THE BODHIDHARMA SYSTEM AND THE RESULTS OF THE MIREX 2005 SYMBOLIC GENRE CLASSIFICATION CONTEST

Cory McKay
Music Technology Area
Faculty of Music
McGill University
Montreal, Quebec, Canada
cory.mckay@mail.mcgill.ca

Ichiro Fujinaga
Music Technology Area
Faculty of Music
McGill University
Montreal, Quebec, Canada
ich@music.mcgill.ca

ABSTRACT

This paper discusses the results of the MIREX 2005 symbolic genre classification contest and describes the Bodhidharma system, which attained the highest classification success rates in all four of the evaluated categories.

Five systems were submitted to this contest, which was conducted independently at the University of Illinois at Urbana-Champagne (UIUC). Each system was evaluated in two different experiments, one involving thirty-eight genre classes and one involving nine classes. Success rates were measured in two ways: one based only on how well each system was able to find the single correct genre of each recording, and the other giving partial scores to incorrect classifications that were relatively close to the correct genre. Evaluations were performed using stratified cross-validation.

Bodhidharma is a sophisticated system that utilizes a combination of flat, hierarchical and round-robin classification strategies based on classifier ensembles consisting of feedforward neural networks and k-nearest neighbour classifiers. Bodhidharma bases its classifications on 111 high-level features that it extracts from MIDI recordings. Each classifier ensemble uses genetic algorithms to evolve a weighted sub-set of the features that are appropriate for that particular ensemble.

Keywords: genre, classification, hierarchical, features, music, MIDI

1 INTRODUCTION

MIREX 2005 was an event in which researchers in music information retrieval submitted systems for the purposes of algorithm evaluation and sharing. This event was inspired by the ISMIR 2004 Audio Description Contest held at the Universitat Pompeu Fabra. The MIREX evaluations were conducted independently at UIUC. Symbolic genre classification was one of nine evaluation topics to which participants submitted entries.

Musical genre is used by retailers, librarians, musicologists and listeners in general as an important means of organizing music. The need for an effective automatic means of classifying music is becoming increasingly pressing as the number of recordings available continues to increase at a rapid rate. Software capable of performing automatic classifications would be highly useful to the administrators of the rapidly growing networked music archives, as their success is very much linked to the ease with which users can browse through and search for music on their sites. These sites currently rely on manual genre classification, a methodology that is slow, unwieldy and often inconsistent

There has been a significant amount of recent research in audio classification. Although this research is certainly very valuable, the current lack of reliable polyphonic transcription techniques makes it difficult to impossible to reliably extract high-level musical features from audio recordings. Most research to date has therefore made use of primarily low-level, signal-processing based features. Some initial success has been achieved with such features, but classification rates seem to have stabilized recently.

The problem of implementing an effective genre classifier that deals with realistic taxonomies has yet to be solved. It is therefore appropriate to take advantage of whatever resources are available in order to improve performance. There is a large body of existing recordings in MIDI format from which high-level musical features can in fact be extracted and experimented with. Such research in high-level features will become highly useful as improvements in signal processing allow them to be extracted from audio recordings. Furthermore, research with MIDI files can be applied to electronic or paper scores for which recordings are not available.

In addition to its practical applications, a system that can automatically classify recordings by genre using high-level musical features has significant theoretical musicological interest as well. There is currently a relatively limited understanding of how humans construct musical genres, the mechanisms that they use to classify music and the characteristics that are used to perceive the differences between different genres. A system that could automatically classify music and reveal what musical dimensions it is using to do so would therefore be of great interest.

Section 2 of this paper describes the Bodhidharma system, which can classify MIDI recordings using high-level features and was submitted to MIREX. Section 3 describes the details of the MIREX evaluation experiments, and Section 4 describes the results. Section 5 discusses the current state of the Bodhidharma and what the plans are for its future development.

2 THE BODHIDHARMA SYSTEM

2.1 Introduction

Bodhidharma was developed as a tool for automatic classification of MIDI files (McKay 2004). It includes an easy-to-use graphical interface and is implemented entirely in Java, making it highly portable.

2.2 Feature extraction

Bodhidharma extracts a total of 111 high-level features from MIDI files. These are based on instrumentation, musical texture, rhythm, dynamics, pitch statistics, melody and chords. These features are described in more detail, along with 49 additional features, in McKay's work (2004).

