ANALYZING AND IMPROVING GENRE AND STYLE CLASSIFICATION IN MUSIC THROUGH EXPERIMENTS

ZAHRA GHASEMAGHAI
Bachelor of Information Technology, Sheikh-bahaee University, 2010

A Thesis
Submitted to the School of Graduate Studies
of the University of Lethbridge
in Partial Fulfillment of the
Requirements for the Degree

MASTER OF SCIENCE

Department of Mathematics and Computer Science
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

Analyzing and Improving Genre and Style Classification in Music Through Experiments

Approved:

Supervisor: John Zhang

Committee Member: Yllias Chali

Committee Member: Gongbing Shan

Chair, Thesis Examination Committee: Howard Cheng

Dedication

To my beloved parents.

Abstract

Music classification is a core task in the field of *Music Information Retrieval* (MIR). Classification refers to recognizing patterns in data. Music classification assigns *genre, style, mood* and etc. to each piece of music, to facilitate managing music data. It is an interesting topic in MIR with potential applications.

There has been a considerable deal of attention focused on variety issues of music classification, such as selection appropriate feature sets, feature selection techniques, classification algorithm, etc.

In this thesis, a series of empirical experiments are conducted to investigate and evaluate the genre and style classification in music. To validate our investigations and evaluations, several methods are proposed to analyze and interpret the results. In addition, we also design and implement an effective classification approach that obtains higher classification accuracy.

Acknowledgments

First and foremost, I would like to thank Dr. John Zhang for his support and guidance. This thesis could not have been written without him. He not only served as my supervisor but also encouraged me throughout my M.Sc. program. I am grateful to have him as my supervisor. Many thanks to my committee members, Dr. Yllias Chali and Dr. Gongbing Shan for taking the time to read my thesis and for their valuable feedbacks on this work, and also Dr. Hadi Kharaghani for his support. I would like to thank Mr. Mohammad Akbari and Mr. Tom Arjannikov for the help they provided to me throughout this thesis. I am extensively indebted everything in my life to my family. Many thanks to my lovely parents who taught me patience, kindness, and hard work; my sister and brothers, Maryam, Reza, and Ali, who taught me love. I would like to thank my best friend, Houman, for his support. And also I would like to thank my friends whom it was with their supports that my stay in Lethbridge become more pleasant.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Music plays an important role in human entertainment, and accessing music is markedly enhanced with the improvement in knowledge and new advancements in digital music sharing [22]. The size of music repositories has grown considerably during the past years. Thus, searching, organizing, and navigating music are an inherent issue when dealing with enormous repositories. The advances of hardware and software have important effects on using huge datasets in Music Information Retrieval (MIR). Therefore, in the near future, all recorded music in human history will be accessible by everyone on the Internet [57]. Most of the research in MIR uses music content. The main advantage of *content-based approaches* is that a music piece can be represented by a set of features, which are computed directly from sound [18, 25].

In MIR, a variety of data mining technologies have been used to explore the modeling of large music datasets and to discover the relations among music pieces [18]. *Data mining* is the science of mining information and knowledge from data repositories in order to transfer it into efficient and understandable structures for future use. In other words, *data mining* is used to discover interesting and non-trivial information from huge amount of data. *Music data mining* can be split into different tasks, such as *music classification*, and *music data management* [18].

One of the important problems in *music data mining* is *classification*. The most general music classification focuses on *genre/style classification, mood classification, and instrument classification*. The focus of this thesis is classifications in music genre and

style. In the next chapters, different issues of classification will be considered and several approaches to improve the classification performance will be discussed.

1.1 Music Information Retrieval

Music Information Retrieval (MIR) is a multidisciplinary area that includes different fields, such as computer science, psychology, audio engineering, cognitive science, and musicology. End users, such as music students, musicology researchers, and musicians, are interested in different aspects of music, such such as whether it can control bodies, motivate people, and cure mental illnesses, among other things [25].

One of the main tasks in MIR is to develop techniques to facilitate access to various music collections [25, 60]. Furthermore, MIR approaches are useful in managing digital music repositories, such as music classification, tag prediction, music recommendation and etc. Working on these tasks often involves techniques from different fields, including machine learning, artificial intelligence, and data mining [23, 25].

1.2 Problems in MIR

The problems in MIR span several areas, ranging from social, to artistic, to computational. This study focuses on a computational problem in MIR, specifically music classification based on *genre* or *style*. Music classification is the process to categorize music pieces into different classes.

Various issues affect the performance of automatic classifiers. Some of the issues are related to cultural factors. Therefore, a number of researchers in MIR have considered the role of cultural factors in identifying the genre for a music piece. While a small number of genres and styles have a clear definition, most of them are ambiguous and inconsistent [22]. To deal with this problem, Barbedo and Lopes [4] designed and developed a *musical genre taxonomy*, which divides basic genres into subgenres, e.g., *rock* into *hard rock* and *country rock*.

