IMPROVING THE RELIABILITY OF MUSIC GENRE CLASSIFICATION USING REJECTION AND VERIFICATION

Alessandro L. Koerich
Pontifical Catholic University of Paraná (PUCPR)
Federal University of Paraná (UFPR)
alekoe@computer.org

ABSTRACT

This paper presents a novel approach for post-processing the music genre hypotheses generated by a baseline classifier. Given a music piece, the baseline classifier produces a ranked list of the N best hypotheses consisting of music genre labels and recognition scores. A rejection strategy is then applied to either reject or accept the output of the baseline classifier. Some of the rejected instances are handled by a verification stage which extracts visual features from the spectrogram of the music signal and employs binary support vector machine classifiers to disambiguate between confusing classes. The rejection and verification approach has improved the reliability in classifying music genres. Our approach is described in detail and the experimental results on a benchmark dataset are presented.

1. INTRODUCTION

For over ten years the problem of classifying music genres has been the subject of intensive research and is the most widely studied area in Music Information Retrieval [2, 13]. Automatically classifying music by genre is a challenging problem considering that music is an evolving art and there are not clear edges between music genres. A variety of features and classification approaches have been proposed in the last years [1, 2, 6, 13]. Relatively high classification accuracies have been reported in recent papers that carry out experiments on benchmark datasets such as ISMIR 2004, GTZAN [15], and LMD [12]. Sturm [13] provides a comprehensive review of the approaches used for evaluating music genre classification. Sturm shows that over 92% of the papers approach evaluation of music genre classification systems by classifying several music excerpts and comparing the labels to a ground truth. Sturm states that the classification accuracy might not be a correct measure since it can not address the problem at all.

A few survey articles provide an overview of features and techniques used for music genre classification and related tasks [2, 9, 13]. Scaringella et al. [9] reviewed the techniques of audio feature extraction and classification for the task of genre classification only. Fu et al [2] provide a comprehensive review on audio-based classification and systematically summarize the state-of-the-art techniques for music classification, stressing the difference in the features and the types of classifiers used for different classification tasks such as music genre classification, mood classification, artist identification, instrument recognition and music annotation. The review paper of Sturm [13] focused on how music genre recognition systems are evaluated. Moreover, the field of music classification research is developing rapidly in the past few years, with new features and types of classifiers being developed and used. However, few works have evaluated the reliability of the current systems or make a deep analysis of the errors [14]. The definition of reliability used in this paper is borrowed from [5] which was referred to the proportion of correct answers among the accepted instances as a function of the rejection rate.

In this paper we propose the use of rejection and verification steps in an attempt to overcome the deficiencies of conventional classification approaches. The rejection focuses on not classifying instances that generate high uncertainty at the output of the classifier, while verification focuses on the confusions that may happen between particular music genres which is revealed by analyzing the confusion matrices. First, a baseline classifier, which takes into account all the possible music genres is used to classify the input music signal represented by a 68-dimensional feature vector. If this baseline classification scheme does not provide an output with high confidence, which is given by the *a posteriori* probability estimated to each possible music genre, the instance is rejected or may be post-processed at a verification stage, which extracts different features from the music signal and employs a set of binary SVM classifiers. We show in this paper that this approach is able to increase the reliability in music genre classification. It is important to notice that we tackle the problem from a pattern recognition perspective where music genres are treated as labels. The musicological aspect of the music genre classification task is not analyzed in this paper.

This paper is organized as follows. Section 2 presents a baseline music genre classification system. Section 3 introduces the concept of rejection in music genre classification. Section 4 introduces the concept of post-processing and verification of the output of a baseline classifier. Section 5 presents the experimental results of the proposed approach for rejection and verification on a benchmark dataset. Finally in the last section some conclusions are stated.

2. BASELINE MUSIC GENRE CLASSIFICATION SYSTEM

The baseline music genre classification system is composed of two main modules: feature extraction and classification as show in Figure 1. The feature extraction module uses the MARSYAS framework to extract seventeen audio features from $46ms$ frames of the audio signal with no overlap. The features are: zero crossing rate, the spectral centroid, roll-off frequency, spectral flux, and 13 Mel frequency cepstral coefficients, including MFCC0. Frame features are collapsed in a two-step process (texture windowing and computation of global mean and standard deviation) into a 68-dimensional feature vector for the whole audio excerpt [8]. For classification, two implementations were considered: A first version of the baseline system employs a multi-class support vector machine (SVM) algorithm with one-against all strategy. A second version of the baseline system employs the same features, but uses a multilayer perceptron neural network (MLP) trained with the backpropagation momentum algorithm. Such classification algorithms were chosen because both of them can provide estimations of the *a posteriori* probability of each class at the output [5, 16].

