2012 IEEE International Conference on Systems, Man, and Cybernetics
October 14-17, 2012, COEX, Seoul, Korea

Music Genre Classification using Dynamic Selection of Ensemble of Classifiers

Paulo Ricardo Lisboa de Almeida
Department of Informatics (DeInfo)
State University of Ponta Grossa (UEPG)
Ponta Grossa (PR), Brazil

Alceu de Souza Britto Jr
Post-graduate Program in Informatics (PPGIa)
Pontifical Catholic University of Parana (PUCPR)
Curitiba (PR), Brazil

Eunelson José da Silva Júnior
Department of Informatics (DeInfo)
State University of Ponta Grossa (UEPG)
Ponta Grossa (PR), Brazil

Luis Eduardo Soares de Oliveira
Post-graduate Program in Informatics (PPGInf)
Federal University of Parana (UFPR)
Curitiba (PR), Brazil

Tatiana Montes Celinski
Department of Informatics (DeInfo)
State University of Ponta Grossa (UEPG)
Ponta Grossa (PR), Brazil

Alessandro Lameiras Koerich
Post-graduate Program in Informatics (PPGIa)
Pontifical Catholic University of Parana (PUCPR)
Curitiba (PR), Brazil

***Abstract* — this paper presents a dynamic ensemble selection method for music genre classification which employs two pools of diverse classifiers. The pools of classifiers are created by using different features types extracted from three distinct segments of each music piece. From these initial pools of weak classifiers, ensembles of classifiers are dynamically selected for each test pattern using the *k*-nearest oracles method. The experiments compare the performance of different selection strategies on the Latin Music Database to those related to the use of best single classifier, and to the combination of all classifiers in the pool. It was possible to observe that the most promising selection strategy evaluated allows improving the classification accuracy from 63.71% to 70.31%.**

Keywords - musical genre classification; ensemble selection.

I. INTRODUCTION

Many efforts have been done to provide to the computers similar abilities than the ones inherent to the human being. Among them, a special ability is the understanding of different sounds. In this context, the challenge related to music genre classification has gained the attention of researchers around the world. The literature offers different approaches to deal with music classification. They usually are based on low-level features extracted from the audio signal. Despite the fact that they use different features and classification methods, the available approaches share the same challenge that is to find the edges among the music genres or rhythms.

An important contribution to this matter was done by Tzanetakis and Cook [1]. They have proposed three sets of features based on the timbral texture, rhythm and pitch. Using an experimental protocol based on 1,000 music pieces from 10 different musical genres, they achieved 61% of classification accuracy. Silla et al. [2] reported 65% of classification accuracy using an approach based on multiple feature vectors and an ensemble strategy conducted according to different space and time decomposition schemes. Their experiments were carried out on a novel dataset called Latin Music Database (LMD) [3], which contains more than 3,000 music pieces categorized into 10 genres. This important dataset has allowed the comparison of different approaches for music genre classification using a similar experimental protocol. The LMD was used for evaluation of thirty-three different algorithms in the MIREX 2009 contest [4] and the reported results range from 38.8% to 74.6%. The best result was achieved by using acoustic features (e.g. Mel Frequency Cepstral Coefficients), Gaussian mixture models and Support Vector Machine (SVM) [4].

Even with important contributions in the literature, the music genre classification is still a challenging problem. McKay and Fujinaga [5] pointed out some problematic aspects of genre and refer to some experiments where human beings were not able to classify correctly more than 76% of the music pieces. In spite of the fact that more experimental evidence is needed, these experiments give some insights about the upper bounds on software performance. McKay and Fujinaga also suggest that different approaches should be proposed to achieve further improvements.

As for many tasks, the construction of a perfect classifier is almost impossible. An alternative way to handle such a problem has been to construct ensembles of classifiers. The main assumption that supports this strategy is that different classifiers make errors on different instances [6]. Ensemble of classifiers is a strategy that has been investigated by many researchers on different classification tasks [7-13].

978-1-4673-1714-6/12/$31.00 ©2012 IEEE

Furthermore, another assumption is that the diversity among the classifiers of an ensemble may contribute to improve the general classification accuracy.

Inspired on that, in this paper, we investigate the use of a dynamic ensemble selection method on the music genre classification problem. For such an aim, the first step consists in creating a pool of weak classifiers trained on features extracted from three distinct segments of the music audio signal. From this initial pool, an ensemble of classifiers is dynamically selected for each test instance. The experiments compare different selection strategies on the Latin Music Database [3] which is a benchmark database.