Lippens et al. [29] compared the performance of human genre classification and automatic genre classification. Since genre classification and style classification are a subjective task, perfect results cannot be expected from any human or automatic classifiers. One of the main reasons is that boundaries between genres, as well as styles, are not clear. Thus, identifying one genre from another is a complex task [29]. Furthermore, there is often a significant overlap among genres; some of them are more similar than compared to others. Another issue in genre classification techniques is that not only new genres are introduced regularly, but also existing genres can change over time. Moreover, the precise definitions are not available for each genre and style [36]. These studies suggested that human classification approaches produced better accuracy when compared to automatic ones. Some researchers use correlations between particular cultural and content-based characteristics for mapping a genre to a music piece [22]. Some suggestions offered to improve music classification are as follows [22]:

- The features should be categorized into three important types: content-based low-level, content-based high-level, and cultural features.
- It should be possible to assign multiple genres to any piece. Moreover, the genred should be weighted to express some general sense of relative similarity.
- A number of realistic candidate genres should be used and organized into an ontological structure.
- Misclassification penalties in both training and testing datasets should increase the similarity among different genres.
- A music piece can belong to different genres. In addition, each segment of a music piece can be tagged by more than one genre.
- The structure of segments over time can be indicative of a genre.

- Different statistical methods, such as principle component analysis, should be used to reduce feature dimensionality.
- Having psychological, musicological, and music theoretical knowledge should be taken as a benefit by researchers in MIR.

These suggestions could help avoid confusing the classifiers and improve their accuracy. In addition, music style classification has similar issues to genre classification. Therefore, these suggestions could improve music style classification as well.

1.3 Contribution

This thesis contributes to the improvement of the accuracy of genre and style classification through data pre-processing. This is achieved via data mining techniques, such as discretization, and feature selection. Some existing feature selection techniques are used to evaluate the feature importance from different perspectives [16, 18]. Moreover, some genres or styles confuse the selected classifiers, because those genres or styles are similar. Differentiating them is a complex task. In this thesis, we analyze the classification results to discover which genres or styles are more difficult to recognize. Furthermore, we consider the performance of each classifier and each feature selection method in our experiments and attempt to find the ones that provide better classification accuracy.

1.4 Outline

The thesis is outlined as follows. In Chapter 2, we review previous works related to feature selection techniques, different classifiers, and also music classification. Furthermore, issues related to the use of music genre and style are investigated. In Chapter 3, we conduct a series of experiments to explore the effects of using different data mining techniques, such as discretization and feature selection techniques. We also consider different classifiers. In Chapter 4, we propose some approaches to reduce the confusion of

classifiers. In addition, we analyze which features could improve the accuracy of music classification based on genre or style. We use different datasets, such as *Million Song Dataset*, *Last.fm Dataset*, and *Latin Music dataset*, to evaluate the proposed approaches. In Chapter 5, we present our conclusion, with discussions on the limitation of our approach and some potential tasks for future research.

Chapter 2

Background

The work in thesis is centered on the classification of musical genres or styles in *Music Information Retrieval* (MIR). Most music classification approaches follow similar process, which involves different data mining and machine learning techniques and consists of several sequential steps.

The first step of this process is music signal processing. In this step, some work is performed on the raw data in order to obtain enough information, such as *melody*, *harmony*, *rhythm*, and *timbre*, for each piece of music to increase the precision of music classification. Preprocessing techniques, such as extracting and discretizing features, make a better dataset for the next steps of processing. Not all parts of each music piece have useful information. Thus, segmentation is another part of the preprocessing step. Audio tracks are divided into different segments. One of the important issues in the segmentation of music tracks is the length of each segment [45]. Some previous studies concluded that using small segments of music could represent sufficient information. Several software frameworks are used to extract *features*, such as MFCC. In feature extraction , feature vectors are extracted from each segment. For extracting features from a segmented piece of music, musicologists and researchers have different opinions since some particular parts of a music piece that could present much more useful information [19, 21, 47, 67].

The next step in music classification is *feature selection*. The goal of this step is to remove redundant, noisy and irrelevant information, this also helps reduce data dimensionality. From all of the features that were extracted during the first step, a subset that produces the

best classification accuracy is selected. This can be done either by trying each possible combination of features and comparing the results or via automatic approaches.
After these steps, the outcome is organized for use in higher-level processing, such as music classification. In addition, for each track of music, musicologists and users have added additional textual information called *tags*, such as *musical style*, *musical genre*, and *musical mood*, that have direct effects on the music classification performance. Moreover, most of the additional information is due to cultural factors and scenarios and also there is no specific definition for each of them.
Finally, we can classify music using the information obtained in the steps above. Classification accuracy is directly affected by what happens in each of these steps. In the following sections, we detail this process and its components and reference some of the related research work.