Let's consider a n-dimensional pattern recognition problem with c classes. The baseline classifiers perform a task of classifying an input music signal, represented by a n-dimensional feature vector $\overline{x} = [x_1, \ldots, x_n] \in \Re^n$ provided at the input and produce at the output *a posteriori* probability for each of the c possible classes, denoted as $P(\omega_1|\overline{x}), \ldots, P(\omega_c|\overline{x})$, where $\omega_1, \ldots, \omega_c$ denotes the c possible classes. Therefore, we can consider that the baseline classifiers produce at their output a c-dimensional vector $[P(\omega_1|\overline{x}), \ldots, P(\omega_c|\overline{x})]^T$ where $P(\omega_j|\overline{x})$ represents the support for the hypothesis that vector $\overline{x}$ submitted for classification comes from class ω_j. Furthermore, we can assume that such an output vector is ordered in a decreasing order according to the probability, from the most probable class, denoted as TOP 1 to the least probable class. The larger the probability, the more likely the class label ω_j. In real-life applications, the classification system has to come up with a single music genre hypothesis at the output or a rejection of the input if it is not certain enough about the hypothesis. Most of the current works on music genre classification does not consider the rejection option and force a decision using the MAX operator to select the music genre hypothesis which has the highest *a posteriori* probability as shown in Figure 1, i.e. the TOP 1 hypothesis, denoted as ω' and computed by Equation 1.

$$\omega' = \arg \max_{\omega_1 \leq \omega_i \leq \omega_c} P(\omega_i|\overline{x}) \qquad (1)$$

The main drawback of using the MAX operator as decision rule is that it is very severe and it often overlooks the uncertainty that may be present at the output of the classifiers. Such an operator will return a class regardless if the highest probability is close to 1.0 or lower than 0.5. The probabilities estimated by the classifier can be associated to how confident it is in assigning a given class to a input vector.

3. REJECTION OF MUSIC GENRES

How can we handle the uncertainty at the output of a multiclass classifier? Maybe the simplest way is to employ rejection [3–5]. The concept of rejection admits the potential refusal of a music genre hypothesis if the classifier is not certain enough about the genre hypothesis. In our case, the probabilities assigned to each output of the baseline classifiers give evidence about the certainty. The refusal of a music genre hypothesis may be due to two different reasons: there is not enough evidence to come to a unique decision since more than one music genre hypothesis among the c possible music genres appear adequate; there is not enough evidence to come to a decision since no music genre among the c genres appears adequate. In the first case, it may happen that the confidence scores do not indicate a unique decision in the sense that there is not just one confidence score exhibiting a value close to one. In the second case, it may happen that there is no confidence score exhibiting a value close to one. Therefore, the confidence scores assigned to the music genre hypotheses in the N best hypothesis list should be used as a guide to establish a rejection criterion.

Bayes decision rule already embodies a rejection rule, namely, find the maximum of $P(\omega_i|\overline{x})$ but check whether the maximum found exceeds a certain threshold value or not. Due to decision-theoretic concepts this reject rule is optimum for the case of insufficient evidence if the closed-world assumption holds and if the *a posteriori* probabilities are known [10]. This suggests the rejection of a music genre hypothesis if the confidence score for that hypothesis is below a threshold λ. In the context of our baseline classifiers, the task of the rejection mechanism is to, based on output vector $[P(\omega_1|\overline{x}), \ldots, P(\omega_c|\overline{x})]^T$ provided at the output of the classifiers, which is ordered in a decreasing order according to the probability, decide whether the best music genre hypothesis, which is so far called the TOP 1, can be accepted or not. Therefore, the conventional classification approach is modified. Now, the highest *a posteriori* probability provided by the MAX operator is not simply accepted, but it is compared with a threshold (λ). If such a probability is greater than λ then the music genre is assigned to the input vector, otherwise, no label is assigned to the input vector $\overline{x}$ and it is rejected. This novel decision scheme is show in Figure 2).

