technische universität
dortmund

Training Set Reduction Based on 2-Gram Feature Statistics for Music Genre Recognition

Igor Vatolkin
Mike Preuß
Günter Rudolph

Algorithm Engineering Report
TR13-2-001
May 2013
ISSN 1864-4503

Faculty of Computer Science
Algorithm Engineering (Ls11)
44221 Dortmund / Germany
http://ls11-www.cs.uni-dortmund.de/

Training Set Reduction Based on 2-Gram Feature Statistics for Music Genre Recognition

Igor Vatolkin, Mike Preuß, Günter Rudolph
Chair of Algorithm Engineering
Department of Computer Science
TU Dortmund
{igor.vatolkin;mike.preuss;guenter.rudolph}@tu-dortmund.de

Abstract

Too large instance and/or feature number for supervised classification requires higher storage demands and computing time, and also the classification quality may suffer from too huge datasets. In our work we examine the reduction of training instance number in music genre recognition where each instance is mapped to a class described by a corresponding 2-gram estimated from the statistical distribution of musical characteristics. Two approaches are integrated: The removal of outlier instances and the limitation of the maximal training instance number from each 2-gram. The experiments show that it is possible to keep the classification performance and even improve it in many cases despite the strong reduction of instance number.

1 Introduction

A substantial fraction of music data analysis research deals with classification - either recognizing instruments, harmony and melody characteristics, structure and so on - or categorizing music recordings into genres and styles, related composers and artists etc. This work is very often done by supervised classification [Duda *et al.*, 2000], i.e. one starts from a labeled feature set and creates a computational model which predicts labels deending on the certain feature distribution. One of the easiest possibilities is to calculate an entropy value [Witten and Frank, 2005], which measures, how well a singular feature may distinguish between separate classes - where more complex methods like Support Vector Machines (SVMs) [Bishop, 2006] even increase the initial feature dimensions for the linear separation of different classes.

No supervised training is possible without labeled training set, or ground truth. Not only the classification itself, but also the associated costs are significant for successful model creation [Last, 2009]. One way to reduce the training set is to decrease the feature number by means of feature selection [Guyon *et al.*, 2006]. This procedure provides the following advantages:

- Storage of less features saves disc space.
- Extraction and processing of features as well as training of models becomes faster.
- Not only training but also classification is done faster, and classification of unlabeled data is usually done more frequently after a model is once created.
- Smaller models created from less features are often more robust and less overfitted towards a certain data set, this means increased generalization ability.

Another possibility is to reduce the number of training instances, which may also lead to several benefits:

- Storage of less instances saves disc space.
- Training of models becomes faster.
- If ground truth is created by humans, this task requires less effort and cost for smaller training sets.
- Removal of outlier instances in the labeled data (e.g. quiet parts at the beginning and end of a song) may reduce the classification error.

Since we have already investigated several studies for feature selection in music classification as the first possibility for training set reduction [Vatolkin *et al.*, 2011; 2012], we have decided to examine the success of training instance number reduction. The motivation was not just to reduce the instance number, but to do it in a more intelligent way. This can be done by unsupervised structuring of music instances (which are here the labeled feature sets describing an audio frame of 4 seconds). The instances with the similar sound (e.g. female vocals with guitar background or quiet piano bridge) can be grouped and sorted according to their occurence in the training data set.

Figure 1 illustrates an example for such organization: similar segments A and B are often existent among N training set songs, whereas Y and Z segments are very seldom and can be considered as outliers. After the sorting of segment shares two complemental targets can be approached: At first, we want to remove the rather rare instances from the training set (Y, Z): The removed segments are marked by the shaded shape on the right subfigure. On the other side, we intend to remove a part of similar instances from the training set with are very frequent (e.g. from several similar vocal segments), since the larger number of similar labeled feature sets does not improve learning quality. The removed instances are marked by the dashed area on the right subfigure. If we compare this approach to the 'baseline' method of simply removing an equally distributed number of instances, it is clear that the number of the most frequent instances will be reduced well, but it becomes not possible any more to detect and remove the outlier segments in a systematic way.