The paper is organized as follows. Section II presents the proposed method for creating two pools of weak classifiers and the dynamic ensemble selection used to evaluate these pools. Section III presents the experimental results. In the last section, the conclusions are stated as well as perspective for future work.

II. PROPOSED METHOD

This section presents the feature extraction and the time decomposition strategy that is employed to extract features from different segments of the music signal. Next, the construction of the base classifiers that makes up the pool of classifiers is described. Finally, the dynamic ensemble selection used in our method for music genre classification is presented.

A. Feature Extraction

Nowadays the music signal representation is no longer analogous to the original sound wave. The analogical signal is sampled, several times per second, and transformed by an analog-to-digital converter into a sequence of numeric values in a convenient scale. This sequence represents the digital audio signal of the music, and can be employed to reproduce the music.

A digital music signal may be represented by a sequence $S = <s_1, s_2, ..., s_N>$, where s_i stands for signal sampled at time i, while N is the total number of samples of the music. This sequence contains a lot of acoustic information, and different types of features can be extracted from it. Initially the acoustic features are extracted from short frames of the audio signal; then they are aggregated into more abstract segment-level features [2]. Therefore, a d-dimensional feature vector $X = < x_1, x_2, ..., x_d >$ can be generated, where each feature x_i is extracted from S (or from some part of it) by an appropriate extraction procedure.

It has been shown in previous works [2][15] that we can obtain a more adequate representation of a music piece if we consider several time segments of the signal. This procedure, named time decomposition, allows treating the great variation that usually occurs along music pieces. The time decomposition can be formalized as follows. From the original music signal $S = <s_1, s_2, ..., s_N>$ we obtain different sub-signals S_{pq}. Each sub-signal is simply a projection of S on the interval $[p,q]$ of samples, or $S_{pq} = < s_p,...,s_q >$. In the generic case that use M sub-signals, we obtain a feature vector from each sub-signal. In our case we employ 30-second sub-signals taken from the beginning, middle and end parts of the original music signal. Such a time decomposition scheme, experimentally defined in [2], is illustrated in Figure 1.

We note that the feature vectors for the sub-signals are computed over intervals delimited by a window of size w. An overlap between adjacent windows may occur, depending on the strategy used to move the window over S. It depends on the number of signal samples (h) that must be skipped to compute the next window.

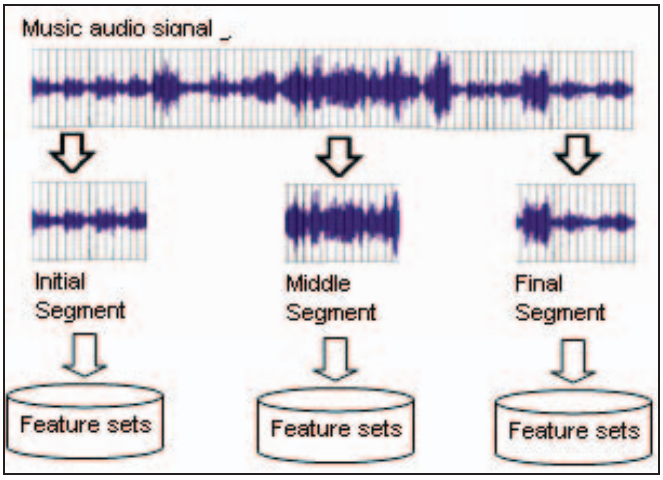

Figure 1. Feature extraction from 3 segments of the music signal adapted from [2].

TABLE I. FEATURE SETS IN THE FIRST GROUP

Feature Sets	Number of Features
Timbral	14
Spectral	12
Mel Frequency Cepstral Coefficients	18
Chroma	13
Spectral Centroid	1
Rolloff	24
Spectral Flux	24
Zero Crossings	1
Spectral Flatness Measure	1
Spectral Crest Factor	1
Line Spectral Pair	3
Linear Prediction Cepstral Coefficients	31
Stereo Panning Spectrum	31
Total	109

The features that are extracted from the digital audio signal are related to the different musical aspects such as melody, harmony, rhythms, timbres and special positions. For such an aim, the Marsyas framework [1] is employed. It includes several individual feature sets proposed in the music information retrieval and audio analysis literature as well as some common combinations of them. The feature sets can be separated into three large groups: time-domain, spectral-domain, and LPC-based. However, in this paper the feature sets were organized into two groups, one group considering all the feature sets available in the Marsyas framework, and another

with the most promising features cited in the MIR (Music Information Retrieval) literature. The first group is composed by thirteen feature sets while in the second group we have only six feature sets. Table I shows the feature sets of the first group and the corresponding dimension of each vector. Table II shows the same aspects for the six feature sets of the second group. When the features sets are concatenated to form a single feature vector, the resulting feature vectors sum up 109 and 61 dimensions for the first and second group, respectively.

