Multi-Objective Feature Selection in Music Genre and Style Recognition Tasks

Igor Vatolkin
TU Dortmund
Faculty of Computer Science
Chair of Algorithm Engineering
igor.vatolkin@tu-dortmund.de

Mike Preuß
TU Dortmund
Faculty of Computer Science
Chair of Algorithm Engineering
mike.preuss@tu-dortmund.de

Günter Rudolph
TU Dortmund
Faculty of Computer Science
Chair of Algorithm Engineering
guenter.rudolph@tu-dortmund.de

ABSTRACT

Feature selection is an important prerequisite for music classification which in turn is becoming more and more ubiquitous since entering the digital music age. Automated classification into genres or even personal categories is currently envisioned even for standard mobile devices. However, classifiers often fail to work well with all available features, and simple greedy methods often fail to select good feature sets, making feature selection for music classification a natural field of application for evolutionary approaches in general, and multi-objective evolutionary algorithms in particular. In this work, we study the potential of applying such a multi-objective evolutionary optimization algorithm for feature selection with different objective sets. The result is promising, thus calling for deeper investigations of this approach.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Selection process, Retrieval models; H.5.1 [**Multimedia Information Systems**]: Evaluation/methodology—*multi-objective evaluation*

General Terms

Algorithms

Keywords

Multi-Objective Optimization of Data Mining, Feature Selection, Music Information Retrieval

1. INTRODUCTION

Music classification is one of the major topics in the Music Information Retrieval (MIR) [Ras and Wieczorkowska, 2010] research field. Its target is to create the categorization models which label the given music data depending on the previously extracted and analyzed features. Several steps required for performing this process are depicted in Figure 1.

The starting point for any classification approach is the calculation of the numerical characteristics of the given data. Signal analysis transforms, extraction of spectral and time domain characteristics [Peeters, 2004, Lartillot and Toiviainen, 2007] or high-level harmony analysis [Müller and Ewert, 2010] are examples of the methods involved in this step. After gathering the raw features, different processing steps are required: normalization, substitution of undefined values, time series analysis for feature consolidation etc. Some of the algorithms reduce the data amount, e.g. in the course of a principal component analysis. Other may even create new dimensions, e.g. calculation of derivatives or moments of feature series. The last and often most visible step is the training and application of classification models. They can be created either by supervised learning guided by the given ground truth (labeled data), or by unsupervised algorithms, e.g. clustering.

Previous works in music feature selection confirmed the suggestion that too large features sets not only slow down the classifiers, but also diminish the obtained quality ([Vatolkin et al., 2009, Bischl et al., 2010]). Pursuing two aims—e.g. high quality and low feature set size—at the same time naturally brings up the idea to apply a multi-objective optimization algorithm. There are many other interesting objective tradeoffs depending on the actually chosen application scenario. We therefore investigate the potential of multi-objective feature selection in two different scenarios and also aim at generalizing the findings of our experimental analysis to rules-of-thumb which may be useful for other researchers.

We are firstly going to provide some background in music feature selection, taking into account related approaches and evaluation metrics. Next, we discuss different application scenarios that may be worthwhile investigating from a multi-objective perspective. This is followed by the setup of our experimental study, and the discussion of results obtained for both scenarios.

2. MUSIC FEATURE SELECTION

The number of available music descriptors - either extracted from the audio signal or e.g. based on the metadata / playlist analysis from the web - is very large and growing continuously. On the other side, the music categories to learn may be very different even for the genres, from 'classic against pop' to specific music styles ('progressive symphonic

GECCO'11, July 12–16, 2011, Dublin, Ireland.

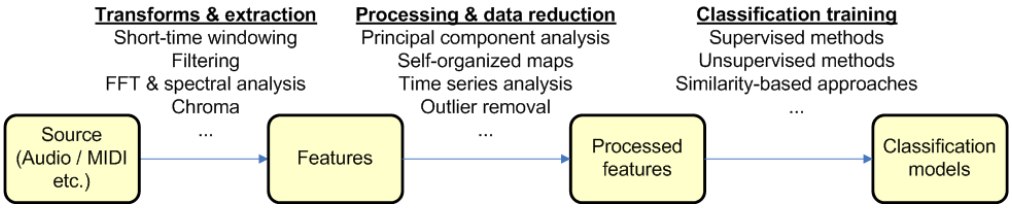

Figure 1: Algorithm Chain in Music Classification

death metal') or user-specific personal categories like 'personal favourites', 'car driving' or numerous available *Last.fm* tags. To enable high performance of the classification models it becomes challenging to select the most important features for the concrete classification task: e.g. the temporal and rhythmic characteristics or the harmonic descriptors may play a role or be completely meaningless depending on the defined category. Another target is to reduce the number of the noisy or highly correlated and thus redundant features. Finally, the smaller feature sets lead to reduced storage and processing time as well as an accelerated classification.