In our work we propose a new approach for the reduction of training sets in music genre classification by unsupervised instance structuring based on n-grams. n-grams are the sequences of n symbols developed and well established mainly for text classification area [Suen, 1979] - but

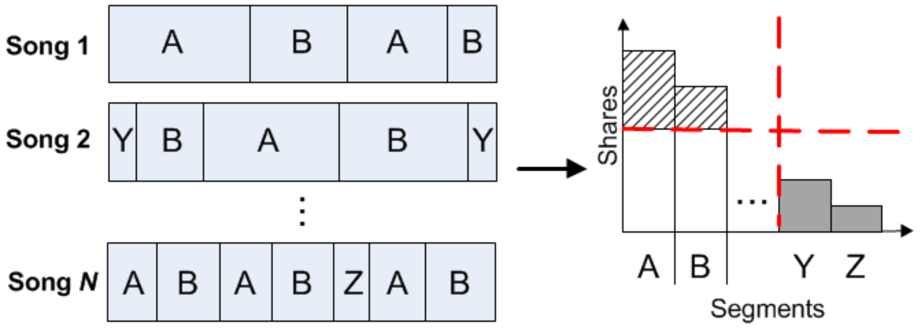

Figure 1: Training instance number reduction by removal of outliers and a part of the most frequent instances.

are also suitable for any other classification tasks, e.g. web page categorization [Mason, 2009]. In [Gao *et al.*, 2000] n-grams are explicitly adopted for removal of outliers.

Music series can be also well mapped to n-gram representations - in one of the first corresponding publications, 'musical words' created from monophonic melodies enabled further application of text-related methods for music retrieval tasks [Downie, 1999]. The mapping of music structure into an n-gram representation is discussed in general in [Müllensiefen, 2009] and is applied to a composer recognition task in [Wołkowicz *et al.*, 2008]. Among other related works it is also worth to mention [Doraisamy and Rüger, 2003], where n-grams were used for indexing of polyphonic recordings, or also the integration of n-grams into chord recognition [Scholz *et al.*, 2009; Sumi *et al.*, 2008].

The target of our study was to examine the possibilities of the n-gram-based training data reduction measuring how far it could be driven keeping or even improving the classification performance at the same time. Several possibilities of instance pre-structuring were developed based on the characteristics typical for especially music data - which however does not restrict the method application to only music classification. The results show that it is indeed possible to select only a rather small fraction of labeled feature sets for successful classification.

The remainder of this paper is organized as follows: At first we outline the complete algorithm chain and introduce in detail the underlying methods for the training set reduction. Then we describe our experimental study and discuss the results after the application of n-gramm-based training set reduction for recognition of three music genres. Finally we conclude with the advantages and challenges of the proposed method and provide a list for future improvements.

2 Algorithm Background

The rough overview of the algorithmic steps is sketched by Fig. 2 and is described in detail in the following subsections. At first, we extract a large set of audio features and preprocess them for further classification. A set of 6 features is selected for later building of n-grams. The building method is based on the optimization of the training set feature statistics. Then the number of instances used for model building is reduced with regard to the both criteria mentioned above: elimination of outliers and reduction of frequent similar instances. Afterwards, the classification models are built based on the reduced training set and features selected to be the most significant for a concrete classification task. Finally, the models are validated on a labeled holdout music song set.

2.1 Feature Extraction and Aggregation

As the source for learning we have extracted a large number of 674 audio features by means of the MIR Toolbox [Lartillot and Toiviainen, 2007], jAudio [McKay, 2010], Yale [Mierswa and Morik, 2005], Chroma Toolbox [Müller and Ewert, 2011] and own implementations. These features correspond to different low-level and high-level signal characteristics such as energy, zero-crossing rate, spectral peak distribution, Mel Frequency Cepstral Coefficients (MFCCs), chroma and tonal centroid vector, fluctuation patterns, rhythmic clarity etc. The values to analyze were saved only from the frames between the previously estimated onset events for description of the more stable sound between the playing notes - this method performed well compared to others [Vatolkin *et al.*, 2010]. Then the mean value and the standard deviation were calculated for intervals of 4s with 2s overlap for the complete songs.

2.2 n-gram Feature Estimation

Our algorithm aggregates the statistics of 6 features for n-gram building. This feature number was manually selected as compromise, so that the number of instance classes will be not too low - in that case no outliers can be identified - or too high - where the strongest instance classes will be represented only by a few number of corresponding instances (see Sect. 2.3 for exact method description). We created 5 different feature sets providing several possibilities to structure the labeled feature sets for this target:

- ENERGY AND TEMPO: Signal RMS, sensory roughness, angles in phase domain, tempo based on onset events, its deviation and rhythmic clarity.
- First 6 MFCCS which describe the cepstrum estimated from the Mel bands with distribution motivated by human perception [Rabiner and Juang, 1993] and are commonly used in speech and music classification.
- 6-dimensional TONAL CENTROID vector [Harte *et al.*, 2006].
- TIMBRE AND HARMONY: Inharmonicity, key and its clarity, normalized energy of harmonic components and two of the first tristimulus values.
- RANDOM SET: In each statistical repetition 6 features were chosen randomly from 674 original features.

