Musical Instruments in Random Forest

Miron Kursa[1], Witold Rudnicki[1], Alicja Wieczorkowska[2],
Elżbieta Kubera[3], and Agnieszka Kubik-Komar[3]

[1] Interdisciplinary Centre for Mathematical and Computational Modelling (ICM), University of Warsaw, Pawinskiego 5A, 02-106 Warsaw, Poland
[2] Polish-Japanese Institute of Information Technology, Koszykowa 86, 02-008 Warsaw, Poland
alicja@pjwstk.edu.pl
[3] University of Life Sciences in Lublin, Akademicka 13, 20-950 Lublin, Poland

Abstract. This paper describes automatic classification of predominant musical instrument in sound mixes, using random forests as classifiers. The description of sound parameterization applied and methodology of random forest classification are given in the paper. Additionally, the significance of sound parameters used as conditional attributes is investigated. The results show that almost all sound attributes are informative, and random forest technique yields much higher classification results than support vector machines, used in previous research on these data.

1 Introduction

In the era of vast and continuously growing audiovisual data available in private repositories and in the Internet, it becomes desirable to be able to automatically browse these data in order to find the specified contents. The user may be interested in finding pieces of music in the desired mood or style, finding tunes, or timbres. This research addresses the problem of automatic identification of timbre, i.e. musical instrument, in audio data. This can be performed using various classifiers, run on digital data. In this paper, we use exemplary feature vector to parameterize sound for timbre classification purposes, and we present random forest technique applied as a classifier. Since the feature vector is arbitrarily chosen, we decided to check importance of these parameters, even though they were already used and reported fairly successful in similar research [23].

1.1 Recognition of Musical Instruments

Musical instrument sound recognition has been investigated last years by various research centers, starting from recognition of isolated sounds, and last years also sounds in mixes. There is no standard set of parameters used, although MPEG-7 features reflect parameterization used in audio research and offer numerous features [11]. The classifiers applied in research on musical instruments include k-nearest neighbors, artificial neural networks, rough set based algorithms, support vector machines (SVM), etc. [5,8,12,16,24]. Parameterization used includes

J. Rauch et al. (Eds.): ISMIS 2009, LNAI 5722, pp. 281–290, 2009.

features describing properties of DFT spectrum, wavelet analysis coefficients, MFCC (Mel-Frequency Cepstral Coefficients), multidimensional scaling analysis trajectories, etc. [10]. The results are difficult to compare because of various numbers of classes used, no. of objects per class, and audio data. Generally, recognition of instruments for isolated sounds can reach 100% for small number of classes, more than 90% if instrument or articulation family is identified, and the accuracy goes down to about 70% or less for recognition of instrument when there are more classes to recognize. Research on identification of instruments in mixes was also performed (same-pitch multi-timbre mixes being the most difficult) [7,9,13,23,26], and the results vary depending on the instruments and sounds chosen. The outcomes can be used for aiding automatic music transcription, but this usually aims at multiple-pitch tracking, and instrument information is only supplementary (for separation of particular voices).

Here we decided to use random forest technique, as it is promising for high-dimensional feature set data, which is often the case in audio signal classification.

Random Forest. Random Forest (RF) is a classifier which comprises of a set of weak, weakly correlated and non-biased classifiers, namely the decision trees. It has been shown that in many cases RF performs equally well or better than other methods on a diverse set of problems [3]. It has been widely used in classification problems as diverse as bioinformatics [2,6,15], medicine [22] or, more recently, material science [4], transportation safety [1], or customer behavior [25].

In addition, RF offers a useful feature that improves our understanding of a classification problem under scrutiny, namely it gives estimate of the importance of attributes for the final prediction. It is often used for analysis when both classifier and identification of important variables are goals of the study [2,15].

2 Material and Methods

Our data originate from MUMS [17], widely used in similar research. They represent instrument sounds, in many cases played with various methods (articulation). We chose 12 sounds - octave no.4 in MIDI – for the following 14 instruments: clarinet; flute; oboe; English horn; trumpet; French horn; tenor trombone; piano; marimba; vibraphone; tubular bells; violin, viola, and cello *vibrato*.

