CONFIDENCE-BASED LATE FUSION FOR MUSIC GENRE CLASSIFICATION

Ming-Ju Wu
Computer Science Department
National Tsing Hua University
Hsinchu, Taiwan
brian.wu@mirlab.org

Jyh-Shing Roger Jang
Computer Science Department
National Taiwan University
Taipei, Taiwan
roger.jang@mirlab.org

ABSTRACT

Using one type of feature for classifier learning may be inadequate to achieve optimal results. In this submission, the confidence-based late fusion is proposed to combine the acoustic and visual features for music genre classification. The experimental results indicated that the proposed method achieved an accuracy improvement of 7.32% and 6.68% respectively for mixed popular genre classification and Latin music genre classification, demonstrating the effectiveness of our approach.

1. INTRODUCTION

A key factor in genre classification is the use of effective features for classification. Using one type of feature for classifier learning may be inadequate to achieve optimal results. In the literature, acoustic [1] and visual features [4,5] are effective features for music genre classification. Because acoustic features are based on spectral analysis and visual features are based on time-frequency analysis, combining both types of feature may benefit music genre classification. In this submission, GSV (Gaussian super vector) [1, 3, 5] is applied as our acoustic features and MLVFs (multi-level visual features) are applied as our visual features. However, combining acoustic and visual features has rarely been attempted (only the early fusion approach was applied [5]). Therefore, a confidence-based late fusion approach is proposed to combine the decisions made by two individual classifiers (based on acoustic and visual features, respectively) to achieve the final prediction.

2. PROPOSED CONFIDENCE-BASED LATE FUSION

For late fusion, the fusion is performed after classification. To perform the proposed confidence-based late fusion, two quantities from SVMs are measured. Figure 1 shows the flowchart of this process. The prediction of the multiclass SVM is based on the one-against-one approach. If the predicted classes of the two multiclass SVMs, ω_{GSV} and ω_{MLVF}, are different, then the confidence measures of the pair of the binary-class SVMs (corresponding to classes $\{\omega_{GSV}, \omega_{MLVF}\}$), c_{GSV} and c_{MLVF}, are computed and compared to complete the final prediction. In other words, the final prediction is taken from the binary classifier with a higher confidence measure. Because different types of feature may exhibit different discriminative powers for a given music clip, confidence-based late fusion selects a presumably more accurate prediction. Next, we describe the basic concept of the SVM and how to compute its confidence measure from two confidence factors.

The goal of a binary-class SVM is to identify the hyperplane (i.e., decision boundary) with the widest separation between two classes of training data, which can be expressed as:

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = \sum_{i=1}^{l} \lambda_i y_i \mathbf{x}_i^T \mathbf{x} + b, \qquad (1)$$

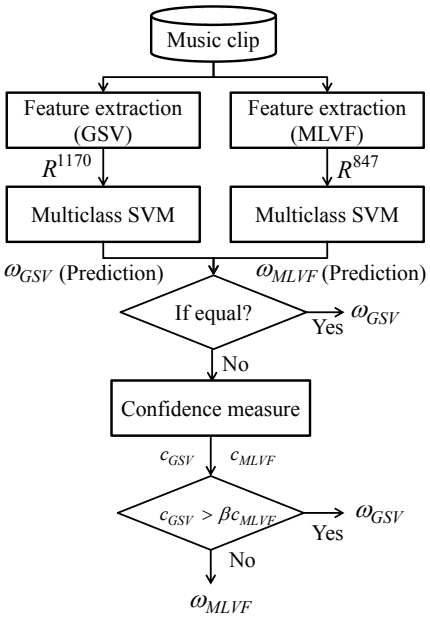

Figure 1. Flowchart of the proposed confidence-based late fusion.

This document is licensed under the Creative Commons Attribution-Noncommercial-Share Alike 3.0 License.
http://creativecommons.org/licenses/by-nc-sa/3.0/
© 2014 The Authors.

where $\mathbf{x}$ is the feature vector of the test instance; $\mathbf{w}$ is a normal vector; b is the bias term in the hyperplane; $\mathbf{x}_i$ is a d-dimensional feature vector of training instances; y_i is the label (ground truth) of $\mathbf{x}_i$, which is set at either 1 or -1 to distinguish between the two classes; l is the number of music clips in the training set; λ_i is the Lagrange multiplier, which can be either zero or positive. Specifically, the optimal hyperplane is the linear combination of $\mathbf{x}_i$ with $\lambda_i > 0$. These $\mathbf{x}_i$ are support vectors, which support the maximum-margin and create the optimal hyperplane. Predicted class ω of test instance $\mathbf{x}$ is either 1 or -1, depending on whether the sign of $\mathrm{g}(\mathbf{x})$ is positive or negative.