2.3 Mapping of Instances to n-Grams

The general mapping of instances to n-grams was implemented as follows: At first we estimated the $min(f_i)$ and $max(f_i)$ value of each n-gram building feature f_i for all instances of the corresponding training set. Then three boundaries b_1, b_2, b_3 were selected between $min(f_i)$ and

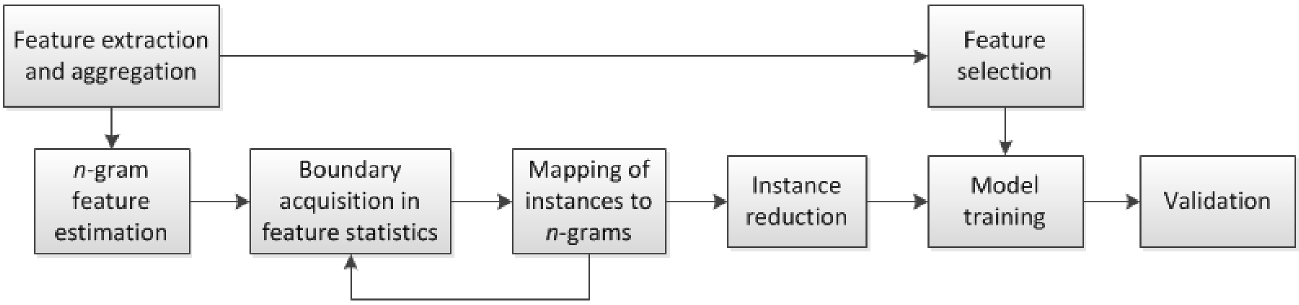
Feature extraction and aggregation
Feature selection
n-gram feature estimation
Boundary acquisition in feature statistics
Mapping of instances to n-grams
Instance reduction
Model training
Validation

Figure 2: Algorithmic steps for training set reduction.

$max(f_i)$ as described in the next subsection. Each feature value for a current instance was then mapped to a binary vector of two values depending on the interval where it belonged to: 00 for $f_i \in [min(f_i); b_1]$, 01 for $f_i \in [b_1; b_2]$, 10 for $f_i \in [b_2; b_3]$ and 11 for $f_i \in [b_3; max(f_i)]$. The combination of three n-gram-building features led to the binary vector of length 6, which was mapped to a symbol. Since the number of possible symbols in this case is equal to $2^6 = 64$, we used 26 low-case letters a-z, 26 upper-case letters A-Z, ten number symbols 0-9 and the symbols plus(+) and point(.). The combination of both three-feature groups led to the 2-gram based on the mentioned symbols.

2.4 Boundary Acquisition in Feature Statistics

Since we wanted to get approximately the same number of frequent and rare n-grams, we provided a straightforward method for the search of the appropriate boundaries b_1, b_2, b_3 based on the optimization of n-gram distribution. The upper and the middle subfigures of Fig. 3 illustrate the distributions of the sorted n-grams for energy and tempo characteristics and MFCCs for the classic training set. For the reasonable balance between the frequent and rare n-grams we calculate the areas of the triangles $\mathcal{A}_i$ spanned between the points $[0; 0]$, $[n_i; I(n_i)]$ and $[n_{i+1}; I(n_{i+1})]$ by the application of Heron's formula, where n_i and n_{i+1} are the consecutive n-grams and $I(n_i), I(n_{i+1})$ the amounts of the corresponding music instances:

$$\mathcal{A}_i = \sqrt{s \cdot (s-a) \cdot (s-b) \cdot (s-c)}, \tag{1}$$

where

$$\begin{aligned} a &= \sqrt{(n_i)^2 + (I(n_i))^2}, \\ b &= \sqrt{(n_{i+1})^2 + (I(n_{i+1}))^2}, \\ c &= \sqrt{(n_{i+1} - n_i)^2 + (I(n_{i+1}) - I(n_i))^2}, \\ s &= (a+b+c)/2. \end{aligned} \tag{2}$$