Our goal was to recognize the instrument dominating in same-pitch mix, as this is the most challenging task, since in this case partials in spectra overlap and separation of sounds is more difficult. The training data describe isolated monophonic instrumental sounds, and the same sounds mixed with the second, artificial sound: triangular wave of the same pitch, saw-tooth wave, white noise, or pink noise. To make sure we recognize the predominant sound, the level of added sound was only a percentage of the main sound level, at 7 versions in equal step in log scale: 50%, $50/\sqrt{2}$%, 25%, $25/\sqrt{2}$%, 12.5%, $12.5/\sqrt{2}$%, 6.25%. The test data represent mixes of the predominant instrument with the remaining instrument sounds of the same pitch from our data set, at the same 7 levels. We also performed experiments on combined training set and combined test set [23].

2.1 Construction of Attributes

The parametrization used here was already applied in similar research [23,26] and the results were promising, so we decided to follow this scheme. The feature set consists of 219 parameters, based mainly on MPEG-7 audio descriptors, and other features used in similar research. Most of the parameters in this feature vector represent average value of frame-based attributes, calculated for consecutive frames of an investigated sound (or mix) using sliding analysis window, moved through the entire file. The calculations were performed for the left channel of digital data for 44.1 kHz sampling rate and 16-bit resolution, using 120 ms analyzing frame with Hamming window (hop size 40 ms), which allows analysis of the low-pitched sounds even for the lowest audible fundamental frequencies, if one would like to investigate full music scale. The feature vector consists of the following parameters, describing sound features in time domain, time-frequency domain, and frequency domain (averaged over all frames) [23,26]:

- MPEG-7 based descriptors [11]: *AudioSpectrumSpread*; *AudioSpectrumFlatness* for 25 out of 32 frequency bands; *AudioSpectrumCentroid*; *AudioSpectrumBasis* features: 165 features for 33 sub-spaces - min, max, mean, distance (summation of dissimilarity, i.e. absolute difference of values, of every pair of coordinates in the vector), and standard deviation of *AudioSpectrumBasis*; *HarmonicSpectralCentroid*, *HarmonicSpectralSpread*, *HarmonicSpectralVariation*, *HarmonicSpectralDeviation*, *LogAttackTime*, *TemporalCentroid*.
- other: *Energy*; *MFCC* (min, max, mean, distance, standard deviation); *ZeroCrossingRate*; *RollOff*; *Flux*; *FundamentalFrequency*; $r_1, ..., r_{11}$ - various ratios of harmonic partials: r_1 – energy of the fundamental to the total energy of all harmonics, r_2: amplitude difference [dB] between 1^{st} and 2^{nd} partial, r_3: ratio of the sum of partials 3-4 to all harmonics, r_4: partials 5-7 to all, r_5: partials 8-10 to all, r_6: remaining partials to all, r_7: brightness – gravity center of spectrum, r_8, r_9: contents of even/odd harmonics in spectrum.

2.2 Random Forest Method

RF is an ensemble of classification trees, constructed using procedure which minimizes bias and correlations between individual trees. Each tree is built using different bootstrap sample of the training set. The elements of the sample are drawn with replacement from the original set; roughly 1/3 of the training data are not used in the bootstrap sample for any given tree. For each tree in RF these elements are called out-of-bag (OOB) elements for the tree (they are different for each tree). Let us assume that objects are described by a vector of P attributes. At each stage of tree building, i.e. for each node of any particular tree in RF, p attributes out of all P attributes are randomly selected, where $p \ll P$ (often $p = \sqrt{P}$). The best split on these p attributes is used to split the data in the node. Each tree is grown to the largest extent possible (no pruning).

By repeating this randomized procedure M times one obtains a collection of M trees, hence a random forest. Classification of each object is made by simple

voting of all trees. The number of trees depends on the problem, usually the number of steps is selected to assure that the classification error is not changed after adding more trees. The classification error on the training set is estimated by counting only votes put by trees on their OOB objects.