Two types of features were used: one-dimensional features and multi-dimensional features. One-dimensional features each consist of a single number that represents an aspect of a recording in isolation. Multi-dimensional features consist of sets of related values that have limited significance taken alone, but together may reveal meaningful patterns. For example, the bins of a histogram consisting of the relative frequency of different melodic intervals were treated as a multi-dimensional feature, but the average durations of melodic arcs were treated as one-dimensional features. The reason for this differentiation is explained in the following section.

2.3 Basic classification methodology

Bodhidharma uses two classification techniques as the basic units in its classification scheme: feedforward neural networks (NN) and k-nearest neighbour classifiers (k-NN). NNs have the advantage of being able to simulate sophisticated logical relationships between features, but can require long training times. k-NN classifiers, in contrast, cannot simulate such logical relationships, but require essentially no training time. The use of both techniques allows one to use NNs where the modelling of more sophisticated relationships between features is likely to be most beneficial, while using k-NN classifiers elsewhere in order to limit training times.

The relative advantages and disadvantages of these two approaches was the motivation behind the one-dimensional and multi-dimensional features discussed in Section 2.2. Each multi-dimensional feature was classified by a separate multi-dimensional neural network, thus increasing the likelihood that appropriate relationships would be learned between the components of each multi-dimensional feature. The one-dimensional features, in contrast, were all processed by a single k-NN classifier. This greatly reduced the training time, as the majority of features were one-dimensional, and training neural networks to process them would have been quite time consuming.

Feature selection was performed in several stages, all of which used genetic algorithms (GAs). To begin with, the least promising features were eliminated for the k-NN classifier and the remainder were weighted. Another selection was then performed among the k-NN classifier and each of the NN classifiers. Finally, weightings were evolved for each of the classifiers.

The result of all of this after training was a weighted ensemble of classifiers consisting of a single k-NN classifier using a weighted subset of all candidate one-dimensional features and a set of NNs representing a subset of all candidate multi-dimensional features. Such a classifier ensemble can be seen as a black box that took in the entire feature set of a recording as input, and output a classification score for each candidate category that it had been trained to recognize. This is illustrated in Figure 1.

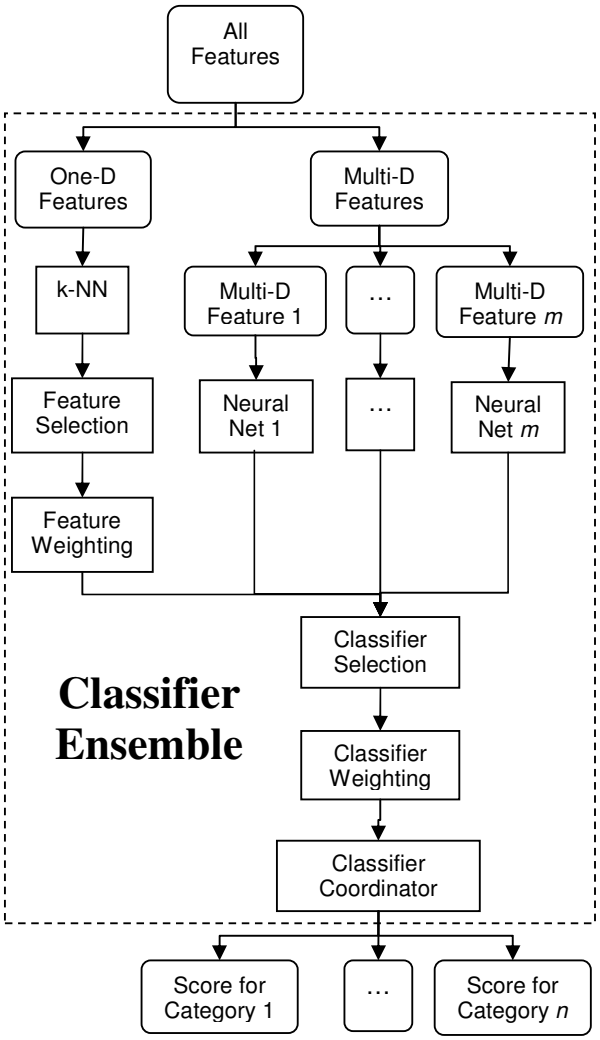

Figure 1: Architecture of Bodhidharma's classifier ensembles.

2.4 Combining the classifier ensembles

A number of these black boxes were trained to perform different tasks. One was trained to perform overall flat classification, one was trained for each possible pair of genres for use in round-robin classification and one was trained for each node in the taxonomical hierarchy in order to perform hierarchical classification.