The area skewness is then calculated by division of the sum of all triangle areas $\sum \mathcal{A}_y$ above the symmetry line $y = x$ by the area sum below this line $\sum \mathcal{A}_x$ (the intersection point marked by circle in the bottom subfigure of Fig. 3 is included in area calculation):

$$\mathcal{A}_s = \sum \mathcal{A}_y / \sum \mathcal{A}_x \tag{3}$$

For the search of the minimum of $|1 - \mathcal{A}_s|$ we run a grid search trying different combinations of b_1, b_2 and b_3 where the interval between $min(f_i)$ and $max(f_i)$ is divided into 11 equal intervals by 10 boundary candidates:

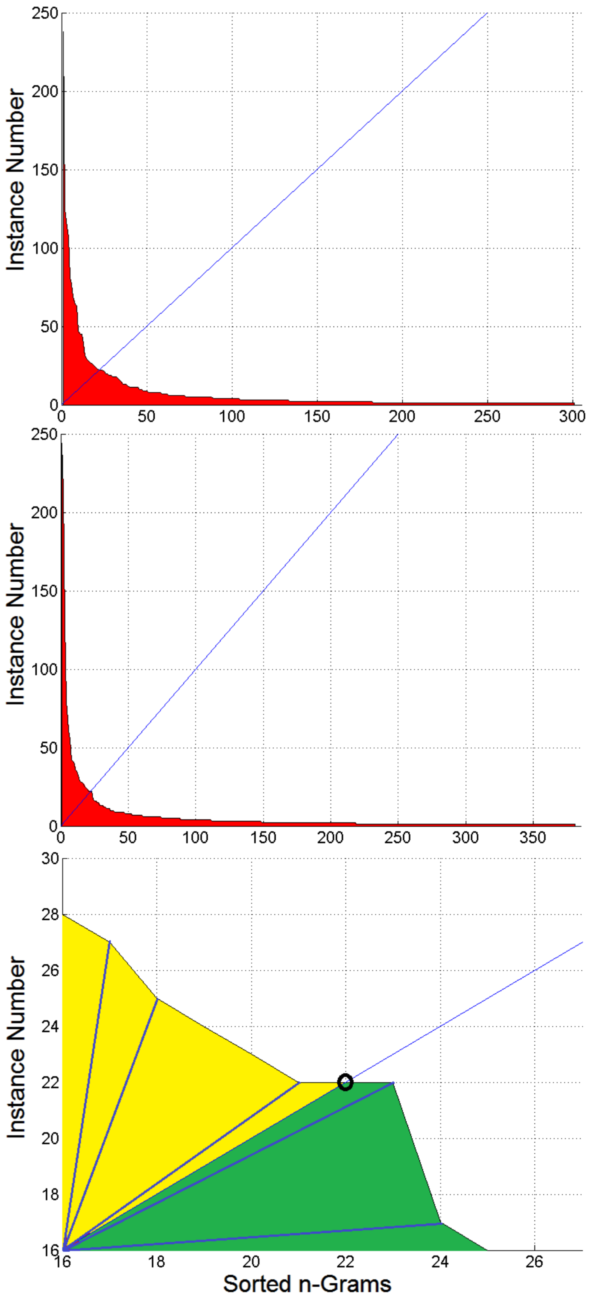

Figure 3: n-grams sorted by the frequencies of the corresponding instances for training set of the classic category. Upper subfigure: Energy and tempo n-gram building features; Middle subfigure: MFCCs; Bottom subfigure: Enlarged region of the middle subfigure as example for triangle area estimation.

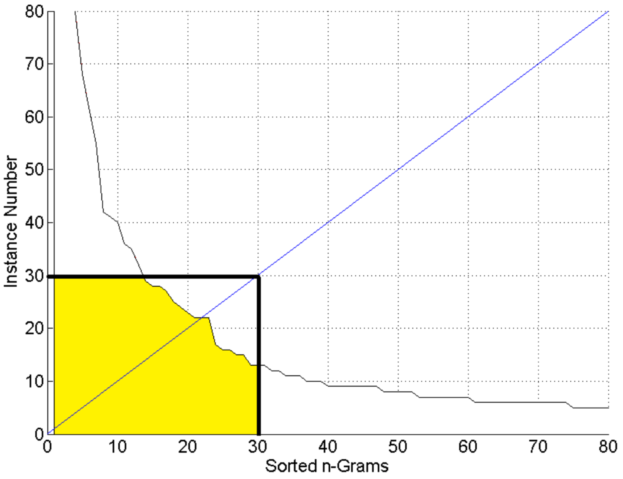

Figure 4: Example for impact of R_x and R_y for training set reduction.

$$\begin{aligned} b_1 &:= min(f_i) + (max(f_i) - min(f_i))/11 \cdot l,\\ b_2 &:= min(f_i) + (max(f_i) - min(f_i))/11 \cdot m,\\ b_3 &:= min(f_i) + (max(f_i) - min(f_i))/11 \cdot n, \end{aligned} \tag{4}$$

where $1 \leq l \leq 10$, $l < m \leq 10$, $m < n \leq 10$ and $l, m, n \in \mathbb{N}$.