This estimate of attributes' importance is performed in the following way. For each attribute, one takes all trees, which were using this attribute Then each tree classifies all its OOB objects. One counts number of correct decisions. Then one permutes values of the attribute between all objects and repeats the procedure. The average difference between number of correct classifications in these two cases is a raw classification score. One can also compute the variance and standard deviation and use it to compute Z-score of the raw score.

2.3 Feature Selection

The validity of the estimate of the variable importance is based on the assumption that the individual trees building the random forest are uncorrelated, which, in most cases is fulfilled only approximately [3]. Also, when the number of variables is large, it is difficult to discern truly important variables from these which gain importance due to random correlations in data. It has been shown by simulations [19,20,21] and observation of experimental data [18] that importance measure for most attributes can be highly variable.

To solve this problem, we developed an algorithm comparing the apparent importance of the original variables with that of the randomized ones [14,18]. This algorithm is a wrapper based on the importance score obtained from the RF method. In this method one creates many times an extended system, where each attribute has a mirror copy. Values of the mirror attributes are randomly permuted between objects, and importance of the original attributes compared with these of the mirror ones. Only attributes of importance consistently higher than the highest importance for the randomized attribute are considered important.

This algorithm finds all attributes, which, for given data set, are correlated more strongly with the decision attribute than random attributes. There is no guarantee that all attributes which are truly related with the decision are found or that the attributes which are found are truly, and not by chance, correlated with the decision attribute. Still, our algorithm is an attempt in this direction and it gives reasonable estimate of the importance of the selected attributes.

3 Results and Discussion

Machine learning methods have been already used for identification of musical instruments; in the previous paper [23] we have shown that SVM yielded results ranging from 55.9% (for learning on single instrument sounds and testing on mixes with added instrumental sounds of 50% level of the main sound level) up to 89.88% correctness (for learning on single instrument sounds with added artificial sounds of 6.25% level and testing on mixes for the same level).

Results of classification for RF trained on pure instrument sounds are shown in Figure 1 (left panel). Horizontal axis represents quantity of added sound [%],

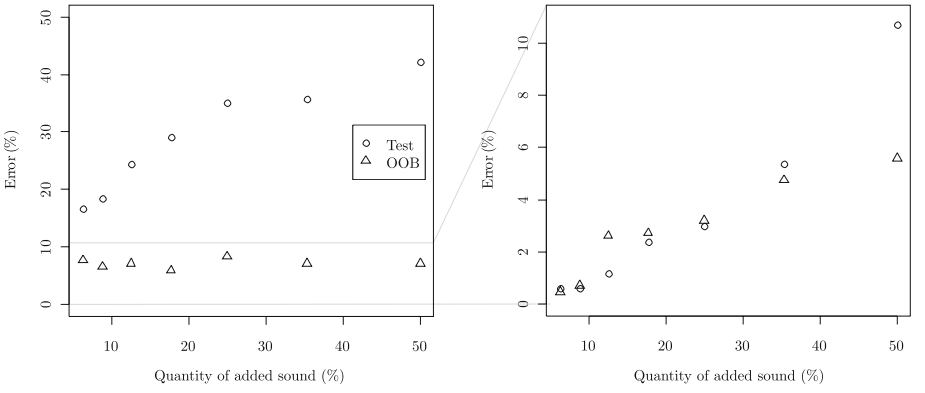

Fig. 1. Classification results using RF for learning on sounds representing only a single instrument (left panel) and learning on sounds containing varying level of additional sounds (right panel); the level is shown as percentage of the main sound level. OOB results are marked with circles, and the results for test sets (mixes of instrument sounds) are marked with triangles. Remark: different scales are used in the left and right panels.

and vertical axis represents recognition error for dominating sound. Learning in the presence of multiple sounds is shown in Figure 1 (right panel). In this case the OOB error and test error of the RF classifier are both small (few percent) and agree very well with each other. As long as the added sound level is small, not much increase of error is observed in tests, so for the low added levels the OOB error estimate agrees well with the test set level. Only in the case where the added sound level is 50% the OOB error estimate and test set estimate diverge. OOB remains relatively small (around 5%), while test set error increases to 10%.