2.1 WEKA

Some software frameworks, such as *Waikato Environment for Knowledge Analysis* (WEKA), were implemented to provide some facilities and tools for mining, analyzing, and managing the dataset in this area to perform the sequential process [16].
WEKA is an open-source Java-based framework that is a unified benchmark for machine learning algorithms and preprocessing techniques. It includes a complete collection of methods and tools, such as classification, clustering, association, and attribute selection techniques. Its main purpose is to provide various techniques in the machine learning area. In addition, researchers could be able to compare and analyze the statistical results of different datasets and attempt to improve them [16]. In WEKA, some filters, such as discretization and feature selection methods, were implemented for preprocessing data. There are two types of filtering techniques: supervised or unsupervised methods. Filters have been used in some previous studies to perform preprocessing techniques and then classifying the dataset. For example, the work in [11] investigated automatic genre

classification, which used the WEKA for selecting features and classifying the *Traditional Malay Music* dataset [11]. In another study, researchers used different classifiers in WEKA to compare the performance of them [54].

2.2 Various Issues in MIR

Some preprocessing steps have been used to design and implement music classification techniques. However, the initial step of all approaches is to find the appropriate data to create a dataset as an input.

2.2.1 Music Data Sets

MIR systems are categorized based on the data that is used in music classification. Meta-based MIR systems use the attached data (which is named *meta-data* and contains different information regarding any type of genre, title, or author of song of a music piece) to find further information in a music collection. However, meta-data often lacks a standard format. Furthermore, some of them have uncertain information. Thus, meta-data has less potential and is less reliable. Music content, such as *melody* or *harmony* features, similarity measures between melodies, and variations of musical structures, are another type of data that is non-verbal [41]. *Content-based MIR systems* are more complex. Due to this, music must be captured and represented in a different manner. Automatically extracting musical information is becoming important as a method to construct and organize the growing amount of music [57]. Therefore, many methods have been implemented to extract the appropriate information from each piece of music.
Raw data in MIR do not have the appropriate formats for high-level tasks, such as classification. There are several methods and software available to extract useful information that can be used to create a dataset, which is used as an input. Therefore, feature extraction is one of the essential steps to obtain the appropriate information from the content of music [18].

2.2.2 Features and Feature Extraction

Feature extraction is a field of signal processing and is the process of calculating the musical content that distinguishes a piece of music from another. Different features could represent details of music from different angles. Features are related to the dimensions of music, such as *timbre* and *rhythm*. Investigating various approaches for obtaining and extracting suitable features is one of the early studies in MIR. In the previous work in MIR, the *Society for Music Perception and Cognition* (SMPC) meeting presented topics, such as *Scanning the Dial*, in MIR for the first time in 1999 [15]. Scanning the Dial means to find a program or a category of programs through different frequencies, features and recognition of music. In other words, it is able to measure human ability to classify music and provides a ground-truth to compare the different classification algorithms' performance [15].

Before Scanning the dial, Charles Seeger (1886-1979), who was a composer, musicologist and teacher, asked a question to an international group of students in the pioneering ethnomusicology program, "How long would a musical expert need to hear a song in order to recognize its culture origin?" Since then, graduate students in ethnomusicology have spent time and efforts to increase their knowledge in melodic and rhythmic music features, and also in different musical styles. They concluded that the computer could use musical details to identify music pieces. Furthermore, long segments of music are better for investigating musical contents. This study was one of the first ones in this area. Later many researchers have been continuing work in the same direction [15].

Some software frameworks, such as *MARSYAS*[1] and *jMIR*[2], were developed to extract music features. Musical Research System for Analysis and Synthesis (MARSYAS) is an open source software framework that is used in the MIR to extract musical features. In several studies [8, 11, 46, 54, 64], MARSYAS was applied into pieces of music to extract features. Another open-source software suite is jMIR that was implemented in Java to

[1] http://marsyas.info.
[2] http://jmir.sourceforge.net.

extract audio and symbolic features. It is used to extract and analysis meta-data, and also uses machine learning algorithms in MIR [33]. The purpose of these two software frameworks is to obtain features from audio, which recognize contents of each segment of music without using any tags, such as the name of artist, or meta-data.
Musicologists and researchers attempt to discover not only which parts of music could present enough information, but also which set of features could help the classifier to group music pieces correctly. Many studies showed that extracting and selecting the proper features are important tasks in music classification. Most of the previous research indicated that creating a new dataset based on a combination of different feature sets would give better outputs. Some features cannot represent all aspects of music. Therefore a combined feature set provides an opportunity to use more details of a piece of music from different aspects. However, not all features might be useful to classify all music. Therefore, classifying large number of feature is a complicated approach [9, 15, 32].
In this study, various benchmarks are used to create datasets. These benchmarks, such as *Million Song Datasets (MSD)* and *Last.fm*, were used to perform the validation experiments of our proposed approaches. They were made based on different features, which were grouped into *timbral* and *rhythm histogram* features.