In summary, the rejection rule is given by:

1. The TOP 1 music genre hypothesis is accepted whenever $P(\omega'|\overline{x}) \geq \lambda$
2. The TOP 1 music genre hypothesis is rejected whenever $P(\omega'|\overline{x}) < \lambda$

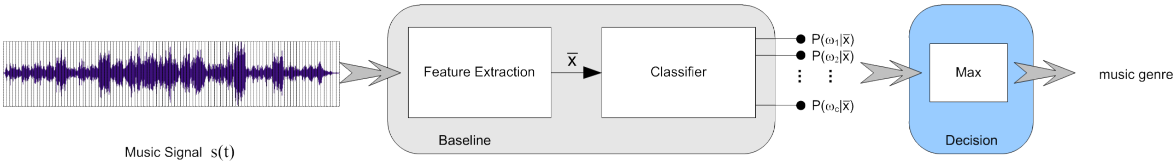

Figure 1. An overview of the baseline music genre classification system

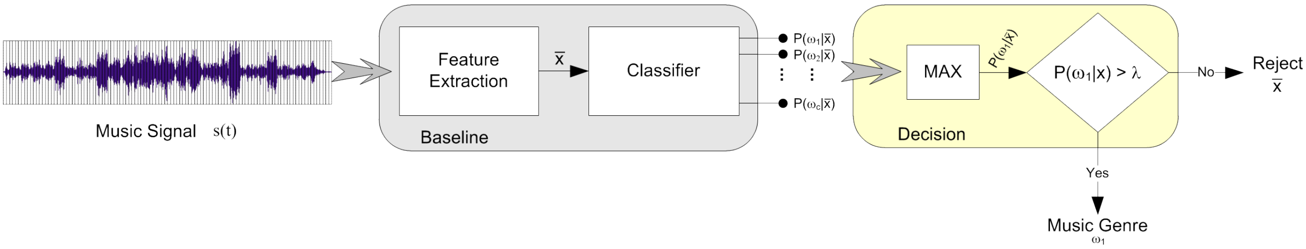

Figure 2. An overview of the rejection scheme for the baseline music genre classification

where λ is a rejection threshold and $P(\omega'|\overline{x})$ is the *a posteriori* probability assigned by the classifier to the best music genre hypothesis ω'.

4. VERIFICATION OF MUSIC GENRES

The question that may arise is if we can do better than simply rejecting the instances that the classifier is not able to classify with confidence. Probably the straightforward way to proceed is to re-classify the rejected instances using a different kind of classifier or even, represent such rejected instances using a different feature set and submit them to further classification steps [3, 5].

The solution that we propose in this paper is to make use of all information provided at the output of the classifier, that is, the ranked list of the N best music genre hypothesis $[P(\omega_1|\overline{x}), \ldots, P(\omega_c|\overline{x})]^T$, to deal with the rejected instances. The main idea is to re-classify the rejected instances using specialized classifiers. To such an aim we assume that the baseline classifiers are somewhat trustworthy. This means that even if the baseline classifier is not able to rank the correct music genre at the TOP 1 position, the correct genre hypothesis will show up among the first best hypothesis. Therefore, we need to analyze carefully the output of the baseline classifier to understand its behavior. In particular, our interest is to find out if there is any relevant information that can be used to guide us in building more specialized classifiers.

To such an aim, we look at the confusion matrix of the baseline classifiers. A confusion matrix shows us how the errors are distributed across the classes. Our prime interest is to find out where the most significant misclassification has occurred. In particular, we look for large off-diagonal entries of the matrix which might indicate a difficult two-class problem that needs to be tackled separately. Based on the analysis of the confusion matrix, we can build two-class classifiers to handle the most significant confusions.

Consider a set $\mathcal{D}$ of L two-class classifiers. Once the TOP 1 music genre hypothesis is rejected we look at the class of the second best music genre hypothesis, denoted as ω'' which is computed by Equation 2:

$$\omega'' = \arg\max2_{\omega_1 \leq \omega_i \leq \omega_c} P(\omega_i|\overline{x}) \qquad (2)$$

where max2 is an operator that returns the second greatest probability provided by the classifier.