To facilitate data separation, a linear mapping ϕ was applied to transform feature vector $\mathbf{x}_i$ into a new space with high dimensionality. According to the kernel trick, the inner product in the high dimensional space can be expressed as kernel function K in the original space. The optimal hyperplane can then be expressed as

$$\begin{aligned}\mathrm{g}(\mathbf{x}) &= \sum_{i=1}^{l} \lambda_i y_i \phi(\mathbf{x}_i)^T \phi(\mathbf{x}) + b \\ &= \sum_{i=1}^{l} \lambda_i y_i \,\mathrm{K}(\mathbf{x}_i, \mathbf{x}) + b.\end{aligned} \tag{2}$$

In this submission, the widely used radial basis function (RBF) kernel was applied.

$$\mathrm{K}(\mathbf{x}_i, \mathbf{x}) = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}\|^2}{2\sigma^2}\right) \tag{3}$$

Because the corresponding linear mapping ϕ transforms data to the Hilbert space (i.e., a vector space with infinite dimensions) for classification, two confidence factors in the Hilbert space were proposed.

1. *Confidence factor 1: The distance between the test instance and the hyperplane in the Hilbert space*
   The goal of an SVM is to identify the hyperplane with the maximal margin between two classes of training data. Consequently, the prediction of the test instance is likely to be correct if the instance is far from the hyperplane. The distance between the test instance and the hyperplane can be expressed as

   $$\frac{|\mathrm{g}(\mathbf{x})|}{\|\mathbf{w}\|}. \tag{4}$$

   To allow this distance to be directly comparable, Equation (4) was normalized by dividing it by the half margin (the distance between support vectors and the hyperplane in the Hilbert space). This normalized distance cf_1 was then used as the first confidence factor:

   $$cf_1 = \frac{\frac{|\mathrm{g}(\mathbf{x})|}{\|\mathbf{w}\|}}{\frac{1}{\|\mathbf{w}\|}} = |\mathrm{g}(\mathbf{x})| \tag{5}$$

   When $cf_1 < 1$, the test instance was inside the margin. When $cf_1 = 1$, the test instance was on the margin. When $cf_1 > 1$, the test instance was outside the margin. Consequently, a greater cf_1 tends to reflect higher confidence.

2. *Confidence factor 2: The distance between the test instance and its nearest neighbor in the Hilbert space*

   As demonstrated in Equation (2), a linear mapping ϕ transforms data to a new space with high dimensions. The relationship between training data $\mathbf{x}_i$ and test instance $\mathbf{x}$ in the new space should also be considered. The distance between $\phi(\mathbf{x})$ and $\phi(\mathbf{x}_i)$ in the Hilbert space can be computed in the original space by using the kernel trick.

   $$\begin{aligned}&\|\phi(\mathbf{x}_i) - \phi(\mathbf{x})\|^2 \\ &= (\phi(\mathbf{x}_i) - \phi(\mathbf{x}))^T (\phi(\mathbf{x}_i) - \phi(\mathbf{x})) \\ &= \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_i) \rangle - 2\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle \\ &+ \langle \phi(\mathbf{x}), \phi(\mathbf{x}) \rangle \\ &= \mathrm{K}(\mathbf{x}_i, \mathbf{x}_i) - 2\,\mathrm{K}(\mathbf{x}_i, \mathbf{x}) + \mathrm{K}(\mathbf{x}, \mathbf{x})\end{aligned} \tag{6}$$

   According to Equation (3) and Equation (6) , the second confidence factor can be expressed as

   $$cf_2 = \min_{i, \text{with } \mathbf{x}_i \text{ in class } \omega} \left\{ 2 - 2\exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}\|^2}{2\sigma^2}\right) \right\} \tag{7}$$

   That is, cf_2 computes the minimal distance between $\phi(\mathbf{x})$ and $\phi(\mathbf{x}_i)$, where $\mathbf{x}_i$ are the training instances of the same class as predicted class ω. A lower cf_2 indicates a higher similarity between $\phi(\mathbf{x})$ and $\phi(\mathbf{x}_i)$, and thus ω should be more convincing.

   The confidence measures c_{GSV} and $c_{MLVF}$are then defined as follows:

   $$\begin{cases} c_{GSV} = \frac{cf_{1\,(GSV)}}{cf_{2\,(GSV)}} \\ c_{MLVF} = \frac{cf_{1\,(MLVF)}}{cf_{2\,(MLVF)}} \end{cases} \tag{8}$$

   Therefore, a greater cf_1 and a smaller cf_2 lead to a higher confidence. The final decision can be determined according to

   $$\text{Apply } \omega_{GSV} (\omega_{MLVF}) \text{ if } \mathrm{c}_{GSV} > (\leq) \beta c_{MLVF} \tag{9}$$

   where β represents the weighting for adjusting the importance of c_{GSV} and c_{MLVF}. When c_{GSV} was larger than βc_{MLVF}, ω_{GSV} was applied as the final decision. Otherwise, ω_{MLVF} was applied. In this submission, β was temporarily set to 1. The well-known SVM tool, LIBSVM [2], with a RBF kernel was applied as the classifier.