Besides the differences among the feature sets, different values for the window size (*w*) and skips (*h*) are also employed to further impose the diversity between the classifiers.

TABLE II. FEATURE SETS IN THE SECOND GROUP

Feature Sets	Number of Features
Chroma	14
Linear Prediction Coefficients	12
Line Spectral Pair	18
Mel-Frequency Cepstral Coefficients	13
Spectral Features	4
Timbral Features	17
Total	61

B. Pool of Classifiers

The base classifier is a one nearest neighbor classifier which is based on the instance based learning paradigm. Such a classifier is considered as a weak classifier and its use is justified by the assumption that it can be helpful to show the improvement by combining classifiers in dynamically selected ensembles.

The strategy used to create the classifiers to compose the initial pools takes into account diversity. The two pools of classifiers were obtained using both groups of features extracted from three 30-second distinct segments taken at beginning, middle and end of the music signal (See Figure 1).

In addition, to further increase diversity we have also varied the windowing scheme used to scan the music signal for feature extraction. Three different window lengths (*w*) were combined with three hop (*h*) values to allow different overlapping between adjacent windows. This allows us to obtain a high number of classifiers to take part of the two pools. The first pool contains 351 classifiers (13 features sets x 3 segments x 3 window size x 3 hop size) while the second one contains 162 classifiers (6 feature sets x 3 segments x 3 window size x 3 hop size).

C. Dynamic Ensemble Selection

The K-Nearest ORAcles (KNORA) method is the strategy used for ensemble selection proposed in [6], which is inspired on dynamic classifier selection methods based on local accuracy [13-14]. In the KNORA method, the idea is simple and corresponds to first find the k-neighborhood of the test instance in a special validation set. The particularity of such a validation set is that for each instance, we know the classifiers from the pool which classify correctly them. The validation set is composed by feature vectors containing all features as described in Tables I and II. Then, in a second step, different strategies may be used to select the classifiers attached to the neighborhood of the test instance in order to compose an ensemble.

Two different selection schemes were considered in this paper: Knora-Eliminate (KE) and Knora-Union (KU). The KE selects only the classifiers that correctly recognize every neighbor of the test instance to compose the ensemble. In case that there is no classifier able to recognize all *k* neighbors, then the value of *k* is decremented and the process is repeated. On the other hand, the KU selects all classifiers that correctly recognize at least one neighbor of the test instance to compose the ensemble. Thus, the same classifier may be selected more than once. In both KE and KU selection schemes, the classifiers dynamically selected are combined using the majority voting rule (MVR). Figures 2 and 3 illustrate the KE and KU strategies, respectively. In both figures, the left side shows the test pattern *X* inside of a hexagon and the corresponding *k* neighbors as darkened circles, while the right side shows the representation of the selected classifiers.

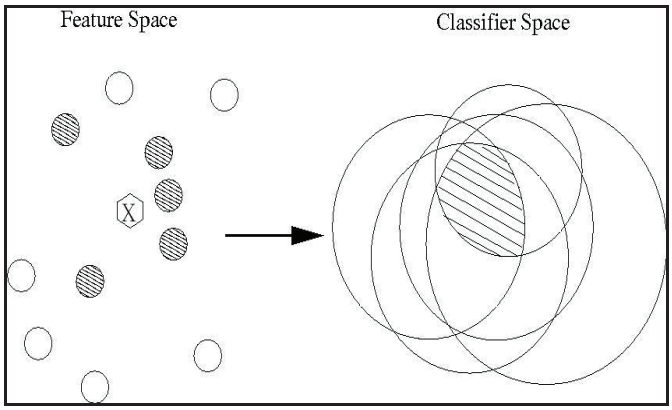

Figure 2. KNORA-Eliminate (KE) selects only the classifers that correctly recognize all nearest neighbors of the test pattern X (figure adapted from [6]).