Let ω' and ω'' the music genre assigned to the TOP 1 and TOP 2 hypothesis respectively, then we verify if there exists a binary classifier, so far called verifier, $D(\omega', \omega'') \in \mathcal{D}$ to handle the confusion between the music genres ω' and ω''. If so, the verification stage is invoked. Otherwise a final decision is taken and the input vector $\overline{x}$ is rejected. The complete classification and verification approach can be summarized by the pseudo-code as follows:

```
CLASSIFY INSTANCE ($\overline{x}, B, \mathcal{D}$)
    // Input: An instance represented by a feature vector $\overline{x}$, a
    // multiclass baseline classifier $B$, a set of two-class
    // verifiers $\mathcal{D}$
    // Output: A music genre assigned to $\overline{x}$ or the
    // rejection of $\overline{x}$
    if $P_B(\omega' \mid \overline{x}) \geq \lambda$ then
        return $\omega'$ as the music genre
    else
        read $P(\omega'' \mid \overline{x})$
        if $D_{\omega',\omega''} \in \mathcal{D}$ then
            classify $\overline{x}$ with $D_{\omega',\omega''}$
            return $\arg\max_{\omega',\omega''}\{P_D(\omega' \mid \overline{x}), P_D(\omega'' \mid \overline{x})\}$ as
                the music genre
        else
            reject the instance $\overline{x}$
        endif
    endif
```

The verification stage is loosely coupled to the baseline classifier. The only information provided by the baseline classifiers is the label of the two best hypothesis, say ω' and ω''. Such information is used to selected the proper verifier. The verifier, may or may not use the same feature set and classification algorithm of the baseline, however, in this paper we have chosen a different feature set which is based on the texture features extracted from the spectrogram of the music signal [1]. This feature set was chosen because it has provided very interesting discriminat-

N best hypothesis	Correct Classification Rate (%)	
	Baseline SVM	Baseline MLP
TOP 1	53.00 ± 0.67	52.11 ± 3.09
TOP 2	71.11 ± 1.39	70.34 ± 2.62
TOP 3	81.33 ± 0.00	79.82 ± 2.93
TOP 5	91.00 ± 1.45	92.10 ± 1.09
TOP 7	95.89 ± 0.84	95.21 ± 2.27
TOP 8	97.78 ± 0.96	98.33 ± 1.03

Table 1. Correct classification rate for the baseline classifiers

ing results in the previous works [1]. This 59-dimensional feature vector is used to train binary SVM verifiers. An overview of the complete approach is shown in Figure 3.

5. EXPERIMENTAL RESULTS

The performance of the classification-verification approach was evaluated on a subset of the LMD [11]. The LMD is made up of 3,227 full-length music pieces uniformly distributed along 10 classes: Axé (Ax), Bachata (Ba), Bolero (Bo), Forró (Fo), Gaúcha (Ga), Merengue (Me), Pagode (Pa), Salsa (Sa), Sertaneja (Se), and Tango (Ta). In our experiments, we use 900 music pieces from the LMD, which are split into 3 folds of equal size (30 music pieces per class). The splitting is done using an artist filter, which places the music pieces of an specific artist exclusively in one, and only one, fold of the dataset. Furthermore, in our particular implementation of the artist filter we added the constraint of the same number of artists per fold. Therefore, all results reported in this section refer to 3-fold cross validation, unless otherwise noted.

5.1 Baseline Classifiers

The first baseline classifier is a 10-class multilayer perceptron neural network with 68 input units, 40 units at the hidden layer and 10 units at the output layer and the following learning parameters: step width = 0.2, momentum = 0.5, flat spot elimination = 0.1, max non-propagated error = 0.1, and 100 learning cycles. After such a number of cycles, the generalization of the network starts to decrease according to the mean squared error measured on a validation dataset. The network provides estimates *a posteriori* probabilities and the value of each output necessary remains between zero and one because of the sigmoidal function used. The second baseline classifier is a 10-class SVM with Gaussian kernel. The gamma and cost parameters were found by a grid search on a validation dataset. Pairwise coupling is used to handle multi-class classification.

The performance of the baseline classifiers was evaluated using the correct classification rate which is defined as the ratio between number of samples correctly classified and the number of samples tested. Table 1 shows the performance of both baseline classifiers taken into account if the correct music genre is among the TOP N best hypothesis. These results support our previous assumption that the baseline classifiers provide a somewhat trustworthy output. For instance, the correct music genre is among the TOP 5 best hypotheses for more than 91% of the cases.

