 (NB). The hardest category to predict is ClubDance, where the smallest m_{BRE} are between 0.1442 (SVM) and 0.1645 (RF).

Similar to results of our previous study on instrument identification [VPR+12], no clear preference for a certain classification method can be done. Only C4.5 contributes rather seldom to global non-dominated fronts, which are built across all classification models. However, C4.5 seems to perform sometimes better than other classifiers for very small feature sets: for Classical, a boundary subset of only 15 features has $m_{BRE} = 0.0508$, for Pop a subset of 15 features has $m_{BRE} = 0.1982$, and for Rap a subset of 16 features has $m_{BRE} = 0.1061$. On the other side, feature subsets with smallest m_{BRE} and largest m_{SFR} belong to different classifiers for different tasks: they contain a RF classification model for Classic, NB models for Rap and ProgRock, and SVM models for Pop, ClubDance, and HeavyMetal. This observation supports our investigations from [VBRW12], where all classification methods contributed to the global non-dominated front, and we recommended to combine several classification algorithms.

4.1 Two-objective performance

To measure an impact of feature selection, the mean dominated hypervolume before and after evolutionary FS can be estimated. Because a hypervolume-related metric $\Delta\mathcal{S}$ is used itself for evaluation during the SMS-EMOA evolutionary loop, we measure a change of hypervolume on the independent holdout song set. Let $\overline{\mathcal{S}}^H_{init}$ be the mean initial hypervolume for the holdout set, averaged for 10 statistical repetitions of experiments, and $\overline{\mathcal{S}}^H_{fin}$ the mean final hypervolume for the holdout set. A relative increase of mean hypervolume is then measured by $\frac{\overline{\mathcal{S}}^H_{fin} - \overline{\mathcal{S}}^H_{init}}{\overline{\mathcal{S}}^H_{init}} \cdot 100\%$. The results are plotted in Fig. 4 for all combinations