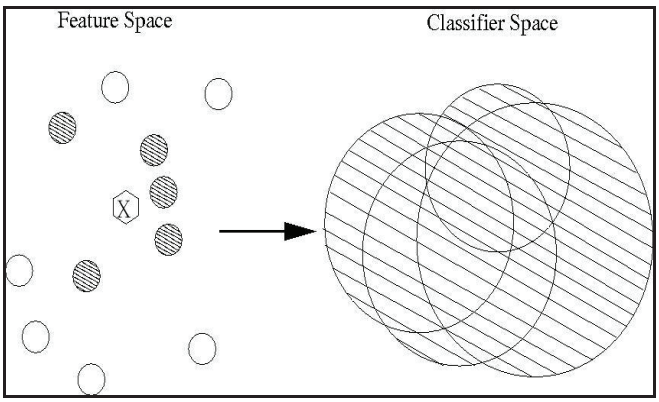

Figure 3. KNORA-Union (KU) selects every classifer that correctly recognize at least one nearest neighbor of the test pattern X (figure adapted from [6]).

As one may see, Figure 2 shows an intersection of the classifiers that correctly recognize the *k*-neighborhood of the test pattern, while Figure 3 shows a union of the classifiers that

Identify applicable sponsor/s here. *(sponsors)*

correctly classify at least one nearest neighbor of the test pattern. Further details of the KNORA method can be found in [6].

III. Experimental Results

This section describes the two experiments undertaken to evaluate the proposed method for music genre classification. Experiment 1 (E1) and Experiment 2 (E2) consider the first and the second group of feature sets, respectively. For both E1 and E2, three different window lengths (w = 256, 512 and 768), defined based on w length (512) reported on [2], were combined with three skip (hop) values to allow different overlapping between adjacent windows (h = 512, 768 and 1024).

A. Database

The experiments were carried out with the Latin Music Database (LMD) [3]. This database contains 3,184 music instances in MP3 file format, distributed on 10 different genres (0-Axé, 1-Bachata, 2-Bolero, 3-Forró, 4-Gaúcha, 5-Merengue, 6-Pagode, 7-Salsa, 8-Sertaneja and 9-Tango). The entire dataset was split into three subsets, named B1, B2 and B3 with 1,064, 1,059 and 1,061 instances, respectively. Each music instance available in the LMD appears in only one subset. In addition, there is no share of artists among the subsets, except for the genre Tango, where there is not enough artists to construct the three different subsets. B1 is used for training, B2 for validation and B3 for testing.

B. Pool Evaluation

Table III summarizes the evaluation of both pools of classifiers: a) the pool created on E1 (351 classifiers); and b) the pool created on E2 (162 classifiers). As one may see the combination of all classifiers using majority voting rule (MVR) provides an interesting improvement in the classification accuracies.

TABLE III. For each music segment - the number of classifiers and the combination result using MVR

Segment	# of classifiers	Combination using MVR (%)
Experiment 1 (E1)		
Initial	117	57.86
Middle	117	56.55
Final	117	52.59
All	351	63.71
Experiment 2 (E2)		
Initial	54	52.59
Middle	54	55.89
Final	54	52.30
All	162	60.50

The best classifiers in E1 and E2, both trained on Chroma features, reached 54.09% and 47.49% of classification accuracy, respectively (see Table IV). However, their performances were surpassed by the corresponding combination of all classifiers, which achieved 63.71% and 60.50%, respectively. Furthermore, the estimated oracle for all segments for both pools is 100%.

TABLE IV. The best classifier at each pool

Experiment	Signal segment	Feature group	Window scheme	Accuracy (%)
E1	Initial	chroma	768	54.09
E2	Middle	chroma	512	47.49

C. Dynamic Ensemble Evaluation

The experiments using the KNORA method considers the k value varying from 1 to 20. The k value is used to determine the size of the neighborhood used to select the classifiers from the validation set (subset B2). Table V presents the best results achieved by the KNORA-Eliminate (KE) and KNORA-Union (KU) and the corresponding k value for both experiments.

TABLE V. Best results of the dynamic selection method and the corresponding K value

Selection scheme	# of classifiers selected	# of of votes	Accuracy (%)
Experiment 1 (E1) Oracle = 100%			
KE (k = 1)	72	72	59.66
KU (k = 10)	249	709	70.31
Experiment 2 (E2) Oracle = 100%			
KE (k = 1)	43	43	57.02
KU (k = 13)	143	573	64.94

As one may see, the pool of classifiers of both E1 and E2 are really diverse. The oracle of both pools is 100%. It means that if the selection strategy succeeds to find the right classifiers to compose the ensemble, then the method may reach 100% of classification accuracy. However, the best result achieved is 70.31% by the KU strategy with k = 10. It surpasses the best classifier in the E1 pool and also the performance obtained by combining all classifiers using the majority voting rule.