Class	Ax	Ba	Bo	Fo	Ga	Me	Pa	Sa	Se	Ta
Ax	41	3	0	0	9	2	10	7	18	0
Ba	2	67	4	5	1	3	3	3	2	0
Bo	1	4	45	6	2	1	10	11	5	5
Fo	0	6	3	43	13	3	8	3	11	0
Ga	13	0	5	8	44	8	6	5	1	0
Me	1	3	1	2	6	66	3	5	3	0
Pa	8	4	11	7	0	1	38	15	5	1
Sa	12	2	9	3	3	12	12	33	4	0
Se	14	1	7	9	7	2	6	9	35	0
Ta	1	0	7	0	3	0	0	0	0	79

Table 2. Confusion matrix for the verification set

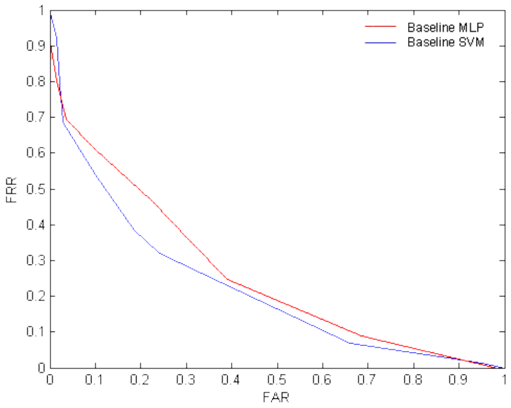

Figure 4. ROC curve for the SVM baseline and MLP baseline.

The information on Table 1 opens up a plenitude of ways to improve the performance of the baseline classifiers. However, this is not the goal of this paper. Here we focus on the concept of rejection in an attempt to improve the reliability of the baseline classifiers, recalling that reliability is the proportion of correct answers among the accepted instances as a function of the rejection rate.

5.2 Rejection Option

In this section we evaluate how the rejection can improve the reliability of the baseline classifiers. We measure the rejection accuracy in terms of its rate of erroneous behavior for each input as false rejection rate $FRR = FR/(FR + CA)$ on instances which had been recognized correctly, and false acceptance rate $FAR = FA/(FA + CR)$ on instances which had been misrecognized, where CA denotes correct acceptance, CR denotes correct rejection, FA denotes false acceptance, and FR denotes false rejection. These two types of error naturally trade off; for example, raising the rejection threshold reduces FAR but at the cost of increased FRR. Therefore, for each measure, we sweep a rejection threshold (λ) across its entire range of values, plotting the two error types as a receiver-operating characteristic (ROC) curve. A curve reaching closer to the origin indicates a superior confidence measure, one enabling low rates of both error types simultaneously. Figure 4 shows the ROC curves as a function of the rejection rate which is defined in terms of the λ value.

Figure 4 shows that the curve reaches close to the ori-

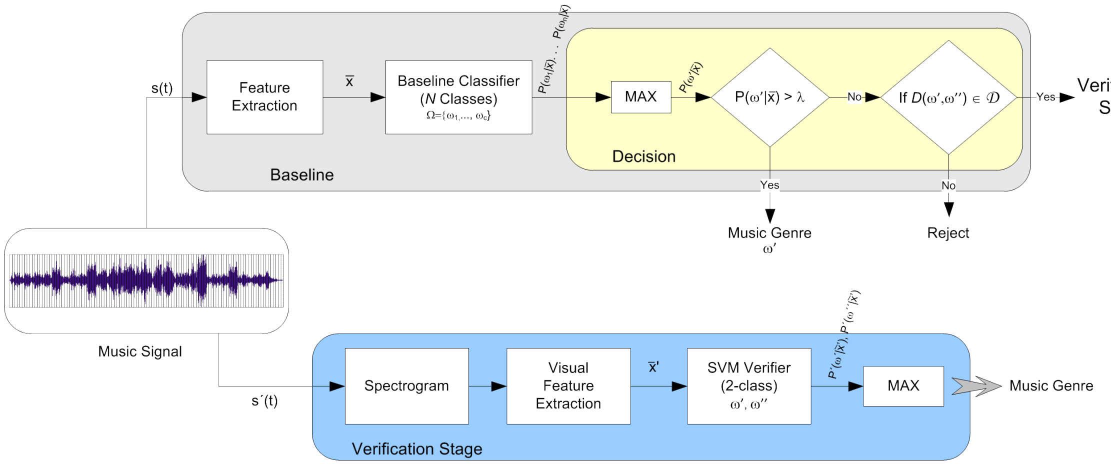

Figure 3. An overview of the complete approach including classification, rejection and verification