RF OOB error of the classifier for pure sounds shows significant improvement over SVM, but no improvement if this classifier is used to recognize mixes (Fig.2).

For all but one levels of added sound, error for RF classifier is much lower than that of SVM, almost by an order of magnitude. This is still about three times better than in the case of SVM classifier, but this increase of error shows the limits of validity for the applied training and testing procedure. It shows that training on noisy data which is not related to the noise added in the test set works well until the level of added noise is smaller than level of pure sound.

Comparing results of the training performed on the samples of pure sound with those obtained for sound mixes, the OOB error in the latter case is sometimes more than an order of magnitude smaller than in the former one. It is clear that adding noise to the training set dramatically improves results of training.

Following the procedure from earlier work, we have also tested the case when training set consisted of all training samples and similarly, the testing set consisted of all testing samples as well. In this case RF classifier worked very well, the OOB error was marginal (0.05%) and test set error was around 1%. Again, these results were an order of magnitude better than that for SVM classifier.

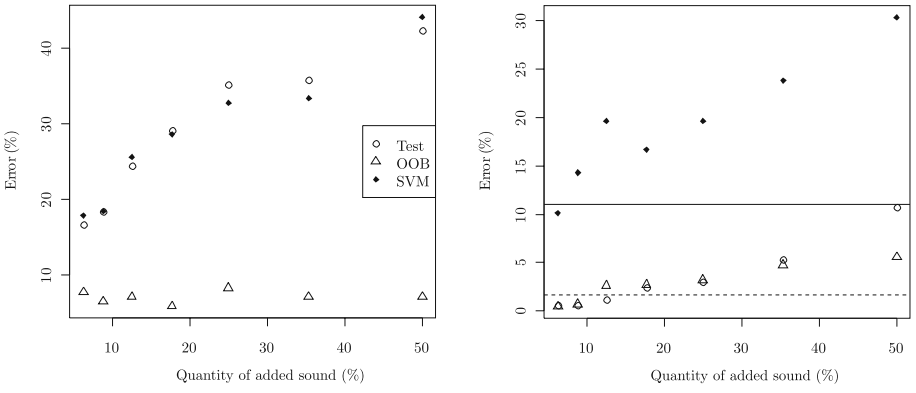

Fig. 2. Classification results using RF and SVM, for learning on single sounds (left) and on sounds containing varying level of additional sounds (right)

Results of the classification show that RF method is very well suited to classification of musical instruments, significantly better than SVM. The difference is of qualitative nature, since in most cases the RF error is one order of magnitude better than SVM. This is mainly due to characteristics of features used for the description of musical instrument sounds: feature values for each instrument are discrete points distributed over wide intervals creating characteristic patterns; intervals corresponding to various instruments overlap significantly (Fig. 3).

As an example all values of two attributes (*TermporalCentroid* and *basis101*) are displayed for the samples taken from clear sound of instruments and for all samples. One can see that adding different sound to the sample of clear instruments does not change the characteristic patterns but moves it slightly, so when one overlays all mixed samples the wider blurred bands are created.

Data of that type fits very well to the tree-based methods, where each leaf can represent separate small interval in 1-dimensional space. On the other hand it poses a challenge for the classifier based on finding continuous intervals. Such method can only succeed using very highly dimensional property space, where one can map all necessary splits on complex multidimensional figures.

Indeed, one can create relatively good RF classifiers using only a single attribute. We have constructed single feature RF classifiers using all features, at 6.25% added sound level. For most features identified as important by our feature selection procedure one can obtain RF classifier with OOB error close to 80%, which is noticeably better than random choice (the reference level of random classification is 92%). If one chooses features which are highly ranked by RF importance ranking algorithm, one can construct significantly better classifiers. For example the RF classifier built using only TermporalCentroid, ranked #1 in the importance ranking, has OOB error 14.5%, whereas the error on the test set is 25%. One can also use less important features and still get classifier which is noticeably better than random choice. For example for *basis5* (ranked #30), the errors of the RF classifier built using this feature are respectively 76.8% and

82.9%. In most cases these classifiers are too weak for any useful classification, but one should remember that in our case the test set consists of samples constructed by mixing different sounds than the original training set, nevertheless the classification error is still noticeably smaller than for random classifier.