Timbral Features

Timbral features could represent the proper details of music segments to recognize musical speech. The *Short-Time Fourier Transform* (STFT) is to compute timbral features for each short segment. Timbre makes two sounds different from another one, even when they have the same loudness and pitch. This kind of feature includes *Spectral Centroid*, *Spectral Rolloff*, *Spectral Flux*, *Mel-Frequency Cepstral Coefficients (MFCCs)*, *Analysis and Texture Window*. In other words, “timbre features describe the spectral characteristics in a given analysis window, whereas spectro-temporal features characterize the temporal evaluation and variety of the timbre features in an analysis window” [28]. Timbre can be

summarized by low-level statistics of the descriptors [28].

Rhythm Content Feature

Rhythm is the timing of events, which happen periodically. Periodicity functions are used to measure the range of tempo, which characterizes the rhythm of a piece of music [48, 57]. In other words, rhythm features are computed based on the relations between the main beat and sub-beats.

In addition, there are also some other types of features, which are used to represent the content-based of different music pieces [9, 28, 32, 59, 66].

Some Feature Categorization

Features are categorized based on different criteria. Some researchers group them based on time, including *short-*, *medium-* and *long-time features*. For example, Gjerdingen and Perrott's study showed that the identification of genre relies on short-term part of music [3]. Moreover, some of the group features are based on content, which groups them into *low-level* and *high-level features* [7]. McKay and Fujinaga [35] categorized features into three groups: *low-level, high-level*, and *cultural*. They defined the *low-level features* that are extracted directly from audio signals, such as spectral or time-domain information. The *low-level features* do not seem to have musical information. The *high-level features* are understandable by musicologists. These features include instruments present, rhythmic density and chord frequencies. *Cultural features* are sociocultural information, which are not based on music contents. In addition, some researchers have used an individual feature set or a multiple feature set to create a dataset. There are various approaches to extract details of music [35].

In another study, Flexer et al. [14] considered the combination of some selected music features to classify ballroom dance music. They combined two types of features: *Spectral* and *Rhythmic features*. In this study, MFCCs were used to represent the spectrum of each music piece. Moreover, rhythmic similarity was a kind of tempo feature that was one of

the fundamental features and yielded a huge amount of information about various types of dance music [14].

Lopes et al. [30] and Shawe-Taylor and Meng [50] used different *short-time features*, such as MFCCs, to classify music [50]. Doraisamy et al. [11] extracted three features: *Short-Time Fourier Transform* (STFT), MFCCs, and *Beat* to classify *Traditional Malay Music* (TMM). In another study, the selected features include *beat spectrum, linear predictive coding* (LPC), *zero crossing rates, spectrum power*, and also MFCCs, which were related to the temporal, spectral, and cepstral domains in music [61]. Wang et al. [59] extracted MFCC, STFT, and also *Daubechies Wavelet Coefficient Histograms* (DWCH) in the preprocessing step of their proposed approach.

According to the study in [2], MFCC and $\Delta MFCC$ were extracted, which were used as numerical representations to model timbre of audio segments. The music classification accuracies of using MFCC and $\Delta MFCC$ with different weights were better than those using just MFCC. The results in Gjerdingen and Perrott [15] indicated that short segmentations did not have enough information. Thus, relying on small parts of music was not reliable.

In the proposed approach by Meng et al. [40], the short-time, medium-time and also long-time features were combined to investigate the best ones for classifying the selected pieces of music. Most of the previous studies used different short-time features, such as MFCC, to improve the performance of music classification. Therefore, MFCCs features were used as *short-time features*. Some feature integration techniques in medium-time scale extract temporal information, such as *Mean and Variance* of the MFCCs, *Filterbank Coefficient*, *Autoregressive* model, *High Zero-crossing Rate Ratio*, and *Low Short-Time Energy ratio*. Beat and vocal were used to extract large time scale features in this study. Furthermore, *beat spectrum* (BS) and *beat histogram* (BH) were other methods to calculate the large time-scale features. Various combinations of feature sets were evaluated to gain the high music classification accuracy. The combination of the short and medium

features, AR and MFCC features, gave a better performance compared to others [40]. Cataltepe et al. [8] used various multiple feature sets, and made a new feature set to obtain better results. MARSYAS framework was used to extract features that were spectral centroid, such as MFCCs. MFCC features obtained very close results, which was not an obvious difference between the performance of using all feature sets and MFCC. However, the results showed that MPITCH and BEAT were unsuitable features, and also the performance of music classification was not improved by using these features.