Rejection Rate (%)	Error Rate (%)	
	Baseline SVM	Baseline MLP
0	47.00	47.89
20	38.59	40.68
40	30.46	32.18
50	23.94	30.34

Table 3. Reduction on the error rate for different rejection rates

gin for FAR and FRR equal 0.3. This corresponds to rejection rates between 20% and 50%. Table 3 shows the corresponding error rates for these rejection levels. Therefore, for the remainder of the paper we consider 40% of rejection rate. At this rejection rate the reliability for music genre increases from 53.00% to 69.54% and from 52.11% to 67.82% for the SVM and MLP classifier respectively. This represents an improvement of about 16% which is very interesting in the context of music genre classification.

5.3 Verifiers

Since the aim of this paper is not to handle all possible confusions but to show how the rejection and verification can improve the reliability of music genre classification systems, we have built very few verifiers among the 45 possible ones. Therefore, there are binary verifiers only to deal with the most significant confusions which were found by analyzing the confusion matrix generated from the output of the baseline classifiers on a validation dataset. The two-class verifiers employ the support vector machines algorithm with Radial Basis Function kernel and trained with the sequential minimal optimization method. A grid-search algorithm was used to optimize the cost and the gamma parameters. The two-class classifiers were built based on the analysis of the confusion matrix shown in Table 2. Since our goal is not to handle all the confusion of the base classifiers, in this table are highlighted the two highest confusions, which are between classes Axé and Sertaneja and between classes Pagode and Salsa. Therefore, two binary verifiers were built to deal with the confusing classes $\mathcal{D} = \{D_{\text{Ax,Se}}, D_{\text{Pa,Sa}}\}$.

Classifier	Correct Classification Rate (%)	
	Ax-Se Class	Pa-Sa Class
Baseline SVM	42.22 ± 0.12	39.44 ± 0.01
Baseline MLP	43.33 ± 0.11	40.56 ± 0.13
Verifier	82.08 ± 13.05	78.33 ± 3.33

Table 4. Correct classification rates for two classes

Approach	Rejection Rate (%)	Reliability (%)
SVM	0	53.00 ± 0.67
MLP	0	52.11 ± 3.09
SVM + Rej	40	69.54 ± 0.99
MLP + Rej	40	67.82 ± 3.68
SVM + Rej + Verif	40	71.35 ± 1.00
MLP + Rej + Verif	40	69.19 ± 3.93

Table 5. Reliability for the baseline SVM and MLP, baseline+rejection (Rej) and baseline+rejection+verification (Verif)

The performance of the two-class verifiers is shown in Table 4. The results refer to the same 3-fold cross validation protocol but for the verifier we have just a subset where each fold holds only the instances labeled with the confusing classes. We include in this table also the performance of the baseline classifiers considering only the joint correct classification rate on the instances of these pair of classes.

Table 4 shows, that as expected, the verifier achieves a much higher rate than the baseline classifiers. However, our main aim is to improve the overall reliability using the verifier. Therefore, we apply the approach proposed in Section 4, where we submit an instance to the verifier if it is rejected and if it is classified by the baseline classifier at the TOP 1 and TOP 2 positions as one of the two pair of classes: Ax-Se or Se-Ax and Pa-Sa or Sa-Pa. Table 5 summarizes the results of the complete approach.

Table 5 shows that both rejection and verification are effective in improving the correct music genre classification rate. Rejection brings about an impressive increasing in the correction classification rate, however, 40% of the instances were rejected and should be handled in a different way, such as by humans. An alternative way is to have a second stage to reprocess the rejected instances. Even if we have not handled all the rejected instances, but only those who felt on the most confusing classes that we have chosen, it is possible to observe a further, but moderated, improvement in the classification rates.

The Friedman test with the post hoc Shaffer's static procedure was employed to evaluate if there are statistically significant differences between the results show in Table 5. The multiple comparison statistical test has show that the p-value is lower than the corrected critical value in most of the cases, showing a statistically significant difference between the baseline, baseline+rejection and baseline+ rejection+verification at 95% confidence level.