Even with a pool that shows the same oracle performance than E1, the E2 achieved just 64.94% of classification accuracy. It shows that the pool of classifiers in E1 is really more diverse than the one in E2. The reason is the larger number of classifiers. In fact, most of these additional classifiers in E1 are really weak in terms of accuracy, however, when combined they improve the classification accuracy.

Tables VI and VII show the confusion matrices related to the experiment E1. Table VI contains the results of the combination of the 351 classifiers in the pool, while Table VII shows the results when using the dynamic ensemble selection based on KU approach. As one may see some classes present a significant improvement, such as: 0 (Axé) and 8 (Sertaneja), while only the class 7 (Salsa) show a negative impact.

TABLE VI. CONFUSION MATRIX RELATED TO THE COMBINATION OF ALL CLASSIFIERS (EXPERIMENT E1)

	0	1	2	3	4	5	6	7	8	9	(%)
0	83	0	4	0	2	6	1	2	6	0	79.8
1	0	101	1	0	0	0	0	1	0	0	98.0
2	3	6	68	2	7	0	3	3	1	11	65.3
3	24	1	11	30	10	3	4	10	9	2	28.8
4	3	7	4	8	41	23	1	7	6	4	39.4
5	2	3	0	0	2	91	0	6	0	0	87.5
6	11	1	18	6	3	4	22	16	14	2	22.6
7	4	1	11	3	4	5	0	70	3	2	67.9
8	13	1	29	1	2	1	8	9	36	4	34.6
9	0	0	0	0	0	0	0	0	0	134	100

TABLE VII. CONFUSION MATRIX RELATED TO THE BEST RESULTS OBTAINED USING THE KU STRATEGY (EXPERIMENT E1)

	0	1	2	3	4	5	6	7	8	9	(%)
0	97	0	2	0	1	4	0	0	0	0	93.2
1	0	101	1	1	0	0	0	0	0	0	98.0
2	2	8	80	0	2	0	2	1	1	8	76.9
3	10	2	13	41	14	3	2	6	11	2	39.4
4	1	4	6	10	42	28	1	6	3	3	40.3
5	1	2	0	1	3	91	0	6	0	0	87.5
6	2	0	8	6	4	5	27	12	31	2	27.8
7	0	0	5	4	3	7	0	61	18	5	59.2
8	1	1	13	1	0	0	5	8	72	3	69.2
9	0	0	0	0	0	0	0	0	0	134	100

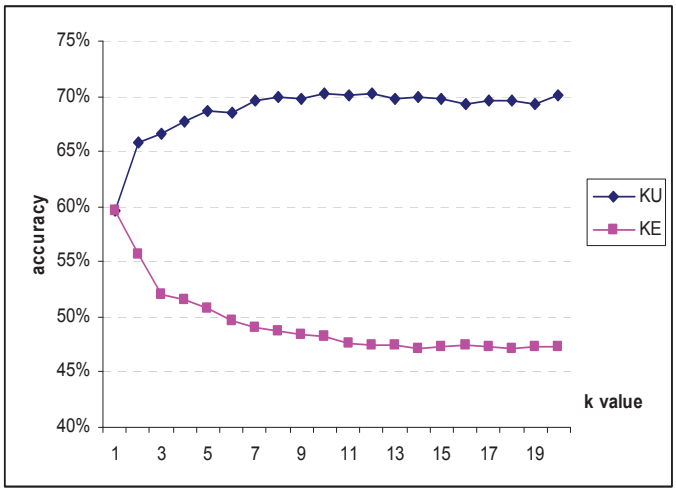

Figure 4. KE and KU accuracies for k from 1 to 20

Figure 4 shows the accuracies of the KE and KU strategies for *k* values from 1 to 20. As one can see the KU accuracy is better when *k* value increases. In other hand, the KE accuracy becomes worst when *k* value increases. This may be explained since KU considers more classifiers as the *k* value is increased, while the KE becomes more selective since the classifiers to be selected must to well classify each one of the *k* neighbors.