Method	Task	Accuracy
Our 2013 submission[a]	Mixed popular genre classification	76.23%
Our 2014 submission[b]	**Mixed popular genre classification**	**83.55%**
Our 2013 submission[a]	Latin genre classification	71.96%
Our 2014 submission[b]	**Latin genre classification**	**78.64%**

Table 1. Performance comparison for early fusion and confidence-based late fusion.

[a] GSV+MLVFs with early fusion.
[b] GSV+MLVFs with confidence-based late fusion.

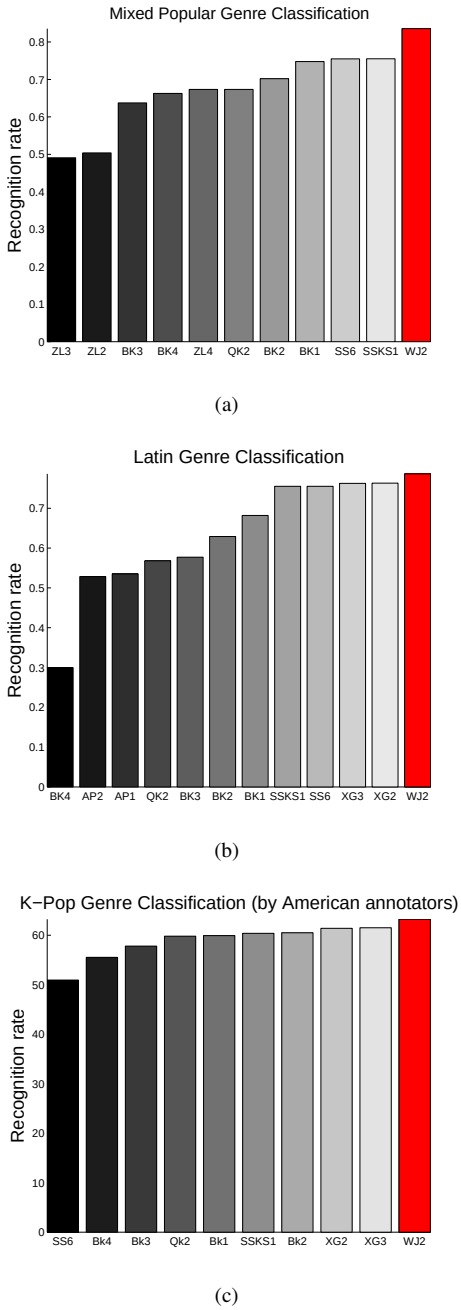

Figure 2. MIREX 2014 Contest Results. Our submission name is WJ2.

3. EXPERIMENTAL RESULTS

Table 1 presents the performance comparison for early fusion and confidence-based late fusion. [1] As can be seen, experimental results indicates that the proposed method achieved an accuracy improvement of 7.32% and 6.68% respectively for mixed popular genre classification and Latin music genre classification. Figure 2 shows that our approaches achieved the highest accuracy rates for mixed popular genre classification, Latin genre classification, and K-Pop genre classification(by American annotators) in MIREX 2014. This indicates the proposed confidence-based late fusion approach can successfully combine both acoustic and visual features. Because both spectral and time-frequency aspects were utilized, considerably increasing the discriminating power of the features. This is vital to the success of music genre classification.

4. CONCLUSION

In this submission, the proposed confidence-based late fusion can effectively utilize both acoustic and visual features. The superior performance indicates the feasibility and robustness of our approaches. Because the proposed confidence-based late fusion is a generic scheme for combining multiple decisions from SVM classifiers using different features, future studies should also explore the possibility of applying the proposed fusion method to other machine learning tasks.

5. REFERENCES

[1] Chuan Cao and Ming Li. Thinkit's submission for mirex 2009 audio music classification and similarity tasks, 2009.

[2] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machine, 2010.

[3] Zhi-Sheng Chen, Jyh-Shing Roger Jang, and Chin-Hui Lee. A kernel framework for content-based artist recommendation system in music. *IEEE Transactions on Multimedia*, 13(6):1371–1380, 2011.

[4] Y.M.G. Costa, L.S. Oliveira, A.L. Koerich, F. Gouyon, and J.G. Martins. Music genre classification using

[1] We can not provide the performance comparison for K-pop genre classification, because this is a newly proposed task in MIREX 2014.

lbp textural features. *Signal Processing*, 92(11):2723–2737, 2012.

[5] Ming-Ju Wu, Zhi-Sheng Chen, Jyh-Shing Jang, Jia-Min Ren, Yi-Hsung Li, and Chun-Hung Lu. Combining visual and acoustic features for music genre classification. In *Proceedings of the International Conference on Machine Learning and Applications (ICMLA)*, volume 2, pages 124–129. IEEE, 2011.