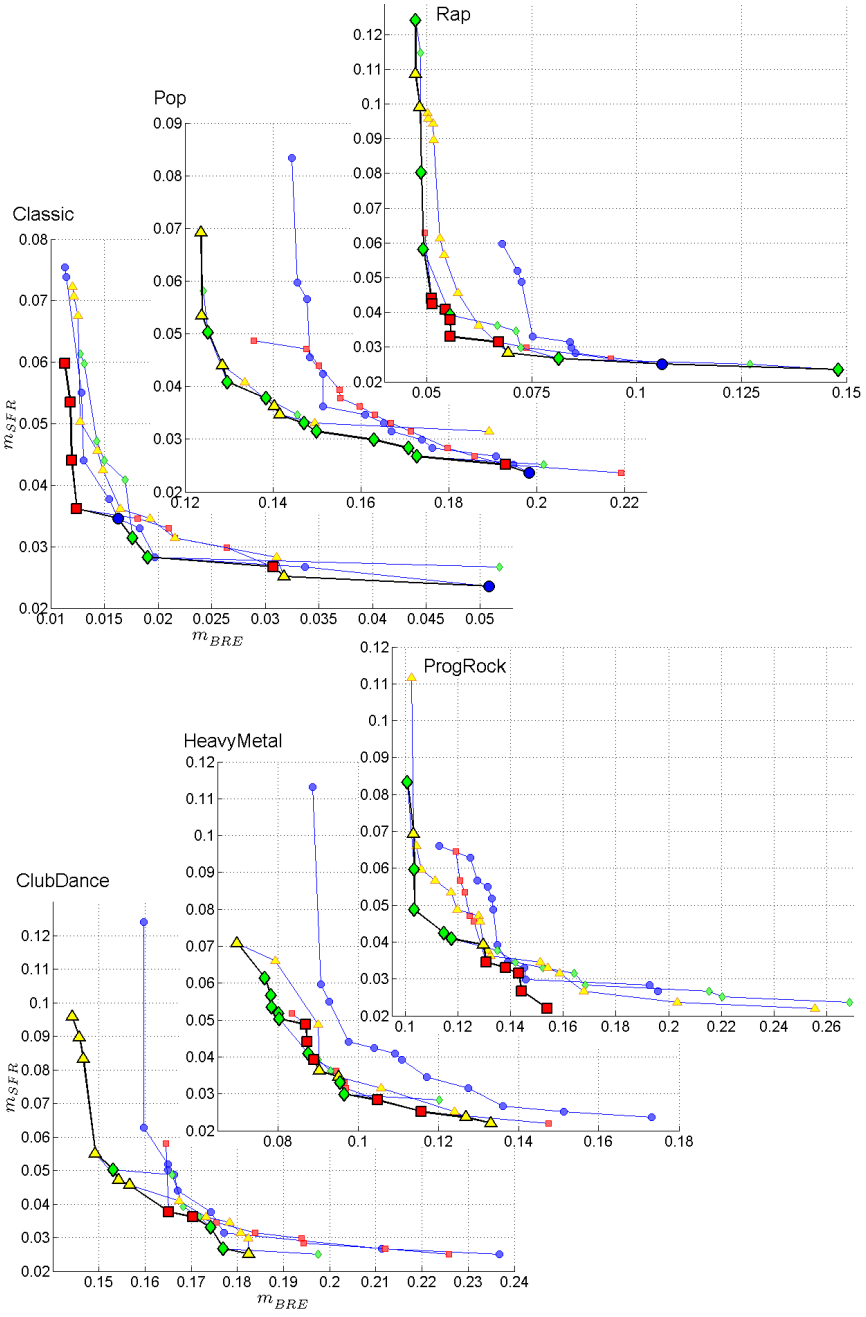

Figure 3: Non-dominated feature subset fronts created from all statistical repetitions. Circles: C4.5, squares: RF, diamonds: NB, triangles: SVM.

of classification methods, categorisation tasks, and two if_r settings.

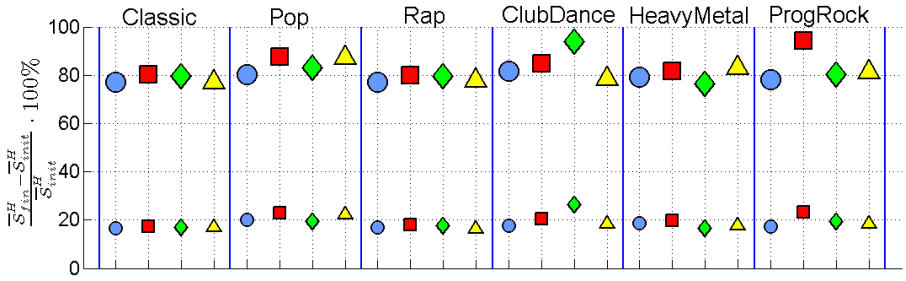

Figure 4: Relative mean holdout dominated hypervolume increase after the optimisation. Circles: C4.5, squares: RF, diamonds: NB, triangles: SVM. Large markers: $if_r = 0.5$, small markers: $if_r = 0.2$.

It can be observed that the mean hypervolume increases in all cases. As it can be expected, this increase is higher for experiments with $if_r = 0.5$: if initial feature subsets contain approximately half of all features (318 features), it is easy to reduce an amount of features keeping or reducing classification error at the same time. But also for $if_r = 0.2$ an increase of the relative hypervolume is about 20% for all tested classifiers and tasks.

For estimation of significance, the Wilcoxon signed rank test is applied. Two observation vectors **u** and **v** are compared. $\mathbf{u}(i, j, k)$ is a 10-dimensional vector of initial holdout dominated hypervolumes for a fixed classifier $i \in \{1, ..., 4\}$, a fixed if_r setting $j \in \{1, 2\}$, and a fixed classification task $k \in \{1, ..., 6\}$. $\mathbf{v}(i, j, k)$ is a similarly built 10-dimensional vector, which contains final holdout dominated hypervolumes. The null hypothesis that **u** and **v** belong to the same distribution is rejected in all cases for a significance level $\alpha = 0.05$. Therefore, we can state that evolutionary multi-objective feature selection leads to a *significant* increase of hypervolume estimated on the independent holdout set.