To evaluate the performance of different proposed feature sets, Tzanetakis and Cook [57] chose timbral texture, rhythmic content, and pitch content. Furthermore, in another study [24], multi-dimensional features of MIDI files were extracted by using JSymbolic software. These features belong to some feature sets, such as instrumentation, texture, rhythm, dynamics, pitch statistics, melody, and chords. The results of these two studies indicated that combining different feature sets could make an improvement in the performance of classifiers. In the proposed approach by Lim et al. [28], the music genre classification used different feature sets. In the testing phase, timbre features were used to classify training dataset. Then, in the training phase, timbre and spectra temporal features were chosen to classify the given dataset.

Some former studies analyzed the effects of using different levels of features on the performance of some proposed music classification techniques. McKay and Fujinaga [37, 38] used separate musical information to extract features, such as *audio* (A), *symbolic* (S), and *cultural* (C) sources. The proposed approach considered not only the effects of feature sets separately, but also the effects of the combination of them on the classifier performance. The results of this experiment compared the average of classification accuracies of three feature types, which were just one type of the feature set (S, A, and C), two types of feature sets (SA, AC and SC), and the combination of all feature sets (SAC). In [37], the results showed the average accuracy of different experiments using SAC was higher than the one from the other feature sets.

The results of the study in [38] showed that combining two feature types (SA, AC and SC) was better than any single feature type. Moreover, combining all three feature types did not give considerable improvements in the performance of classifiers when compared to using the combination of two feature types. In another experiment by McKay and Fujinaga [38], the results indicated that the *low-level features* represent appropriate information to classify the dataset. Symbolic data was used to complete the low-level information that was categorized into high-level features. However, extracting high-level features from audio data is a complex task. As a result, the improvement could be made by using both audio data and symbolic data types, together low-level and high-level information [38].
In [34], McKay and Fujinaga used different techniques to classify the musical genres based on different levels of hierarchical features. Some studies indicated that high-level features conducted better classification performance compared to low-level features. In this experiment, after extracting features, 109 features were selected based on low and high-level features to characterize and classify recordings, which consisted of seven categories, including *Instrumentation, Musical Texture, Rhythm, Dynamics, Pitch Statistics, Melody, and Chords*. Genetic algorithms were used to select features. The results indicated that using a hierarchical feature set improved the performance of classification compared to using a flat feature set. Each classifier used some particular features to classify the main genres, and also sub-genres.
Lidy et al. [26] attempted to improve the performance of music classification by combining two approaches, which were designed based on audio and symbolic music analysis and retrieval techniques. To extract audio features, different techniques were used, such as *Rhythm Pattern, Rhythm Histograms, Statistical Spectrum Descriptors, and Onset detection*. Moreover, symbolic features were extracted from audio files to complete the audio features. By comparing the results of using individual feature sets, the *Rhythm Pattern* set resulted in higher accuracy. In the second experiment, the multiple feature set got considerably more precise outcomes [26].

Various earlier studies have used different approaches to identify the best feature sets. As mentioned above, the results indicate that using a multiple feature set causes positive effects on the performance of music classification. However, choosing the appropriate features for classifying the dataset is one of the significant challenges in MIR studies [1, 11, 28]. In addition, a large number of features has some disadvantages, such as increasing running time, and also confusing classifiers. How to provide the best set for every approach is thus one of the important subjects in MIR. One approach is to use a feature selection technique after combining the appropriate feature sets. However, several feature selection techniques are available, and choosing the best one is another problem in MIR.

2.2.3 Benchmarks in MIR

One of the important issues in MIR is the ability to compare the outcomes. Therefore, proper benchmarks could help researchers create the dataset. In addition, they can share and compare the outcomes of their approaches. There are different benchmarks available that contain different extracted features. Thus, most researchers in MIR have attempted to use the appropriate ones. The size of benchmarks is another significant consideration, and it has to be close to the amount of data in the real world. Most datasets do not have enough data. Thus, some companies attempt to create a huge dataset for solving this issue in MIR area[3]. Different datasets are available for research proposes. In this study, *Million Song Dataset* (MSD)[4], *Latin Music Database*[5], and *Last.fm*[6] are selected to create the dataset. In the following sections, these datasets will be discussed.

[3] http://labrosa.ee.columbia.edu/millionsong/.
[4] http://www.ifs.tuwien.ac.at/mir/msd/.
[5] http://www.ppgia.pucpr.br/ silla/lmd/.
[6] http://labrosa.ee.columbia.edu/millionsong/lastfm.