IV. CONCLUSIONS

In this paper we presented a dynamic ensemble selection method for music genre classification which employs two pools of diverse classifiers. The diversity is imposed by training the classifiers with different feature sets.

The experiments have shown that the oracle performance of both pools of classifiers generated with the proposed two groups of feature sets is 100% of classification accuracy. This means that even generating weak classifiers (best single accuracy was 54.09%), the strategy used to create them has assured diversity. Thus, the created classifiers show some kind of complementarity. This evidence is very important for applications where the number of samples is not enough for training a single and robust classifier.

We have also observed that the KU scheme to build a dynamic ensemble (70.31%) surpasses the performance of the best single classifier in the initial pool (54.09%), as well as the performance of the combination of all classifiers using the majority voting rule (63.71%). In addition, the KU strategy was also better than KE in both experiments. These results corroborate the hypothesis that investigating different strategies for ensemble selection from pools of diverse classifiers is really promising.

Further work may be done to evaluate pools composed of SVMs (Support Vector Machines) and to test other strategies to select the classifiers from the pools.

ACKNOWLEDGMENT

The authors thank CNPq - Brazil (grants 307567/2011-7 and 472238-2011-6) and CAPES (grant 595) for financial support.

REFERENCES

[1] G. Tzanetakis and P. Cook. Musical Genre Classification of Audio Signals. IEEE Transactions on speech and Audio Processing. 2002.

[2] C. N. Silla; A. L. Koerich; and C.A. Kastnert. A Machine Learning Approach to Automatic Music Classification. Journal of the Brazilian Computer Society, v. 14, p. 7-18, 2008.

[3] C. N. Silla Jr., A. L. Koerich, and C. A. A. Kaestner. The latin music database. In International Conference on Music Information Retrieval, pages 451–456, 2008.

[4] J. S. Downie. MIREX 2009: evaluate your audio DSP algorithms on musical material. In MIREX abstracts, International Conference on Music Information Retrieval, Japan, 2009.

[5] C. McKay and I. Fujinaga. Musical genre classification: Is it worth pursuing and how can it be improved?" in Proc. of the 7th Int. Conf. on Music Information Retrieval, 2006.

[6] H. R. Ko, R. Sabourin., A. S. Britto Jr. From dynamic classifier selection to dynamic ensemble selection. *Pattern Recognition*, 41(5):1718–1731, 2008.

[7] E. M. D. Santos, R. Sabourin and P. Maupin. A dynamic overproduce-and-choose strategy for the selection of classifier ensembles. *Pattern Recognition*, 41(10):2993–3009, 2008.

[8] L. I. Kuncheva and J. J. Rodrígues. Classifier ensembles with a random linear oracle, IEEE Transactions on Knowledge and Data Engineering, Vol. 19, No. 4, pp.500–508, 2007.

[9] J. Xiao and C. He. Dynamic classifier ensemble selection based on GMDH, International Joint Conference on Computational Sciences and Optimization, Vol. 1, pp.731–734, 2009.

[10] G.I. Webb and Z. Zheng. Multistrategy ensemble learning: reducing error by combining ensemble learning techniques', IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 8, pp.980–991, 2004.

[11] A. Santana; R.G.F. Soares; A.M.P. Canuto; and M.C.P. Souto. A dynamic classifier selection method to build ensembles using accuracy and diversity, in Proceedings of the Ninth Brazilian Symposium on Neural Networks (SBRN'06), pp.36–41, 2006.

[12] D. Opitz and R. Maclin. Popular ensemble methods: an empirical study, Journal of Artificial Intelligence Research, Vol. 11, pp.169–198, 1999.

[13] W. P. Kegelmeyer Jr. and K. Bawyer. Combination of Multiple Classifiers using Local Accuracy Estimates. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 19, No. 4, pp. 405-410, 1997.

[14] L. Didaci, G. Giacinto, F. Roli, and G. L. Marcialis. A study on the performances of dynamic classifier selection based on local accuracy estimation. Pattern Recognition 38(11): 2188-2191, 2005.

[15] T. Lidy, C.N. Silla Jr., O. Cornelis, F. Gouyon, A. Rauber, C.A.A. Kaestner and A.L. Koerich. On the Suitability of State-of-the-Art Music Information Retrieval Methods for Analyzing, Categorizing, Structuring and Accessing Non-Western and Ethnic Music Collections. Signal Processing, Vol.90, No.4, pp.1032-1048, 2010.