4.2 Classification error performance

Obviously, classification quality is more important than m_{SFR} performance. We may measure an impact of evolutionary multi-objective feature selection on the single-objective performance according to the balanced classification error. For each statistical repetition of an experiment with a fixed combination of the classification task, the classification method, and the if_r setting, boundary non-dominated front solutions with smallest m_{BRE} and largest m_{SFR} are analysed. Then, the relative error decrease compared to the baseline method with classification using a full feature set can be measured by $\frac{m_{BRE}(\Phi_{all}) - \overline{m}_{BRE}}{m_{BRE}(\Phi_{all})} \cdot 100\%$ (Φ_{all} denotes the full feature set).

As plotted in Fig. 5, the error decrease has a strong variation between 10.08% (Rap, SVM, $if_r = 0.2$) and 77.39% (ProgRock, NB, $if_r = 0.2$). It depends on a classification task and a classifier. Almost always NB benefits strongly from feature selection. C4.5 has already an integrated feature pruning technique–and the error decrease for this method is indeed

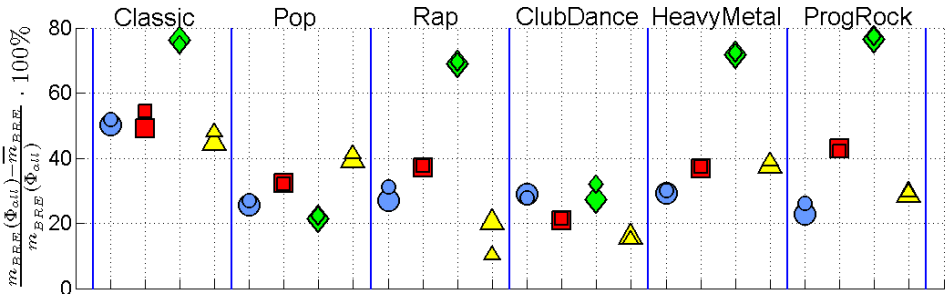

Figure 5: Relative balanced classification error decrease compared to classification with a complete feature set. Circles: C4.5, squares: RF, diamonds: NB, triangles: SVM. Large markers: $if_r = 0.5$, small markers: $if_r = 0.2$.

the lowest for categories HeavyMetal and ProgRock. However, this observation does not hold for other categories. Also, for C4.5 a relative error decrease is always above 20%.

Again, for the estimation of significance of the feature selection impact, we apply the Wilcoxon signed rank test. Both 10-dimensional observations are built in the following way: $\mathbf{u}(i,j,k)$ is a 10-dimensional vector of minimal m_{BRE} values from statistical repetitions for a fixed classifier $i \in \{1,...,4\}$, a fixed if_r setting $j \in \{1,2\}$, and a fixed classification task $k \in \{1,...,6\}$. $\mathbf{v}(i,j,k)$ is a similarly built 10-dimensional vector, which contains m_{BRE} values for full feature sets. The null hypothesis that **u** and **v** belong to the same distribution is rejected in all cases for a significance level $\alpha = 0.05$. p-value is only in one case (Rap, SVM, $if_r = 0.2$) equal to 0.049 and is slightly below the boundary p = 0.05. It means that for all combinations of a task, classifier, and if_r setting, the application of evolutionary multi-objective feature selection leads to a *significant* decrease of balanced classification error compared to the baseline method using a complete feature set.

5 Summary and outlook

In this study, we measured the impact of evolutionary multi-objective feature selection on two-objective and single-objective performance for three genres and three styles. The application of the Wilcoxon signed rank test confirmed a significant increase of dominated holdout hypervolumes compared to the baseline method without feature selection: the feature subsets identified by evolutionary FS are built with less features and have smaller classification errors. Furthermore, a significant decrease of classification errors was confirmed by the Wilcoxon signed rank test, compared to the classification with the full feature set. Although the study design was limited (only 6 genres and styles, four classifiers, no tuning of classification parameters), and the results can not be generalised for a wide range of other MIR classification tasks, the advantage of feature selection is clearly observed. Similar results were obtained also for instrument recognition [VPR$^+$12]. The success of feature selection can be explained by theoretical observations as well: with an increasing number of features the probability increases that some of them are by chance recognised

as relevant.