Million Song Dataset (MSD)

The MSD Dataset is a large-scaled dataset that contains audio features for a million popular pieces of music: metadata such as song names, artists and albums and audio analysis. The Echo Nest services extracted the features of the music, such as loudness, tempo, and MFCC-like features from MSD. The wide collection of well-known features in the MIR domain is available, as well as ground truth data with a set of training/testing splits datasets. The purposes of creating MSD are not only to use a large dataset, which is scaled to commercial sizes, but also to create a reference dataset for comparing the results of studies [6]. Various feature sets were extracted by jAudio feature extraction software, the MARSYAS feature extractor, and the Rhythm Patterns family of feature sets. The jAudio software extracted features based on the frequency and time domains, which include MFCCs, low-level spectral features, and method moments [6]. The MARSAYS was used to create three benchmarks: MFCCs, Chroma, and also timbre in MSD. The spectral representation was used to extract the Rhythm patterns and related feature sets. All of these datasets are available in the *MSD* [7] [27].

Latin Music Database (LMD)

The *Latin Music Database* contains 3,160 music samples of 10 different genre types including *Bolero, Forro, Gaucha, Merengue, Pagode, Salsa, Sertaneja*, and *Tango*. The samples were labeled by a group of human experts; therefore, the quality of the dataset is high, compared to the datasets that were labeled by amateur users. The MARSYAS[8] framework was used to extract features. In the feature extraction, the first, middle, and end portions of each piece of music were used. Therefore, LMD has three datasets; each of them contains extracted features from a particular segment (first, middle, and end segment) of all music. Moreover, training, validation, and testing datasets contained 300 different music samples that were chosen randomly from each selected genre [30, 54].

[7] http://labrosa.ee.columbia.edu/millionsong/
[8] http://marsyas.info.

Last.fm Dataset

Last.fm[9] users labeled a piece of music based on various tags, such as artist name, name of album, and mood of music. It contains 943,347 tracks, matched to MSD. The dataset includes several tags for each track, and some of these tags are not acceptable. In this thesis, the tags of 10 top genres and styles are used to evaluate the performance of the proposed approaches [39].

2.2.4 Data Transformation

The format of music data may require some transformations before being applied in any music classification technique. Therefore, some methods may apply to consolidate pieces of music into the proper format, which includes smoothing, generalization, aggregation, normalization, and discretization. In other words, some preprocessing techniques are applied to prepare the dataset. In this study, we use discretization and feature selection techniques.

2.2.5 Discretization and Feature Selection

Nominal attributes could present some issues for classifying and clustering algorithms. Some of these algorithms could not handle numerical attributes. Thus, discretization techniques are used to categorize attributes into a number of distinct ranges. Furthermore, some learning algorithms could handle the numeric attributes. But generally discretization could provide better results and also decrease the running time. In other words, data discretization is a form of reducing data: it divides the numerous data into different ranges, which are tagged by interval labels. Therefore, the interval labels can be used as actual values. The discretization algorithms are categorized into two types: supervised and unsupervised. The supervised algorithms take the class of each attribute to discretize all continuous values. On the other hand, the classes in the unsupervised algorithms are unknown, and these algorithms categorize the features without using the classes of them

[9] http://www.last.fm.

that are useful for dealing with clustering problems [12, 31].

In this thesis, the datasets are discretized by the supervised and unsupervised methods ini WEKA to group feature values into distinct ranges [18]. WEKA uses either Fayyad & Irani's MDL method or Kononeko's MDL criterion to implement the supervised discretization technique. Furthermore, the unsupervised discretization technique in WEKA uses the simple binning [16].

2.2.6 Feature Selection

As discussed above, using a combination of two or more feature sets resulted in better music classification accuracy compared to using only one. However, most of the datasets, which contain a combination of different features, have a huge number of features. Not only could some features give very little useful information, but they could also confuse the classifier and increase the classification running time. Therefore, feature selection is one of the important steps in music classification that has significant effects on the performance of classification. The main rule of all feature selection techniques is to remove inappropriate features, which do not represent enough information for music, and select or rank the appropriate ones.

There are two main generic methods to select features: the *wrapper* and the *filter* methods. Each of them uses different search methods to choose the proper set of features. The *wrapper* method, such as *Information Gain* (IG), uses a ranker as a search method, which ranks features by some individual evaluations such as Gain Ratio and entropy [28]. It also uses a learning method to separate useful features from redundant and irrelevant ones. The *filter* method uses a learning method to evaluate the features, and then selects the features with the high ranks. For example, the *Correlation Feature Selection* (CFS) with search strategy is the filter method. The huge dataset with irrelevant features has effects on the performance of the filter method. Therefore, the *wrapper* method is known to outperform the *filter* method; however, it is more expensive and slower. The main goal of feature

selection techniques is to select the best features [25, 60].