6. CONCLUSIONS

In this paper we have presented a rejection and verification approach to automatic music genre classification that post-process the output of a baseline classifier in an attempt to improve the reliability in classifying music genres. The output of the baseline classifiers is evaluated and the probabilities provided by these classifiers serve as a guide to either reject or accept the input instance. Furthermore, the rejected instances may be re-classified at a verification stage using a different approach if they were previously classified by the baseline classifier as belonging to specific classes. The performance resulting from the combination of the baseline, rejection and verification is significantly better than that achieved by the baseline classifiers alone. For instance, the baseline classifiers alone achieves a classification rate of 53%. The rejection stage improves the reliability to 69.54% at a rejection level of 40% and the verification stage further improves it to 71.35%. In spite of the current verification stage deal with only two pair of confusing classes, it was able to improve the reliability in almost 2%. Given the high number of confusions between other classes, as show in Table 2, we expect to achieve further improvement by adding more verifiers out of the 45 possible ones.

Compared with previous works that use the same features, the same dataset, and the same experimental protocol [1, 7], the results reported in this paper represent a significant improvement in terms of classification rate and reliability. It is difficult to compare the performance of the proposed approach with other results available in the literature due to the differences in the experimental conditions.

In spite of the good results achieved, there are some shortcomings related to the use of the second stage. For instance, the second stage depends on the results of the first stage and on the availability of a binary classifier to handle the confusions between specific classes. As future work we plan to validate the proposed approach on other datasets, such as the Magnatagatune and the Million Song Dataset.

7. ACKNOWLEDGMENTS

This research is partially supported under CAPES/Fulbright grant BEX1770/712-9, FA grant 203/12 and CNPq grants 306.703/2010-6 and 472.238/2011-6.

8. REFERENCES

[1] Y.M.G. Costa, L.E.S. Oliveira, A.L. Koerich, F. Gouyon, and J.G. Martins. Music genre classification using LBP textural features. *Signal Processing*, 92(11):2723–2737, 2012.

[2] Z. Fu, G. Lu, K. M. Ting, and D. Zhang. A survey of audio-based music classification and annotation. *IEEE Trans. on Multimedia*, 13(2):303–319, 2011.

[3] A.L. Koerich. Rejection strategies for handwritten word recognition. In *Int'l Workshop on Frontiers in Handwriting Recognition*, pp.479–484, Tokyo, Japan, 2004.

[4] A.L. Koerich, L.E.S. Oliveira, and A.S. Britto Jr. Verification of unconstrained handwritten words at character level. In *Int'l Conf. on Frontiers in Handwriting Recognition*, pp.39–44, Kolkata, 2010.

[5] A.L. Koerich, R. Sabourin, and C. Y. Suen. Recognition and verification of unconstrained handwritten words. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(10):1509–1522, 2005.

[6] T. Lidy, C.N. Silla, 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 and accessing non-western and ethnic music collections. *Signal Processing*, 90(4):1032–1048, 2010.

[7] M. Lopes, F. Gouyon, A.L. Koerich, and L.E.S. Oliveira. Selection of training instances for music genre classification. In *Int'l Conf. on Pattern Recognition*, pp.4569–4572, Istambul, Turkey, 2010.

[8] S.R. Ness, A. Theocharis, G. Tzanetakis, and L.G. Martins. Improving automatic music tag annotation using stacked generalization of probabilistic SVM outputs. In *ACM Multimedia Conf.*, pp.705–708, Beijing, China, October 2009.

[9] N. Scaringella, G. Zoia, and D. Mlynek. Automatic genre classification of music content - a survey. *IEEE Signal Processing Magazine*, 23(2):133–141, 2006.

[10] J. Schurmann. *Pattern Classification: A Unified View of Statistical and Neural Approaches*. John Wiley and Sons, 1996.

[11] C.N. Silla, A.L. Koerich, and C.A.A. Kaestner. The Latin music database. In *Int'l Conf. on Music Information Retrieval*, pages 451–456, Philadelphia, USA, 2008.

[12] C.N. Silla, A.L. Koerich, and C.A.A. Kaestner. A machine learning approach to automatic music classification. *Journal of the Brazilian Computer Society*, 14(3):7–18, September 2008.

[13] B. L. Sturm. A survey of evaluation in music genre recognition. *Adaptive Multimedia Retrieval*, 2012.

[14] B. L. Sturm. Two systems for automatic music genre recognition: What are they really recognizing. In *Int'l ACM Workshop on MIR with User-centered and Multimodal Strat.*, pp.69–74, Nara, Japan, 2012.

[15] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Trans. on Speech and Audio Processing*, 10(5):293–302, 2002.

[16] T.-F. Wu, C.-J. Lin, and R.C. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005, 2004.