Important attributes. Importance of the attributes for prediction was estimated using Boruta Algorithm, aiming at finding all truly informative features. This algorithm compares importance of all features with a reference importance, i.e. maximal apparent importance of the randomly permuted mirror features.

The results yielded by Boruta are different for systems with different levels of added sound. Generally, the number of important attributes grows with the level of added sound; for the combined data set (all levels), all attributes were informative. Still, for the 6.25% level, 146 out of 219 features were informative. There was no clear cut-off value of the average importance between important and non important features, and the importance of a given feature in different iterations

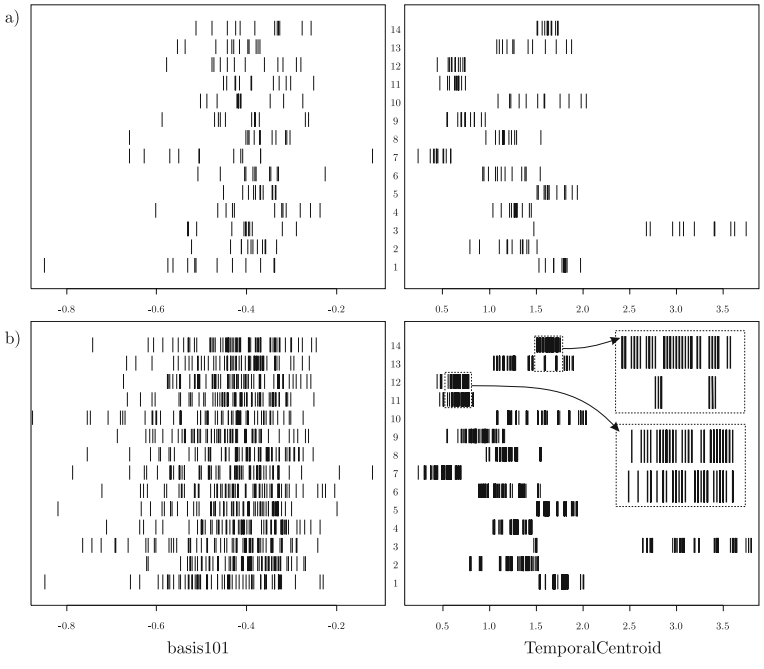

Fig. 3. Attribute values for one of unimportant descriptors (*basis101*) and the most important descriptor (*TermporalCentroid*) for sounds containing only a single instrument a), and for sound containing 50% of added (mixed) sounds b) for all instruments. The instruments are marked on vertical axis as follows: 1.clarinet, 2.cello, 3.trumpet, 4.English horn, 5.flute, 6.French horn, 7.marimba, 8.oboe, 9.piano, 10.trombone, 11.tubular bells, 12.vibraphone, 13.viola, 14.violin. The details for two pairs of instruments which are well discerned (viola and violin) and poorly discerned (bells and vibraphone) by classifier using single attribute are shown in inset in right panel for the case of mixes.