2.1 Related Approaches

For non-trivial classification tasks, feature selection by means of evolutionary algorithms provides a valuable alternative to simple greedy methods as these fail or request a large number of evaluations. Feature selection by genetic algorithm for different classification tasks was applied in [Raymer et al., 2000]. Some ideas concerning the design of hybrid evolutionary feature selection methods are discussed in [Zhu et al., 2010]. A general and detailed methodology of feature selection techniques is introduced in [Guyon et al., 2006]. However, these methods are still rarely applied in music categorization, let alone investigated in detail. In [Fujinaga, 1998], a genetic algorithm was applied as feature selection technique for instrument identification. A hybrid approach was investigated in [Vatolkin et al., 2009] and it was stated that the classification with full feature sets led to a significant decrease of decision tree performance. The best found feature sets formed a compromise between very small, not sufficient, and larger feature groups. In [Bischl et al., 2010], further improvements have been suggested (asymmetric mutation, success rule adaptation, greedy heuristics etc.) and several algorithms have been compared. Another approach incorporated the generation of very large feature sets optimized for the different music categories by means of genetic programming [Mierswa and Morik, 2005].

2.2 Evaluation Metrics

The performance of the applied feature selection method must be thoroughly measured. Keeping in mind the special case of music classification evaluation, we propose the following categorization of convenient metrics which are to be optimized with a multi-objective approach.

Common quality-based metrics are calculated based on the confusion matrix data. Often used measures are accuracy, precision, recall, specificity and f-measure. Let TP be the number of true positives (music songs which belong to the category and classified as belonging to it), TN the number of true negatives (songs not belonging to category and classified as not belonging to it), FP - false positives (songs not belonging to category but classified as belonging to it) and FN - false negatives (songs belonging to category but classified as not belonging to it). The following formulas describe the measures which are optimized in our experimental study:

$$recall = \frac{TP}{TP + FN} \tag{1}$$

$$specificity = \frac{TN}{FP + TN} \tag{2}$$

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

Especially the expectations on the balance between the performance on positive and negative examples can differ (see Sect. 3). If the given data is strongly unbalanced—which can be the case for large music collections—metrics from medicine applications can be valuable [Sokolova et al., 2006].

Specific quality-based metrics are useful for applications that do not allow to measure performance directly or where specific aspects of the task shall be examined, e.g. song segmentation [Lukashevich, 2008] or audio synchronisation [Fremerey et al., 2010] evaluation measures. For the music genre and category classification the design of new metrics can be also promising. Consider an automatic recommendation system that presents the sorted playlist to several users at once (e.g. on the dancefloor). Here, the measured diversity of the songs in a given time interval becomes very important, so that the most of the audience (which may have different tastes) is satisfied.

Resource metrics consider algorithm hardware demands. Dealing with runtime and storage requirements is unavoidable if music classification is employed in practice. However, they are not easily measured. Even if the runtime is given in some works, it is almost impossible to compare the results across different studies because of the broad variety of available hardware. Even on the same machine, a metric can provide different results depending on CPU load from operating system activities. One possibility is to use profiling software as used e.g. in [Seppänen et al., 2006] for the calculation of CPU cycles during beat tracking.

Model complexity metrics measure the balance between the highly aligned and possibly overfitted complex models against more simple classification models which are better suited for general performance. One of the easiest possibilities is to measure the rate of the selected features. Let N be the number of all and m of selected features. Then:

$$feature\ rate\ f_r = \frac{m}{N} \tag{4}$$

The models built with a very large feature number tend

to be overfitted and also may require larger efforts to apply them on uncategorized data. The comparison of the metric on training and test sets can measure the general performance of the model. Classifier-specific measures can be defined: e.g. for SVM, a successful approach for the optimization of tradeoff between the maximization of the separating margin and the minimization of the training error number was developed in [Mierswa, 2007].

User interaction metrics are useful if interaction with users or their influence on the classification must be evaluated. Since only a small part of the music classification studies deals with this metric group, the measures are often (re-)defined for particular applications, making it hard to compare the performance across multiple investigations. An extensive list is of these metric is provided in [Liu and Hu, 2010]. If user interaction is considered, the classifier system may ask the user for the rating of songs where the direct categorization is doubtful. This is e.g. useful for songs which belong to clusters which are not contained in the training examples. So the tradeoff between classification quality and the number of user interactions can be measured.

3. MULTI-OBJECTIVE APPROACH

The above discussed metric groups with often contradictory targets enable a multi-objective evaluation. The tradeoff between low computation time against high accuracy can be considered, or the general performance of the classification models can be positioned against the quality-based measures. Even if only the confusion matrix-based measures are taken into account, they may be fairly uncorrelated as was shown in [Vatolkin, 2010] for music classification tasks.

The balance between surprise effect and safety can differ between listeners. A classification model recommending only user-preferred songs (with the highest true negative rate) can be unacceptable for another listener who wants to be surprised by slightly different music from time to time. It can become boring if the algorithm with the highest accuracy recommends too many songs from the same album or artist — the similarity criterion can reach the ceiling of its advantage for a human listener.