Many methods are implemented in some software like WEKA that can be used to select features. Each feature selection technique has some advantages and disadvantages. Typically, feature selection techniques are used to analyze the correlation between distinguish-ability of the instances and classes. For example, Lopes et al. [30] used the selection of the best features and the selection of random features. They were performed on a subset of LMD. The best performance of the approach in of Doraisamy et al. [11] was obtained using *Multilayer Perceptron* (MLP) as a classifier, which incorporated the CFS and a genetic search strategy. Also, feature selection caused a significant improvement in the *Artificial Immune Recognition System* (AIRS) accuracy. As a result, the accuracy of classifiers, such as MLP, *Sequential Minimal Optimization* (SMO), and AIRS, in this study was improved considerably by using feature selection technique.

Ariyaratne et al. [1] investigated another method. Features are selected for each class separately. The two main techniques split multi-class into a small collection of binary problems, which are One-Vs.-One and One-Vs.-All. For the first approach, the classes were grouped by two, and then features were selected for classifying each group separately. The classifier attempted to classify one class against all classes. As a result, the first technique provided a better performance compared to the other [1]. Additionally, forward and backward selection techniques were used as feature selection methods in [64]. The results of using these techniques concluded that there are considerable improvements in the music classification performance.

Lim et al. [28] attempted to make improvements in the *Support Vector Machines* (SVM) accuracy by using a linear SVM and feature selection technique. The proposed music genre classification used in two databases, GTZAN (100 songs) and ISMIR2004 (729 songs) databases. GTZAN has 10 genres, including *Blues, Classic, Country, Disco, Hip-Hop, Jazz, Metal, Pop, Reggae*, and *Rock*. The 10-fold cross validation was performed for the first one. ISMIR2004 was used for training and testing procedures, which contains

6 genres (*Classical, Electronic, Jazz/Blues, Metal/Punk, Rock/Pop*, and *world songs*). The results indicated that SVM classifier returns better results for higher dimensionality of the feature vectors compared to other feature selection techniques [28].
In this study, a variety of feature selection strategies, including *Information Gain* (IG), *Gain Ratio* (GR), *Symmetric Uncertainly* (Sym), and *Chi-squared* (Chi) were chosen to evaluate the performance of the given classifiers. Moreover, *Correlation-based Feature Selection* (CFS) is chosen as a filter-based feature selection technique that selects a subset of features, which is highly correlated with different classes.

Correlation-based Feature Selection

Many previous studies have employed *Correlation-based Feature Selection* (CFS) methods to obtain the proper subset of features, which has a high correlation between the selected features and the target concept. CFS uses the *Best-First Search method*, which is a heuristic search method. It creates all feasible feature subsets from an empty set until the best one is found. Each possible subset is evaluated, and then compared with preceding results. It will proceed sequentially to obtain the best improvement [17]. Furthermore, in the proposed approach by Kofod and Ortiz-Arroyo [24] and Doraisamy et al. [11], CFS was used to reduce dimensionality and choose the relevant features. These studies concluded that CFS could increase considerably the accuracy of classifiers.

Information Gain

Another popular method is *Information Gain* (IG) that is used as a feature selection technique frequently in machine learning. Information theory is calculated to evaluate the significance of each feature. In other words, “IG uses the number of bits of information obtained for category prediction by knowing the presence or absence of a term in a data” [63].

Gain Ratio

Gain Ratio (GR) is a feature selection technique, which is a modification of IG. The GR uses the ratio between the IG and the intrinsic value. The GR "takes number and size of branches into account when choosing an attribute" [20, 44]. It corrects the information gain by taking into account the *intrinsic information* of a split into account, which is an entropy of distribution of instances into branches (i.e. how much information do we need to tell which branch an instance belongs to) [20, 44].

Symmetrical Uncertainty

One of the best measures for selecting features is *Symmetrical Uncertainty* (Sym). This method analyses the correlation between the features and the classes based on two aspects: analyzing whether the feature is pertinent with the class or not, and reducing the feature, whether is pertinent with the other selected features or not [65].

Chi-Squared

The *Chi-Squared* (Chi) method analyzes features individually based on their chi-squared statistics. The feature with higher Chi is more important than the others with the lower ones. This feature selection technique evaluates the dependency of each feature and the class for selecting accepted features [65].

Issues Related to Feature Selection

To represent raw data, we need to use more features. However, some of the features could not represent enough details for music pieces. Redundant features have harmful effects on increasing the classification running time, and also the predictive accuracy of classification. In the real world, this situation is one of the problems in machine learning. Thus, feature selection techniques are used to select a small set of features, which are sufficient to represent enough information for describing the proposed concept [25, 60]. However, there are several feature selection techniques. Choosing the proper one is an

important task in music classification.

The results of the former studies used different techniques to obtain an acceptable classification performance [25, 60]. Therefore, we will discuss the effects of using some of them on the music classification accuracy in the following sentence.