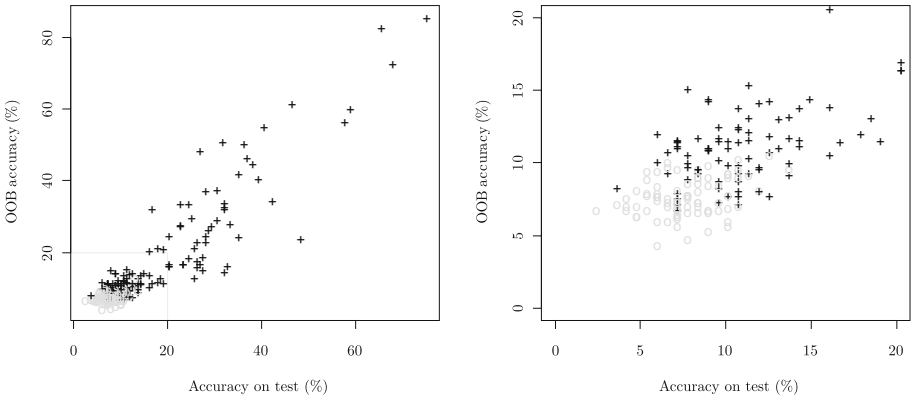

Fig. 4. Importance of attributes vs. the results of the classifiers built using a single attribute. Selection of important attributes was performed for the classifier trained for recognition of musical instrument sounds with mixes (at 6.25% level), using this attribute as the only conditional attribute. Important attributes are marked as "+", non-important – as "o". Position of the attribute on the plot corresponds to the accuracy of the classifier for the test set and OOB accuracy on the training set. The results for all attributes are shown in the left panel, the close-up of the is shown on the right.

of Boruta may vary very significantly; e.g., the maximal value registered for the feature which is unimportant by design was higher than average importance of 118 out of 146 important features. Still, the minimal value registered for the feature confirmed to be important was lower than average importance of 68 out of 73 unimportant features. It means that in any single RF run, an important feature can attain value lower than average value for most unimportant features. Still, the unimportant feature can reach the value that is higher than average importance measure for most of the important features. Therefore the result of the importance measure for a single RF run should not be considered reliable.

It is interesting to compare importance of the attributes with the error achieved by RF built on a single attribute (Fig.4). Feature selection algorithm finds all attributes that are alone sufficient for construction of a good classifier. Also, the algorithm can find in the cloud of weak attributes (of low predictive power) a set of features that together with other features yield a good classifier.

4 Summary and Conclusions

Results for classifiers trained on pure instrumental sounds are quite low both in case of RF and SVM. Adding mixed sounds to the training set significantly improves classification accuracy in both cases, but the improvement is much higher for RF. The classification results show spectacular superiority of RF over SVM, even though SVM is commonly considered to be a very good classifier – RF is an order of magnitude better than SVM in most cases. The advantage of

using RF in comparison with SVM is caused by sparse distribution of attribute values. They cannot be mapped on large continuous intervals – a large number of small intervals must be used for representation of the attributes. This structure fits very well trees, whereas the SVM may construct inconvenient representation. When additional sounds are mixed with the main one, single attribute values split into several distinct values occupying an interval. The intervals pertaining to different instruments often overlap, and discernment becomes more difficult. This increased difficulty is reflected in the size of the trees. The number of nodes depends on the level of added sounds; for single sounds, trees had 39-83 nodes, for the highest level of added sounds: 227-381 nodes, and for combined data 545-1089 nodes. The trees trained on pure sounds are too simple and cannot properly classify mixes, but the trees trained on mixes perform well also for simpler cases.

Our analysis of the importance shows that most of MPEG-7 based features may be used for the classification, and reasonably good results can be obtained with RF, using merely a single descriptive attribute for all trees in a given RF.

Acknowledgements. This project was partially supported by ICM grants 501-64-13-BST1345 and G34-5, and the Research Center of PJIIT, supported by the Polish National Committee for Scientific Research (KBN).

References

1. Abdel-Aty, M., Pande, A., Das, A., Knibbe, W.: Assessing Safety on Dutch Freeways with Data from Infrastructure-Based Intelligent Transportation Systems. Transp. Res. Rec. 2083, 153–161 (2008)
2. Bureau, A., Dupuis, J., Falls, K., Lunetta, K., Hayward, B., Keith, T., Eerdewegh, P.: Identifying SNPs Predictive of Phenotype Using Random Forests. Gen. Epidem. 28 (2005)
3. Breiman, L.: Random Forests. Machine Learning 45, 5–32 (2001), `http://www.stat.berkeley.edu/~breiman/RandomForests/cc_papers.htm`
4. Carr, D.A., Lach-Hab, M., Yang, S.J., Vaisman, I.I., Blaisten-Barojas, E.: Machine learning approach for structure-based zeolite classification. Micropor. Macropor. Mat. 117, 339–349 (2009)
5. Cosi, P., De Poli, G., Lauzzana, G.: Auditory Modelling and Self-Organizing Neural Networks for Timbre Classification. J. New Music Research 23, 71–98 (1994)
6. Diaz-Uriarte, R., Alvarez de Andres, S.: Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7, 3 (2006)
7. Dziubinski, M., Dalka, P., Kostek, B.: Estimation of musical sound separation algorithm effectiveness employing neural networks. J. Intel. Inf. Syst. 24(2-3), 133–157 (2005)
8. Fujinaga, I., McMillan, K.: Realtime recognition of orchestral instruments. In: Proceedings of the International Computer Music Conference, pp. 141–143 (2000)
9. Goto, M.: A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals. ISCA 43(4), 311–329 (2004)
10. Herrera, P., Amatriain, X., Batlle, E., Serra, X.: Towards instrument segmentation for music content description: a critical review of instrument classification techniques. In: International Symposium on Music Information Retrieval ISMIR (2000)