The impact of algorithm runtime and memory demands differs from application to application: if the classification must be run on mobile devices, the limited capacities and processor power must be beared in mind [Blume et al., 2008]. Or consider again the automatic song structure identification: this task can be done offline on a server farm run by a music vendor or must be done in real time if applied to songs played by less known bands or during a live performance.

Especially the human aspects can lead to further promising evaluation criteria: if a listener is willing to adjust the models by interaction then an active learning scheme can be applied [Huang et al., 2008]. For this case, less powerful models with little training can be suitable. On the other hand, rating music for a long time is tiring. A listener who wants an algorithm to learn a personal category may decide between the lower number of training examples and the higher accuracy based on a larger number of examples which must be provided to the classification software. Since it requires more effort to find negative (and also not similar!) examples during category learning, lower accuracy results can be also accepted if one-class learning methods are applied [Tax, 2001].

The real situation is even worse: if we run the empirical studies and demonstrate that a certain method chain with certain parameter settings is the *first choice* for some combination of useful metrics, it will not help a certain user or decision maker because her or his preferences may be very different. This aspect is rarely mentioned in current research on music classification and we hope that it will gain more attention in future.

Application of multi-objective optimization to music classification seems to be fairly unexplored until present. If we take a sight out of MIR but remain in the data mining area, some publications are available. The multi-objective scenario for the evaluation of feature selection in three different data sets (physical, medicine and texture) is investigated in [Reynolds et al., 2010]. Multi-objective tuning of classifiers is applied in [Mugambi and Hunter, 2003] for decision trees and in [Mierswa, 2007] for Support Vector Machines (SVM). Strategies for avoiding model overfitting and complexity are developed in [Radtke et al., 2009, Mierswa, 2007].

4. EXPERIMENTAL STUDY DESIGN

4.1 Classification Tasks

We selected 6 AllMusicGuide genres and categories (sorted by increasing complexity): Classical, Pop, Rap, Heavy Metal, Electronic and R&B. The complexity was measured by the mean accuracy after a large number of classification trials with randomly selected features. Each categorization task was binary: e.g. classical music pieces against all other songs.

The classification models were built from the 20 or 10 music tracks which represented the corresponding category. These comparatively small sets were motivated by the application situation - the user does not want to adjust the classification system for a long time, and, on the other side, human listeners can capture very well the music style from a small number of examples.

The trained models were optimized on a set of 120 music tracks randomly chosen from 120 music albums[1]. Although the performance progress against the evaluation number on this optimization set can be clearly seen, the more important criterion is the performance on the independent test set that was not involved in the feature selection procedure. For the test set we used also 120 songs from the same albums, but disjoint from the ones used for training. The idea to use a test set instead of the common known cross-validation evaluation is discussed in [Fiebrink and Fujinaga, 2006] in terms of music classification; this work is motivated by general discussion of the overfitting danger from [Reunanen, 2003].

4.2 Algorithm Setup

We employed an initial 286-dimensional audio feature vector with temporal, spectral, phase domain and cepstral characteristics. Some of these are low-level (e.g. zero-crossing rate or spectral centroid), other correspond to the high-level musical descriptors (tempo, chroma harmony statistics, tonal centroid etc.). Most features are described in detail in [Peeters, 2004, Müller and Ewert, 2010, Martin and Nagathil, 2009] and the user manual of the MIR Toolbox [Lartillot and Toiviainen, 2007] and have been extracted with AMUSE framework [Vatolkin et al., 2010].

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

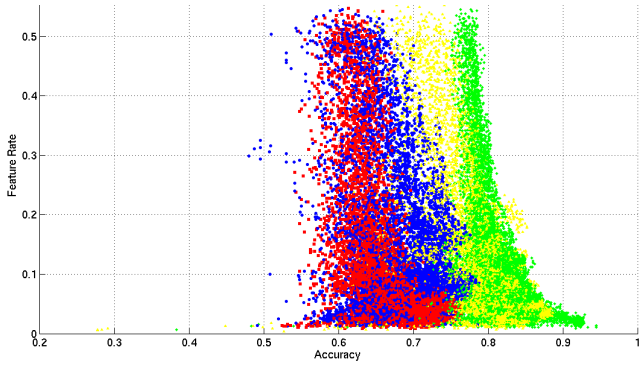

Figure 2: All solutions evaluated on the test set for category Heavy Metal

After the initial extraction, we replaced undefined values by the mean and selected only the time frames from the middle between the onset events. Then mean and standard deviation were calculated for each feature dimension over 4 second partitions with 1/2 overlap, so that the final feature set cardinality was increased to 572.

Four classification methods were used for the model training: decision tree C4.5, Random Forest (RF), Naive Bayes (NB) and SVM with linear kernel [Bishop, 2006]. Tuning of the classifier parameters was not the target of our investigations; however it was important in our opinion to distinguish between several classification methods. Some algorithms such as decision trees have already integrated feature selection techniques, other (SVM) increase the dimensionality of feature space. Therefore we expected different performance characteristics during the feature selection process.