2.5 Instance Reduction

After the optimal boundaries are found and the instance mapping to n-grams is created and sorted as depicted in Fig. 3, two reduction steps can be applied keeping in mind both targets discussed at the beginning of this section. The number of the n-grams used for training can be reduced to $R_x \cdot N$ ($R_x \in [0;1]$, N is the overall n-gram number in the current training set), so that only $R_x \cdot 100\%$ of the n-grams sorted by the frequencies of the corresponding features are selected. The appropriate setting of R_x values removes the 'outlier' instances from the training data. On the other side, the setting of $R_y \in [0;1]$ limits the maximal number of instances selected from each n-gram to $R_x \cdot M$ (M is the number of instances in the strongest n-gram), so that the similar instances of the most frequent n-grams are not proceeded too often during the classification model training.

Fig. 4 illustrates this approach: if R_x and R_y are set to the values so that maximal 30 instances from the 30 most frequent n-grams are selected for training, only the instances from the marked shaded area are used for training. In each classification trial we shuffle the instance order in each n-gram: If $R_y \cdot M < I(n_i)$ (which holds for all n-grams $n_2, .., n_N$), $R_y \cdot M$ of $I(n_i)$ instances of the n-gram n_i are randomly selected for training.

2.6 Feature Selection

Since many of the original features may be compeletely irrelevant for a certain classification task, we ran evolutionary feature selection as introduced in [Vatolkin *et al.*, 2011] for each classification task before the training of the models and saved the rather small feature sets for each genre which produced the smallest classification errors. The classification using these category-specific features and all training set instances may be treated as a baseline method. The reduction of the training set based on smaller R_x and R_y may indeed keep or also improve the performance as stated later in the discussion of the study results.

2.7 Model Training

The model training is done as follows. As the ground truth we use 20 labeled songs for recognition of classic, jazz and rap from our music collection of 120 albums with classical, electronic, jazz, pop/rock, r&b and rap music pieces[1]. For each of three genres to identify exactly 10 positive songs (belonging to this genre) and 10 negative songs (belonging to other genres) build the training set. The number of songs is relative small and is motivated - as well as in our previous studies - by the real-word situation where genre prediction systems have to interact with music listeners. If the model training for each music genre or any other personal category requires labeling hundreds of songs by an user, it is very time consuming and exhausting.

However 20 training songs mean here a significantly larger number of training instances - since the labeled feature vectors are created by the aggregation of audio features from audio intervals of 4s with 2s overlap. For the reduction of instance number we then test 400 different shares of the initial instances using all combinations of $R_x, R_y \in \{0.05; 0.1; ...; 0.95; 1.0\}$. This process is repeated ten times. The classification is done by random forest with 100 trees using the category-specific features after the feature selection procedure described in the previous section since this classifier provided a very good compromise between classification error, runtime and generalization ability in our former investigations [Vatolkin *et al.*, 2011; 2012].

2.8 Validation

The validation is done by the estimation of classification error e for the validation set of 120 songs (test set TS120), where exactly one song was drawed randomly from each album and these songs were not contained in any training set.

3 Experiments

By the subsequent decrease of R_x and R_y we reduced the training set instance number involved in model creation and measured the mean classification error e across 10 statistical repetitions. The main focus of this study was to examine how far the reduction of the training data can be run and how much the classification quality suffers (or possibly gains) from this method. Fig. 5 plots the experiment results. In general in can be observed, that the proposed reduction method even led to smaller classification errors comparing to the complete training data set - the major difficulty is however that it is clearly dependent on the category, which R_x and R_y values and also which n-gram feature building group should be recommended.

The n-gram building by random features produces very similar results for all three classification tasks (right column): the selection of only 5% of instances from the 5% of the most frequent n-grams leads to significant decrease of classification quality. However the R_x and R_y values below 0.25 achieve good performances for certain combinations of R_x and R_y and for jazz the solution from $R_x = R_y = 0.1$ even outperforms the training with all instances. No strong difference between the reduction due to smaller R_x or rather R_y values can be stated.