Classification tasks in MIR may be very different. User-related categories may differently depend on temporal, timbral, harmonic, and other characteristics. Manual selection of relevant features for each task requires too high personal efforts. A more promising solution for automatic classification is to start with a sufficiently large feature set and to integrate feature selection for the recognition of the most relevant features for a particular task.

In the future, other studies and improvements are possible, and we name here only several directions and ideas. The performance of evolutionary multi-objective feature selection can be analysed for other high-level categories and other music-related characteristics. A more extensive comparison to deterministic feature selection methods is reasonable. Also, the single-objective performance according to classification error can be compared to the performance of a single-objective evolutionary feature selection. Further combinations of evaluation metrics make sense, such as sensitivity and specificity, as applied in our previous study [VPR11]. Finally, the reduction of computing demands of the feature selection itself can be addressed.

References

[CLV07] C. A. Coello Coello, G. B. Lamont, and D. A. Van Veldhuizen. *Evolutionary Algorithms for Solving Multi-Objective Problems*. Springer, New York, 2007.

[EBN05] M. Emmerich, N. Beume, and B. Naujoks. An EMO Algorithm Using the Hypervolume Measure as Selection Criterion. In C. A. Coello Coello, A. H. Aguirre, and E. Zitzler, editors, *Proceedings of the 3rd Conference on Evolutionary Multi-Criterion Optimization (EMO)*, volume 3410 of *Lecture Notes in Computer Science*, pages 62–76, 2005.

[EHM00] C. Emmanouilidis, A. Hunter, and J. MacIntyre. A Multiobjective Evolutionary Setting for Feature Selection and a Commonality-Based Crossover Operator. In *Proceedings of the IEEE Congress on Evolutionary Computation (CEC)*, volume 1, pages 309–316, 2000.

[FF06] R. Fiebrink and I. Fujinaga. Feature Selection Pitfalls and Music Classification. In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR)*, pages 340–341, 2006.

[Fuj98] I. Fujinaga. Machine Recognition of Timbre Using Steady-State Tone of Acoustic Musical Instruments. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 207–210, 1998.

[GCK03] M. Grimaldi, P. Cunningham, and A. Kokaram. An Evaluation of Alternative Feature Selection Strategies and Ensemble Techniques for Classifying Music. In *Proceedings of the Workshop on Multimedia Discovery and Mining hold within the 14th European Conference on Machine Learning (ECML)/the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, 2003.

[GNGZ06] I. Guyon, M. Nikravesh, S. Gunn, and L. A. Zadeh, editors. *Feature Extraction. Foundations and Applications*, volume 207 of *Studies in Fuzziness and Soft Computing*. Springer, Berlin Heidelberg, 2006.

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2009.

[Hur02] D. Huron. *Listening Styles and Listening Strategies. http://www.musiccog.ohio-state.edu/Huron/Talks/SMT.2002/handout.html*, 2002. Online resource, date of visit: 30.04.2013.

[HW99] M. Hollander and D. A. Wolfe. *Nonparametric Statistical Methods*. Wiley-Interscience, New York, 1999.

[JPE02] M. Jelasity, M. Preuß, and A. E. Eiben. Operator Learning for a Problem Class in a Distributed Peer-to-peer Environment. In J. J. Merelo Guervós, P. Adamidis, H.-G. Beyer, J. L. Fernández-Villacañas Martín, and H.-P. Schwefel, editors, *Proceedings of the 7th International Conference on Parallel Problem Solving from Nature (PPSN)*, volume 2439 of *Lecture Notes in Computer Science*, pages 172–183. Springer, 2002.

[KJ97] R. Kohavi and G. H. John. Wrappers for Feature Subset Selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

[KS00] M. Kudo and J. Sklansky. Comparison of Algorithms that Select Features for Pattern Classifiers. *Pattern Recognition*, 33(1):25–41, 2000.