For the multi-objective feature selection we implemented the S-Metric Selection Evolutionary Multiobjective Algorithm (SMS-EMOA) [Beume et al., 2007]. The population size was set to $p = 30$ individuals. The number of the evaluations was limited to 5000 and we run $r = 5$ repeats for each combination of category task, classification algorithm and target metrics. These restrictions were made because of the long optimization runs: training and evaluation of one classification model requires up to about 30 seconds depending on classification method and feature number. Asymmetric mutation [Jelasity et al., 2007] was applied with $p_{01} = 0.1$ and the mutation bit probability $\gamma = 32/N$, $N = 572$. The initialization was done randomly.

The following two-objective cases have been considered:

- (a) Optimization of recall (1) and specificity (2)
- (b) Optimization of accuracy (3) and the rate of the selected features (4)

5. DISCUSSION OF RESULTS

5.1 Overall Performance on the Test Set

Fig. 2 contains an example plot of all solutions generated during all runs for one category. C4.5 runs are marked by blue circles, RF by red squares, NB by green diamonds and SVM by yellow triangles. The plot describes the structure of the optimization problem. It can be seen, that the increasing number of the selected features leads to an accuracy drop; the best accuracies are achieved for feature rates below 5 percent. This underlines the statement that the classification algorithms are overstrained by too much features.

Table 1: $N(\mathcal{X}, \mathcal{Y})$, in percent ($\mathcal{X}$: classifiers in rows; $\mathcal{Y}$: classifiers in columns)

Recall vs. Specificity				
	C4.5	RF	NB	SVM
C4.5	x	31.22	75.33	20.27
RF	1.07	x	8.95	0
NB	0	0.03	x	0.02
SVM	5.68	3.00	63.47	x
Accuracy vs. Feature Rate				
	C4.5	RF	NB	SVM
C4.5	x	19.61	55.3	49.71
RF	7.41	x	38.5	33.33
NB	0	0.01	x	8.15
SVM	0.12	29.47	0.45	x

For the further discussion of results we concentrate only on the individuals of the last population after 5000 SMS-EMOA evaluations. Because of the metric calculation on the independent test set it can not be guaranteed, that these solutions contain also all non-dominated solutions - but they are presented to the decision maker after the optimization, and they help to understand the generalization performance of the SMS-EMOA-driven feature selection.

For the comparison of performance across the different classifiers we can calculate the non-dominance relation between the solutions of two algorithms. Let x_i be the i-th solution of the last front of the classifier $\mathcal{X}$ and y_j the j-th of the classifier $\mathcal{Y}$. Then we can calculate the mean percent number of solutions generated by classifier $\mathcal{Y}$ which dominate the mean solution of classifier $\mathcal{X}$:

$$N(\mathcal{X}, \mathcal{Y}) := \frac{1}{p \cdot r} \cdot \sum_{i=1}^{p \cdot r} \left(\frac{1}{p \cdot r} \cdot \sum_{j \in \{1,\ldots,p \cdot r\}; x_i \prec y_i} 1 \right) \quad (5)$$

Table 1 lists the $N(\mathcal{X}, \mathcal{Y})$-relation of classifiers for category Heavy Metal [2]. It is interesting to see for recall and specificity optimization runs, that *no* single SVM solution dominates RF solutions of the last five fronts. The same holds for C4.5 and NB. Due to the stochastic characteristics of optimization runs one can not guarantee, that a certain classifier will always outperform another one: e.g. 31.22% of RF solutions dominate on average the mean C4.5 solution; however some C4.5 solutions exist which dominate RF, even if the correspoding $N(\mathcal{X}, \mathcal{Y}) = 1.07$ is rather small.

If we average the $N(\mathcal{X}, \mathcal{Y})$-values over the rows of the Table 1, we shall get $\widehat{N}(\mathcal{X}) \in [0; 1]$, which measures approximately, how often the solutions of the current classifier $\mathcal{X}$ are dominated by the other classifier solutions (6), $c = 4$ is the number of classifiers. $\widehat{N}(\mathcal{X}) = 0$ means: it can be expected, that the classifier solutions are not dominated by any solutions of other classifiers.

$$\widehat{N}(\mathcal{X}_i) := \frac{1}{c-1} \cdot \sum_{j \in \{1,\ldots,c\} \setminus i} N(\mathcal{X}_i, \mathcal{Y}_j) \quad (6)$$

[2]Further tables and figures: http://ls11-www.cs.tu-dortmund.de/_media/rudolph/gecco2011supplementary.zip

Table 2: $\widehat{N}(\mathcal{X})$, in %; n: music category number

Recall vs. Specificity				
	C4.5	RF	NB	SVM
Classical	12.88	7.63	24.95	6.69
Pop	22.83	1.1	2.43	3.76
Rap	24.81	6.23	5.99	1.51
Heavy Metal	42.28	3.34	0.02	24.05
Electronic	35.8	9.23	6.4	0.05
R&B	6.39	3.61	4.19	14.00
$\frac{1}{n} \cdot \sum_{i=1}^{n} \widehat{N}(\mathcal{X}_i)$	24.17	5.19	7.33	8.34
Accuracy vs. Feature Rate				
Classical	19.44	15.64	8.49	8.98
Pop	41.06	11.52	2.14	6.2
Rap	18.3	16.77	2.13	8.06
Heavy Metal	41.54	26.41	2.72	10.01
Electronic	32.05	20.85	1.96	3.95
R&B	47.83	28.4	0.11	19.34
$\frac{1}{n} \cdot \sum_{i=1}^{n} \widehat{N}(\mathcal{X}_i)$	33.37	19.93	2.93	9.42