For the different manually selected feature groups the situation is not so clear. One of the most interesting columns

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

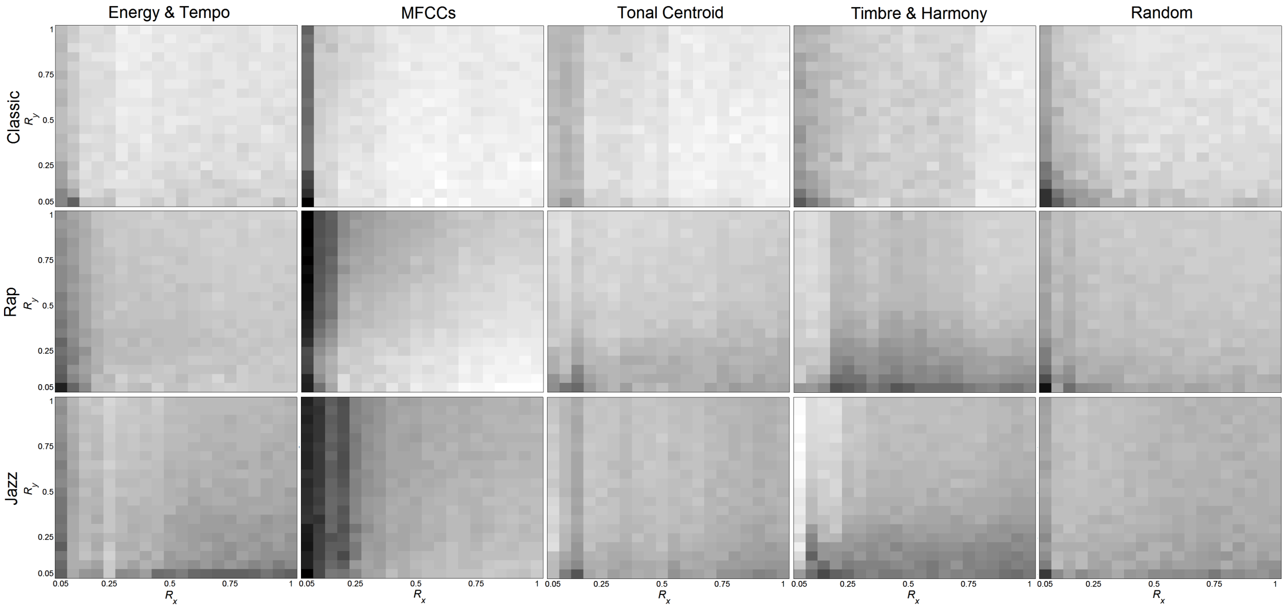

Figure 5: Mean classification errors from 10 statistical repetitions for three genre categories (rows) and for five n-gram building feature sets (columns). The color is scaled between white (smallest error across the different n-gram building features for the same category) and black (largest error).