[Lar12] O. Lartillot. *MIRtoolbox 1.4 User's Manual*. Finnish Centre of Excellence in Interdisciplinary Music Research and Swiss Center for Affective Sciences, 2012. Online resource, date of visit: 30.04.2013.

[MD10] M. Mauch and S. Dixon. Approximate Note Transcription for the Improved Identification of Difficult Chords. In J. S. Downie and R. C. Veltkamp, editors, *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, pages 135–140, 2010.

[ME10] M. Müller and S. Ewert. Towards Timbre-Invariant Audio Features for Harmony-Based Music. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):649–662, 2010.

[MM05] I. Mierswa and K. Morik. Automatic Feature Extraction for Classifying Audio Data. *Machine Learning Journal*, 58(2-3):127–149, 2005.

[MN09] R. Martin and A. M. Nagathil. Cepstral Modulation Ratio Regression (CMRARE) Parameters for Audio Signal Analysis and Classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 321–324. IEEE, 2009.

[PC00] F. Pachet and D. Cazaly. A Taxonomy of Musical Genres. In J.-J. Mariani and D. Harman, editors, *Proceedings of the 6th International Conference on Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications, RIAO)*, pages 1238–1245, 2000.

[RBK12] G. Rozenberg, T. Bäck, and J. N. Kok, editors. *Handbook of Natural Computing*. Springer, Berlin Heidelberg, 2012.

[SEL11] P. Saari, T. Eerola, and O. Lartillot. Generalizability and Simplicity as Criteria in Feature Selection: Application to Mood Classification in Music. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1802–1812, 2011.

[Tor06] K. Torkkola. Information-Theoretic Methods. In I. Guyon, M. Nikravesh, S. Gunn, and L. A. Zadeh, editors, *Feature Extraction*, volume 207 of *Studies in Fuzziness and Soft Computing*, pages 167–185. Springer, Berlin Heidelberg, 2006.

[VBRW12] I. Vatolkin, B. Bischl, G. Rudolph, and C. Weihs. Statistical Comparison of Classifiers for Multi-Objective Feature Selection in Instrument Recognition. In M. Spiliopoulou and L. Schmidt-Thieme, editors, *Proceedings of the 36th Annual Conference of the German Classification Society (GfKl), 2012*, 2012. to appear.

[VPR11] I. Vatolkin, M. Preuß, and G. Rudolph. Multi-Objective Feature Selection in Music Genre and Style Recognition Tasks. In N. Krasnogor and P. L. Lanzi, editors, *Proceedings of the 13th Annual Genetic and Evolutionary Computation Conference (GECCO)*, pages 411–418. ACM Press, 2011.

[VPR+12] I. Vatolkin, M. Preuß, G. Rudolph, M. Eichhoff, and C. Weihs. Multi-Objective Evolutionary Feature Selection for Instrument Recognition in Polyphonic Audio Mixtures. *Soft Computing*, 16(12):2027–2047, 2012.

[VTB10] I. Vatolkin, W. Theimer, and M. Botteck. AMUSE (Advanced MUSic Explorer) - A Multitool Framework for Music Data Analysis. In J. S. Downie and R. C. Veltkamp, editors, *Proceedings of the 11th International Society on Music Information Retrieval Conference (ISMIR)*, pages 33–38, 2010.

[VTB12] I. Vatolkin, W. Theimer, and M. Botteck. Partition Based Feature Processing for Improved Music Classification. In W. A. Gaul, A. Geyer-Schulz, L. Schmidt-Thieme, and J. Kunze, editors, *Proceedings of the 34th Annual Conference of the German Classification Society (GfKl), 2010*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 411–419, Berlin Heidelberg, 2012. Springer.

[VTR09] I. Vatolkin, W. Theimer, and G. Rudolph. Design and Comparison of Different Evolution Strategies for Feature Selection and Consolidation in Music Classification. In *Proceedings of the IEEE Congress on Evolutionary Computation (CEC)*, pages 174–181, 2009.

[WF05] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, San Francisco, 2005.