The results in table 2 summarize $\widehat{N}(\mathcal{X})$ for all categories. Since no zero entries are existing and no clear winner can be seen, we conclude that it makes sense to involve several different classifiers during the feature selection process. Another observation is that the 'best' algorithms are not always the same for the different optimization tasks for a certain category. After the averaging of the results across all categories, NB, RF and SVM provide the best contribution to the non-dominated solutions, followed by C4.5.

5.2 Best Fronts on the Test Set

Figs. 3 and 4 list only the non-dominated solutions from the last populations over 5 repeats. For the optimization of recall and specificity, it can be clearly distinguished between the complexities of the different categories. However, it is very difficult to name the best classification methods: e.g. RF is required for the recall-high regions of the Pareto front for the most simple category Classical as well as for the rather complex R&B. The same region is dominated by NB solutions for Pop and Electronic. The central regions of the Pareto fronts contain often either NB solutions (as for Heavy Metal and R&B) or SVM (Rap, Electronic, Pop). C4.5 seems to be the worst algorithm and achieves rather seldomly the best front. It can be also observed, that RF solutions often outperform C4.5 concerning recall, whereas some single C4.5 tree models are usually better than RF concerning specificity except for Rap.

The classifier impact can differ depending on optimization criterions: e.g. for Electronic and accuracy vs. feature rate the largest part of the Pareto front is specified by NB solutions; for recall vs. specificity it is dominated by SVM. NB and SVM are often the only members of the Pareto front for more complex categories for accuracy vs. feature rate runs. However the 'complexity' ranking is not easy - if the optimization process is applied for a new user-predefined category, it is very hard to recommend the choice of appropriate classifiers.

5.3 Analysis of Selected Features

Of course, we are also interested in the concrete features chosen during optimization. If a robust feature set with stable best performance over different categories can be found, further optimization of feature selection does not make sense any more. However Fig. 5 underlines another situation. Here we calculated how often the features have been selected for different categories and different classifiers across all corresponding runs. The highest selected rates are marked by dark red color. The frequently selected features are not the same for the different categories and classification methods. For the latter, some of them are more or less equally distributed among all classifiers, and some of them are preferable for particular methods. E.g. the mean of the 5th fluctuation pattern characteristic (ID 547) is often selected by RF and SVM; NB selected frequently the standard deviation of MFCC 1 (ID 84) and the standard deviation of the 1st bark scale magnitude (ID 186).

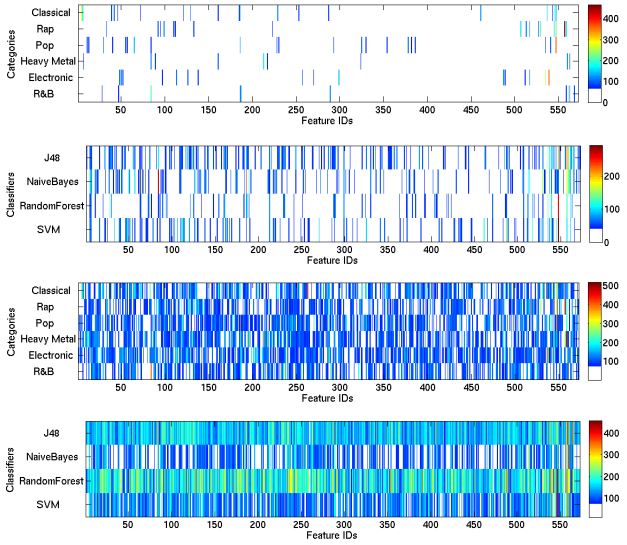

Figure 5: Selected features after the optimization. Both upper subfigures: accuracy vs. feature rate runs; both bottom subfigures: recall vs. specificity

For the recall vs. specificity runs, the differences are not that evident. Here, feature number reduction was not a target in the optimization process, and it seems that many different features are responsible for the creation of tradeoff solutions. Again, no clear tendencies for the feature role in the categorization can be seen.

6. SUMMARY AND OUTLOOK

In our experimental study we applied a multi-objective approach to feature selection for several different music categorization tasks. The analysis of the Pareto fronts after the large number of SMS-EMOA evaluations supports the suggestion that the calculation of at least two objectives makes sense for these tasks. No clear winner can be stated for the four classification methods; all of them produce contributions for Pareto fronts for at least some of the categories. However, by leaving out C4.5, not much quality would be lost. The selected features are also different depending on category and classifier. Therefore, it is reasonable to apply feature selection each time a listener defines a new category and wishes to optimize the preferred metrics.