11. ISO: MPEG-7 Overview, http://www.chiariglione.org/mpeg/
12. Kaminskyj, I.: Multi-feature Musical Instrument Classifier. MikroPolyphonie 6 (2000)
13. Klapuri, A.: Signal processing methods for the automatic transcription of music. Ph.D. thesis, Tampere University of Technology, Finland (2004)
14. Kursa, M., Jankowski, A., Rudnicki, W.: Boruta – a system for feature selection. In: Nguyen, H.S., Huynh, V.N. (eds.) SCKT-08 Hanoi Vietnam (PRICAI 2008), pp. 122–133 (2009)
15. Lunetta, K.L., Hayward, L.B., Segal, J., Eerdewegh, P.V.: Screening Large-Scale Association Study Data: Exploiting Interactions Using Random Forests. BMC Genetics 5, 32 (2004)
16. Martin, K.D., Kim, Y.E.: 2pMU9. Musical instrument identification: A pattern-recognition approach. 136 meeting Acoustical Soc. America, Norfolk, VA (1998)
17. Opolko, F., Wapnick, J.: MUMS – McGill University Master Samples. CD's (1987)
18. Rudnicki, W., Kierczak, M., Koronacki, J., Komorowski, J.: A Statistical Method for Determining Importance of Variables in an Information System. In: Greco, S., Hata, Y., Hirano, S., Inuiguchi, M., Miyamoto, S., Nguyen, H.S., Słowiński, R. (eds.) RSCTC 2006. LNCS (LNAI), vol. 4259, pp. 557–566. Springer, Heidelberg (2006)
19. Strobl, C., Boulesteix, A., Zeileis, A., Hothorn, T.: Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8, 25 (2007)
20. Strobl, C., Zeileis, A.: Danger: High Power! – Exploring the Statistical Properties of a Test for Random Forest Variable Importance. Tech. Rep.17. Univ. Munich (2008)
21. Strobl, C., Boulesteix, A.-L., Kneib, T., Augistin, T., Zeileis, A.: Conditional Variable Importance for Random Forests. Tech. Rep. 23. Dept. Stat., Univ. of Munich (2008)
22. Ward, M.M., Pajevic, S., Dreyfuss, J., Malley, J.D.: Short-Term Prediction of Mortality in Patients with Systemic Lupus Erythematosus: Classification of Outcomes Using Random Forests. Arthritis and Rheumatism 55, 74–80 (2006)
23. Wieczorkowska, A., Kubera, E., Kubik-Komar, A.: Analysis of Recognition of a Musical Instrument in Sound Mixes Using Support Vector Machines. In: Nguyen, H.S., Huynh, V.N. (eds.) SCKT 2008 Hanoi, Vietnam (PRICAI 2008), pp. 110–121 (2008)
24. Wieczorkowska, A.: Rough Sets as a Tool for Audio Signal Classification. In: Raś, Z.W., Skowron, A. (eds.) ISMIS 1999. LNCS (LNAI), vol. 1609. Springer, Heidelberg (1999)
25. Xie, Y.Y., Li, X., Ngai, E.W.T., Ying, W.Y.: Customer churn prediction using improved balanced random forests. Expert Syst. Appl. 36, 5445–5449 (2009)
26. Zhang, X.: Cooperative Music Retrieval Based on Automatic Indexing of Music by Instruments and Their Types. Ph.D thesis, Univ. North Carolina, Charlotte (2007)