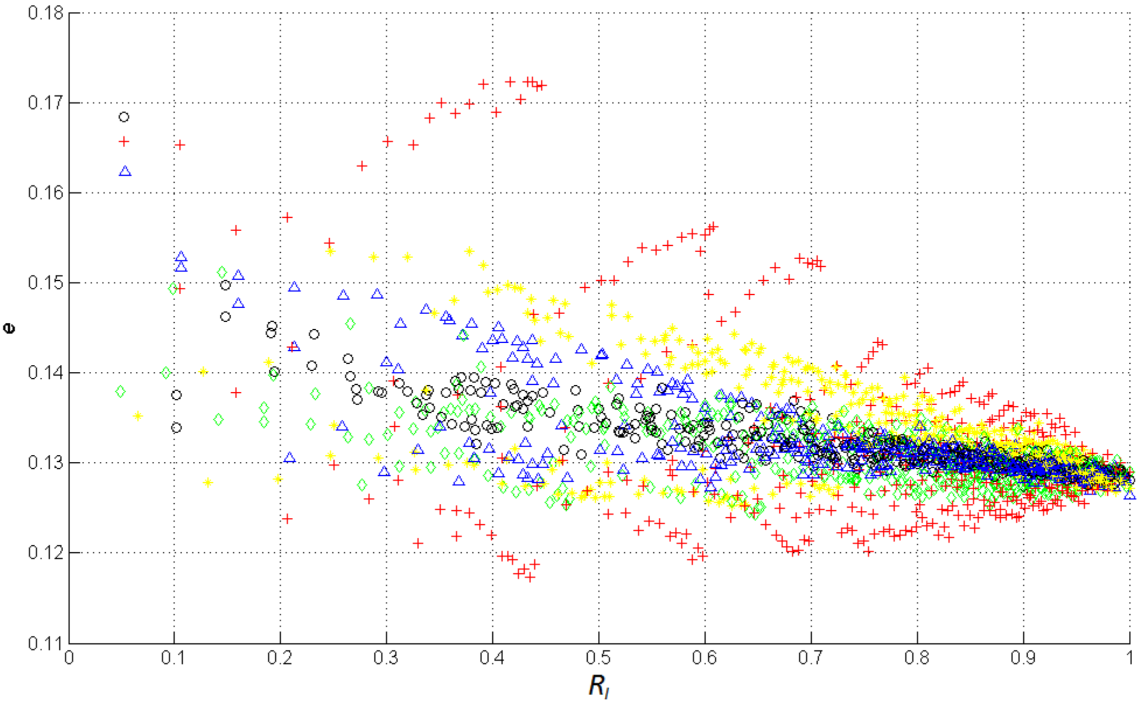

Figure 6: Training set share R_I mapped to median error e for all n-gram building features and (R_x, R_y) combinations for category rap. Triangles: Energy & tempo n-gram building features; Plus signs: MFCCs; Diamonds: Tonal centroid; Asterisks: Timbre & harmony; Circles: Random feature set.

is the 2nd from the left: the organization of n-grams by MFCCs provides very large errors if only the most frequent n-grams are taken for classification model training. This effect increases from the easiest category classic (upper subfigure of this column) to the most complex category jazz (bottom subfigure). For timbre & harmony n-gram building features it seems to hold that the smallest examined R_y value increases the danger of larger errors rather than the small R_x values. For jazz the smallest errors are achieved for $R_x = 0.05$ and $R_y < 0.15$.

The very different behaviour of n-gram building feature sets 1-4 apart from the random feature selection can be explained by the impact of these features on the classification task itself: if the n-gram building features are themselves very relevant for separation of positive and negative instances, the instances of a single n-gram may belong almost completely to one class: e.g. for MFCCs-building features and category classic the first n-gram with 244 instances (cf. Fig. 3, middle subfigure) consists of only 29 classical instances and 215 non-classical labeled feature sets. Therefore the choice of n-gram-building features must be treated very carefully, and it may indeed have some advantages to use not very relevant features for the concrete categorization task. If both positive and negative songs have similar energy and tempo distributions, the first group of energy and tempo features may be a good choice: In that case these n-gram-building features are not relevant for categorization task but may provide justified structuring grouping instances of the approximately equal energy and tempo together.

Since the numbers of instances in different n-grams vary very strongly, the exact fraction of the training set instances saved for learning can not be directly computed from R_x and R_y. Fig. 6 illustrates this fraction R_I for the category rap. Here it can be clearly observed, that the same R_I may correspond to the small and large errors, since R_I is approximately the same if both R_x and R_y values are switched. The largest errors above 0.15 are produced in most cases by MFCCs-n-grams with $R_x = 0.05$ and different R_y values (cf. Fig. 5).

The deviation for error distribution of randomly selected features is very small, and it does not occur, that the switched R_x and R_y may correspond to the very different errors. In other words, for random n-gram-building features there exist no preference of selecting small R_x or rather R_y values. This behaviour can be even seen as an advantage of this method: the classification error change during the reduction of R_x and R_y is more predictable for different categories if random features are drawn for n-gram building.

Also it can be observed, that the classification error can be reduced even if the training set is reduced up to 15% of the original size using timbre & harmony n-gram aggregation and up to 20% for also MFCCs and energy & tempo method.

4 Conclusions and Outlook

In our work we developed a novel training data reduction method using 2-grams for unsupervised pre-structuring of data instances - motivated by the specific characteristics of musical pieces, where some audio segments are repeated often and must not be involved into building of training sets for each repetition. On the other side, segments, which are rather rare and not really significant for a concrete genre, can be omitted for classification model building. As the experimental study showed, the method is indeed reasonable in practice, and the strong reduction of labeled feature sets produces even smaller errors in many cases. Also some general observations can be stated - such as that the selection of random features for n-gram-building does not decrease the expected success by too strong variance. The selection of concrete feature sets led on the other side to several interesting effects, some of them holding for different tested genre categories: MFCCs-n-gram building method did not perform well for very small R_x values, because too many instances of the same class were located in the n-grams with the largest instance number.