In future, we want to investigate further metric combi-

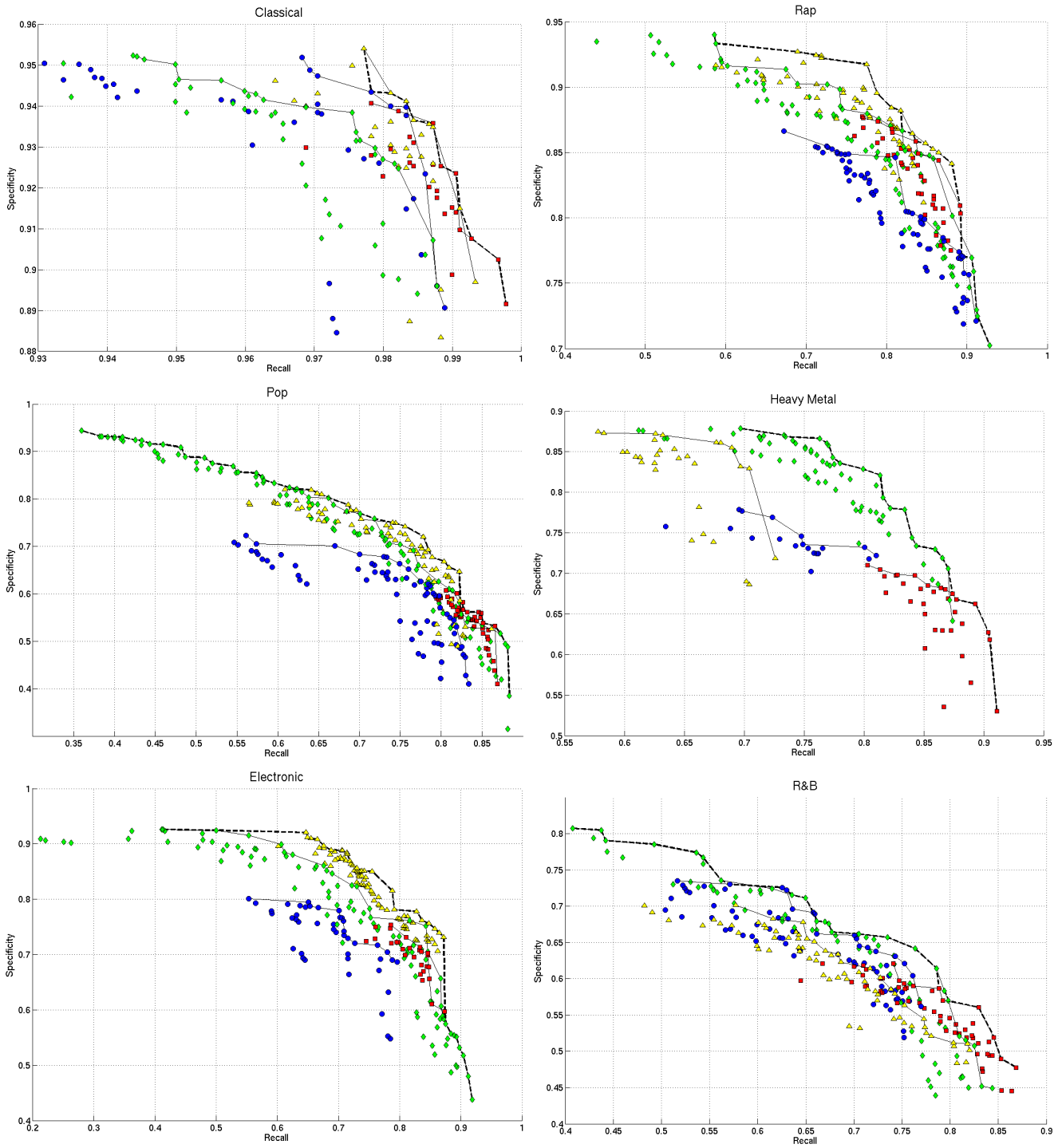

Figure 3: Non-dominated solutions of the last populations. Best fronts for each classifier across all repeats are marked by lines. The Pareto front built from all runs is marked by thick dashed line