The hierarchical classification was particularly effective. Recordings were first classified into broad categories, such as Jazz or Classical, and classification then proceeded to the next level of the hierarchy, where only the sub-genres of the winners of the previous stage of classification were considered as candidates. This continued iteratively down the hierarchy of genres until only leaf genres (i.e. genres with no sub-categories) remained, and these were chosen as the winning genres.

The classification at each level of the hierarchy involved separately trained specialist classifier ensembles of the type presented in Section 2.3. Each of these classifier ensembles was trained only on recordings belonging to their candidate genres, and therefore developed feature selections and weightings especially suited to their genres. A root classifier would therefore likely be good at

making coarse classifications, but a Jazz classifier would likely be better at classifying a recording into specialized sub-genres of Jazz once the recording had been labelled as Jazz by the root classifier.

Hierarchical classification has the potential weakness that a mistake made at a broad level of the hierarchy can lead to a decent through an entirely erroneous branch of the hierarchy. Basic flat classification was therefore performed as well in order to counteract this, and round-robin classification was also performed in order to use highly specialized classifiers that could make particularly difficult classifications when necessary. The results of all of these classification schemes were combined using a weighted sum with pre-set weights. 60% of the weight was assigned to the hierarchical classifier.

3 CONTEST FRAMEWORK

3.1 Original contest proposal

In order for each MIREX contest to take place, it had to be proposed by an interested participant, who submitted a draft proposal. This was Cory McKay in the case of the symbolic genre classification contest. This proposal was then discussed and refined on-line by all interested researchers, and then submitted to the MIREX organizers for review. Further changes were then discussed and implemented in order to arrive at the final proposal and several calls were made requesting all participants to submit annotated MIDI files that could be used in the evaluation.

3.2 Details of the contest

Submissions were evaluated using two hierarchical taxonomies. The first, shown in Figure 2, was a relatively artificial taxonomy consisting of three parent classes and nine leaf classes. This taxonomy was used because it is similar in size to the taxonomies that many researchers have used to evaluate their systems, and is therefore useful for benchmarking.

Classical	**Jazz**	**Popular**
Baroque	Bebop	Country
Modern	Jazz Soul	Punk
Romantic	Swing	Rap

Figure 2: The smaller taxonomy used.

The second taxonomy (see Figure 3) consisted of 9 parent classes and 38 unique leaf classes. Several of these leaf classes were allowed to have multiple parents in order to allow for the overlap that is often found when dealing with genre. This much larger taxonomy was significantly more realistic than the first taxonomy and was proposed in order to evaluate how close systems were to being able to perform real-world classification tasks.

A total of 950 MIDI recordings were used for training and testing, 25 for each unique leaf class. Since the smaller taxonomy was a sub-set of the larger taxonomy, the experiments using this taxonomy only involved 225 recordings. Although two collections of MIDI files were submitted to the organizers at UIUC, they decided to use only one of these, as the other was not diverse enough relative to the taxonomies used in the contest.

The initial contest proposal suggested allowing each recording to belong to a variable number of classes, in order to more realistically simulate the ways in which genre behaves in reality. However, it was decided by the community to instead require that each recording belong to one and only genre for the purpose of this contest, as this reflected how most systems had been implemented. Bodhidharma therefore simply output the genre that it had assigned the highest score to, in contrast to its usual practice of outputting all genres with a score over a threshold.

Jazz	**Western Classical**	**Rhythm and Blues**	**Rock**	**Modern Pop**
Bop	Baroque	*Blues*	*Classic Rock*	Adult Contemp.
Bebop	Classical	Blues Rock	Blues Rock	*Dance*
Cool	*Early Music*	Chicago Blues	Hard Rock	Dance Pop
Fusion	Medieval	Country Blues	Psychedelic	Pop Rap
Bossa Nova	Renaissance	Soul Blues	*Modern Rock*	*Techno*
Jazz Soul	Modern Classical	Funk	Alternative Rock	Smooth Jazz
Smooth Jazz	Romantic	Jazz Soul	Hard Rock	
Ragtime		Rock and Roll	Metal	
Swing		Soul	Punk	

Modern Pop	**Worldbeat**	**Western Folk**	**Country**
Adult Contemp.	*Latin*	Bluegrass	Bluegrass
Dance	Bossa Nova	Celtic	Contemporary
Dance Pop	Salsa	Country Blues	Traditional
Pop Rap	Tango	Flamenco	
Techno	Reggae		
Smooth Jazz			

Figure 3: The larger taxonomy used.