However we are aware that a lot of work still remains to provide a robust method sufficient for real-world application: At the current stage we cannot guarantee reasonable performance for any music genre category. Therefore we aim at continuing the development of the n-gram training set reduction method and also to examine further parameter settings - e.g. omitting the feature selection procedure before classification training, application of other classification algorithms or building of words with the length $n > 2$. The comparison with other data reduction algorithms may help to further improve our method. Another promising approach is to estimate the n-gram-building features by heuristics or provide other sets using expert knowledge. The adaptation of both theoretical targets (outlier removal and omitting of too many similar instances) can be improved if the training sets will be more thoroughly constructed, for example including music tracks with many repetitive segments. Then the systematic method optimization can be applied, so that all of these segments are actually mapped to the same n-grams.

Acknowledgments

We thank the Klaus Tschira Foundation for the financial support.

References

[Bishop, 2006] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

[Doraisamy and Rüger, 2003] S. Doraisamy and S. Rüger. Robust polyphonic music retrieval with n-grams. *Journal of Intelligent Information Systems*, 21(1):53–70, 2003.

[Downie, 1999] J. S. Downie. *Evaluating a Simple Approach to Music Information Retrieval: Conceiving Melodic N-Grams as Text*. PhD thesis, The University of Western Ontario, Faculty of Information and Media Studies, 1999.

[Duda *et al.*, 2000] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley-Interscience, New York, 2000.

[Gao *et al.*, 2000] J. Gao, M. Li, and K.-F. Lee. n-gram distribution based language model adaptation. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP/INTERSPEECH)*, pages 497–500. ISCA, 2000.

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

[Harte *et al.*, 2006] C. Harte, M. Sandler, and M. Gasser. Detecting harmonic change in musical audio. In *Proceedings of the 1st ACM workshop on Audio and Music*

Computing Multimedia (AMCMM), pages 21–26. ACM, 2006.

[Lartillot and Toiviainen, 2007] O. Lartillot and P. Toiviainen. Mir in Matlab (ii): A toolbox for musical feature extraction from audio. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, pages 127–130, 2007.

[Last, 2009] M. Last. Improving data mining utility with projective sampling. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 487–496, New York, 2009. ACM Press.

[Mason, 2009] J. E. Mason. *An n-gram Based Approach to the Automatic Classification of Web Pages by Genre*. PhD thesis, Dalhousie University, Faculty of Computer Science, 2009.

[McKay, 2010] C. McKay. *Automatic Music Classification with jMIR*. PhD thesis, McGill University, Department of Music Research, 2010.

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

[Müllensiefen, 2009] D. Müllensiefen. *Statistical techniques in music psychology: An update*, pages 193–215. Peter Lang, Frankfurt, 2009.

[Müller and Ewert, 2011] M. Müller and S. Ewert. Chroma toolbox: Matlab implementations for extracting variants of chroma-based audio features. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, pages 215–220, 2011.

[Rabiner and Juang, 1993] L. Rabiner and B.-H. Juang. *Fundamentals of speech recognition*. Prentice-Hall, Inc., New York, 1993.

[Scholz *et al.*, 2009] R. Scholz, E. Vincent, and F. Bimbot. Robust modeling of musical chord sequences using probabilistic n-grams. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 53–56, Washington, 2009. IEEE.

[Suen, 1979] C. Y. Suen. n-gram statistics for natural language understanding and text processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):164–172, 1979.

[Sumi *et al.*, 2008] K. Sumi, K. Itoyama, K. Yoshii, K. Komatani, T. Ogata, and H. G. Okuno. Automatic chord recognition based on probabilistic integration of chord transition and bass pitch estimation. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR)*, pages 39–44, 2008.

[Vatolkin *et al.*, 2010] I. Vatolkin, W. Theimer, and M. Botteck. Partition based feature processing for improved music classification. In W. Gaul, A. Geyer-Schulz, L. Schmidt-Thieme, and J. Kunze, editors, *Challenges at the Interface of Data Analysis, Computer Science, and Optimization. Proc. of the 34th Annual Conference of the Gesellschaft für Klassifikation e. V.*, pages 411–419. Springer, 2010.

[Vatolkin *et al.*, 2011] 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 Genetic and Evolutionary Computation Conference (GECCO)*, pages 411–418. ACM Press, 2011.

[Vatolkin *et al.*, 2012] 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 - A Fusion of Foundations, Methodologies and Applications*, 2012. to appear in print.

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

[Wołkowicz *et al.*, 2008] J. Wołkowicz, Z. Kulka, and V. Keselj. n-gram-based approach to composer recognition. *Archives of Acoustics*, 33(1):43–55, 2008.