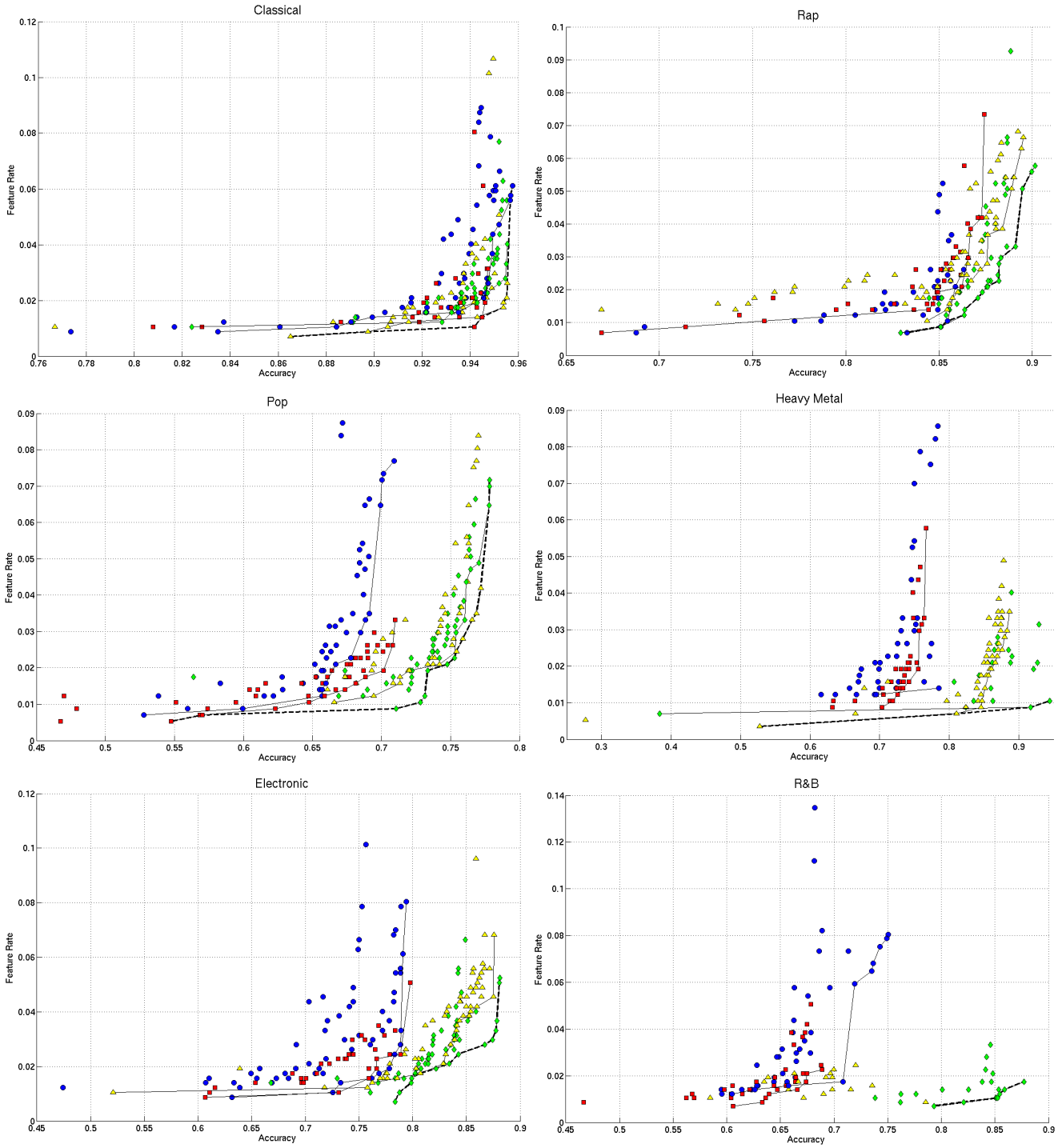

Figure 4: Non-dominated solutions of the last populations. Best fronts for each classifier across all repeats are marked by lines. The Pareto front built from all runs is marked by thick dashed line

nations (cf. Sect. 2.2). The optimization of more than two objectives is another promising approach. Other possibilities are to adjust the parameters of the classification chain – enlarge the feature number, tune the classification method parameters, or solve other tasks such as instrument or harmony recognition. Since the application area of the multi-objective optimization for music data analysis tasks is currently not investigated in detail, we hope that we can motivate other researchers from both MIR and optimization domain to explore it further.

7. ACKNOWLEDGMENTS

This work was partly supported by Klaus Tschira Foundation.

8. REFERENCES

[Beume et al., 2007] Beume, N., Naujoks, B., and Emmerich, M. (2007). SMS-EMOA: Multiobjective selection based on dominated hypervolume. *European Journal of Operational Research*, 181(3):1653–1669.

[Bischl et al., 2010] 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)*, pages 314–323. Springer.

[Bishop, 2006] Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

[Blume et al., 2008] Blume, H., Haller, M., Botteck, M., and Theimer, W. (2008). Perceptual feature based music classification - a dsp perspective for a new type of application. In *Proc. of the 8th Int'l Symp. on Systems, Architectures, Modeling and Simulation (SAMOS)*, pages 92–99.

[Fiebrink and Fujinaga, 2006] Fiebrink, R. and Fujinaga, I. (2006). Feature selection pitfalls and music classification. In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR)*, pages 340–341.

[Fremerey et al., 2010] Fremerey, C., Müller, M., and Clausen, M. (2010). Handling repeats and jumps in score-performance synchronization. In *Proc. of the 11th Int'l Conf. on Music Information Retrieval (ISMIR)*, pages 243–248.

[Fujinaga, 1998] Fujinaga, I. (1998). Machine recognition of timbre using steady-state tone of acoustic musical instruments. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 207–210. ICMA.

[Guyon et al., 2006] Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. (2006). *Feature Extraction, Foundations and Applications*. Springer.