3.3 Evaluation procedure

Evaluations were performed independently by the MIREX organizers at UIUC where they performed experiments using the M2K framework (Downie 2004). The organizers provided several template M2K itineraries that could be used to prepare submitted systems for evaluation.

A separate experiment was performed with each taxonomy. Stratified cross-validation was used to evaluate each system. Two different classification success rates were calculated for each experiment:

1. *Raw accuracy:* Each system was given a full point for each correct leaf genre and no points for each incorrect classification. This evaluated the ability of the systems to make precise classifications.
2. *Hierarchical accuracy:* Each system was given a full point for each correct leaf genre. Partial points were given if the selected leaf genre was incorrect, but was in a correct branch of the taxonomy tree. This measure evaluated whether the systems made minor mistakes or major mistakes when they were incorrect.

4 CONTEST RESULTS

The results for the two experiments are shown in Tables 1 and 2:

Table 1: Ranked average results across folds for the taxonomy in Figure 2.

Contest Entry	Raw Accuracy	Hierarchical Accuracy
Bodhidharma	*84.4%*	*90.0%*
Basili et al. (NB)	72.0%	81.6%
Li	72.0%	80.2%
Basili et al. (J48)	65.3%	76.7%
Ponce de Leon & Inesta	37.8%	50.7%

Table 2: Ranked average results across folds for the taxonomy in Figure 3.

Contest Entry	Raw Accuracy	Hierarchical Accuracy
Bodhidharma	*46.1%*	*64.3%*
Basili et al. (NB)	45.0%	62.6%
Basili et al. (J48)	41.0%	57.6%
Li	39.8%	54.9%
Ponce de Leon & Inesta	15.3%	24.9%

These results indicate excellent performance by the Bodhidharma system. Bodhidharma placed first in all four evaluation categories, and performed particularly well in the experiment shown in Table 1, with a raw accuracy 12.4% higher than the second place system.

The large number of features extracted and the sophisticated classification methodology described in Section 2 are most likely responsible for Bodhidharma's excellent performance. Bodhidharma's approach made it possible to perform specialized classifications between sub-sets of genres using specialized features.

The processing time for training and testing was only released for three of the five submissions, and evaluations were run on different machines, thereby making precise time comparisons impossible. Bodhidharma did take significantly longer than the submissions of Li and of Ponce de Leon and Inesta, however. This is not too serious a drawback, as the long training time needed by Bodhidharma must only be dealt with once, and classifications can then be performed quickly once the system is trained. The other systems certainly deserve credit for their efficiency, however, particularly in the case of the submission of Ponce de Leon and Inesta.

Although the results on the small taxonomy were likely good enough for practical use, none of the submissions achieved practically usable results on the larger taxonomy. So, while it is certainly encouraging that results far better than chance were attained by all of the systems, it is clear that there is still much work to be done before automatic symbolic genre classification can be used in a real-world context involving large taxonomies.

5 FUTURE WORK

Bodhidharma is continually being maintained and improved. It is currently being integrated with the general-purpose ACE classification framework (McKay et al. 2005). This framework experimentally applies a variety of machine learning techniques, including classifier ensemble approaches such as Bodhidharma's, to arbitrary classification problems. Bodhidharma's feature extraction system has already been adapted to ACE, and its classification methodology is in the process of being generalized and integrated into ACE.

ACKNOWLEDGEMENTS

The financial support from the Social Sciences and Humanities Research Council of Canada, the Fonds Québécois de la recherche sur la société et la culture, the Centre for Interdisciplinary Research in Music, Media and Technology (CIRMMT) and the McGill Alma Mater Fund is greatly appreciated. The time and effort generously donated by Stephen Downie and the many others involved in organizing and implementing the MIREX events is also very much appreciated.

REFERENCES

Downie, J. S. 2004. International music information retrieval systems evaluation laboratory (IMIRSEL): Introducing D2K and M2K. *Demo Handout at the 2004 International Conference on Music Information Retrieval.*

McKay, C. 2004. Automatic genre classification of MIDI recordings. *M.A. Thesis*. McGill University, Canada.

McKay, C., R. Fiebrink, D. McEnnis, B. Li, and I. Fujinaga. 2005. ACE: A framework for optimizing music classification. *Proceedings of the International Conference on Music Information Retrieval*.