[Huang et al., 2008] Huang, T., Dagli, C., Rajaram, S., Chang, E., Mandel, M., Poliner, G., and Ellis, D. (2008). Active learning for interactive multimedia retrieval. *Proceedings of the IEEE*, 96(4):648–667.

[Jelasity et al., 2007] Jelasity, M., Preuß, M., and Eiben, A. (2007). Operator learning for a problem class in a distributed peer-to-peer environment. In *Proc. of the 7th Parallel Problem Solving from Nature (PPSN)*, pages 172–183.

[Lartillot and Toiviainen, 2007] Lartillot, O. and Toiviainen, P. (2007). Mir in matlab (ii): A toolbox for musical feature extraction from audio. In *Proc. of the 8th Int'l Conf. on Music Information Retrieval (ISMIR)*, pages 127–130.

[Liu and Hu, 2010] Liu, J. and Hu, X. (2010). User-centered music information retrieval evaluation. In *Proc. of the Joint Conf. on Digital Libraries (JCDL) Workshop: Music Information Retrieval for the Masses*.

[Lukashevich, 2008] Lukashevich, H. (2008). Towards quantitative measures of evaluating song segmentation. In *Proc. of the 9th Int'l Conf. on Music Inform. Retr. (ISMIR)*, pages 375–380.

[Martin and Nagathil, 2009] Martin, R. and Nagathil, A. (2009). Cepstral Modulation Ratio Regression (CMRARE) Parameters for Audio Signal Analysis and Classification. In *Proc. of the Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 321–324. IEEE Press.

[Mierswa, 2007] Mierswa, I. (2007). Controlling overfitting with multi-objective support vector machines. In *Proc. of the Gen. and Evol. Comput. Conf. (GECCO)*, pages 1830–1837. ACM.

[Mierswa and Morik, 2005] Mierswa, I. and Morik, K. (2005). Automatic feature extraction for classifying audio data. *Machine Learning Journal*, 58(2-3):127–149.

[Mugambi and Hunter, 2003] Mugambi, E. and Hunter, A. (2003). Multi-objective genetic programming optimization of decision trees for classifying medical data. In *Proc. of the 7th Int'l Conf. on Knowledged-Based Intelligent Information and Engineering Systems (KES)*, pages 293–299.

[Müller and Ewert, 2010] Müller, M. and Ewert, S. (2010). Towards timbre-invariant audio features for harmony-based music. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):649–662.

[Peeters, 2004] Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the cuidado project. Technical report, IRCAM, France.

[Radtke et al., 2009] Radtke, P., Sabourin, R., and Wong, T. (2009). Solution over-fit control in evolutionary multi-objective optimization of pattern classification systems. *Int'l Journal of Pattern Recognition and Artif. Intelligence*, 23(6):1107–1127.

[Ras and Wieczorkowska, 2010] Ras, Z. and Wieczorkowska, A. (2010). *Advances in Music Information Retrieval*. Springer, Berlin.

[Raymer et al., 2000] Raymer, M., Punch, W., Goodman, E., Kuhn, L., and Jain, A. (2000). Dimensionality reduction using genetic algorithms. *IEEE Transactions on Evolutionary Computation*, 4(2):164–171.

[Reunanen, 2003] Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research*, (3):1371–1382.

[Reynolds et al., 2010] Reynolds, A., Corne, D., and Chantler, M. (2010). Feature selection for multi-purpose predictive models: A many-objective task. In *Proc. of the 11th Int'l Conf. on Parallel Problem Solving From Nature (PPSN)*, pages 384–393. Springer.

[Seppänen et al., 2006] Seppänen, J., Eronen, A., and Hiipakka, J. (2006). Joint beat and tatum tracking from music signals. In *Proc. of the 7th Int'l Conf. on Music Information Retrieval (ISMIR)*, pages 23–28.

[Sokolova et al., 2006] Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006). Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Proc. of the AAAI'06 workshop on Evaluation Methods for Machine Learning*, pages 1015–1021.

[Tax, 2001] Tax, D. (2001). *One-Class Classification*. PhD thesis, Delft University of Technology, Pattern Recognition Laboratory.

[Vatolkin, 2010] Vatolkin, I. (2010). Multi-objective evaluation of music classification. In *Proc. of the German Classification Society Conference (GfKl)*.

[Vatolkin et al., 2010] Vatolkin, I., Theimer, W., and Botteck, M. (2010). Amuse (advanced music explorer) - a multitool framework for music data analysis. In *Proc. of the 11th Int'l Society for Music Inform. Retr. Conf. (ISMIR)*, pages 33–38.

[Vatolkin et al., 2009] Vatolkin, I., Theimer, W., and Rudolph, G. (2009). Design and comparison of different evolution strategies for feature selection and consolidation in music classification. In *Proc. of the 2009 IEEE Congress on Evolutionary Computation (CEC)*, pages 174–181.

[Zhu et al., 2010] Zhu, Z., Jia, S., and Ji, Z. (2010). Towards a memetic feature selection paradigm. *IEEE Computational Intelligence Magazine*, 5(2):41–53.