2.3 Music Classification

In music classification, choosing the appropriate classifier(s) is one of the important steps. Various classifiers are available that were implemented based on different machine learning techniques. The previous works show that different classifiers have been used to obtain the optimal performance for each classifier. The accuracy of each classifier represents its ability to predict the class of a music piece. Various sets of experiments have been used to indicate which algorithms and parameters are best suitable for music classification. In some works, various classifiers were chosen to evaluate the performance of classification algorithms. For example, the performance of the studies in Soares et al. [56], Doraisamy et al. [11] and Cataltepe et al. [8], which used different supervised and unsupervised classifiers, is compared separately. Additionally, some studies used multi-layer classifiers. Xu et al. [61] improved the classification accuracy by using different supervised classifiers in each layer of the proposed approaches. The *Multi-Layer Support Vector Machines* (SVMs) classifier made an enhancement in the performance of the classification.

2.3.1 Classifiers

The machine learning approaches, unsupervised and supervised, have garnered increasing interest. The unsupervised approach is to cluster data based on objective similarity measures. The unsupervised classifiers are two types: time invariant, such as measuring distance between two features, and time variant like hidden Markov-models. The simplest and most popular clustering algorithm is k-means. The Other machine is the supervised

approach that takes a known set of input and output data, and seeks to create a predictor model. The predictor model is used to generate reasonable predications for the new outputs that are unlabeled data. Different supervised learning algorithms are used for music classification, such as *Decision Tree* (DT), *Support Vector Machines* (SVMs) and *Artificial Neural Networks* (ANNs) [16, 18, 48].

Previous studies indicate that the performance of some classifiers is better than others. In this study, five classifiers have been chosen to compare classification performance, and to investigate impacts of using the feature selection techniques. The selected classifiers include the following: *Sequential Minimal Optimizations, Decision Tree, Neural Network, Naive Bayes*, and *Bayes Net*. In the following, we discuss some results related to them.

Sequential Minimal Optimizations

Support Vector Machines (SVMs) are supervised learning methods that use statistical learning theory to classify a dataset. The SVMs are used in a N-dimensional hyperplane to separate data into groups. According to the study in [16], in SVM, "a predicator is called an attribute, a transformed attribute is called a feature, and a set of features is called vector". The *Sequential Minimal Optimizations* algorithm is used to train an SVM classifier [16]. It has been successfully used in music genre classification. Therefore, various studies have chosen SMO as a classifier, such as the works in [28, 50, 54, 62].

Xu et al. [62] investigated the pure and vocal music classification, music characterization, and SVM learning. Different features were extracted in pure and vocal music. The SVM method was used training dataset, and then several methods, such as SVM, traditional Euclidean distance method, and *Hidden Markov Model* (HMM), were used for testing the dataset. Individual users evaluated the music summarization results. It appeared that the SVM learning method increased the accuracy considerably, when compared to HMM methods and traditional Euclidean distance methods [62]. The results of experiments in [28] indicated that the SVM classifier obtained better accuracy, when compared to other

classifiers.

Decision Tree

A popular machine learning method, *Decision Tree* (DT), is a supervised classification learning method. The purpose of this method is to find the model of input that predicts a target variable (class). Most of the DT, such as *Iterative Dichotomiser* (ID3), *Classification and Regression Trees* (CART) and C4.5 (a version of ID3), have a top-down structure. A top-down approach starts with a training dataset that is subdivided into smaller subsets iteratively as the tree is being built. The difference between various DT methods is based on two important aspects: how to choose the attributes of a training set and how to prune the tree. J48 is a decision tree classifier that is implemented by using popular decision tree algorithms, such as C4.5 algorithm and Quinlan algorithm [5].
DT is one of the classifiers, which has been used in several experiments. Generally, its performance is acceptable. However, the DT classifier spends considerable time to classify subjects. In other words, the running time of classifying huge datasets is inefficient. For example, the performance of DT was investigated in several studies [24, 43, 54]. The outcomes by Norowi et al. [43] concluded that different parameters, such as data size, the length of music piece, and the number of cross validate folds, have considerable effects on the classification performance. Consequently, the proposed approach increased the accuracy of DT by upgrading each parameter.

Naive Bayes

Naive Bayes (NB) is a simple Bayesian classifier that uses a statistical analysis of a training dataset to create maximum likelihood estimators and conditional probabilities based on given features and classes [5]. NB can predict the probability of belonging attributes into a specific class. This means that the effects of each attribute value on the class are independent of the other attribute values. Furthermore, NB obtains sufficient accuracy when the given dataset is huge. In the previous study [54], NB was not good

enough to classify subjects with high accuracy when compared to others. The performance of NB was compared with DT in a study by Kofod and Ortiz-Arroyo [24]. In this study, the selected dataset were classified more accurately by using NB.

Neural Networks

The *Neural Networks* (NNs) learning algorithm is based on knowledge from the psychology and neurobiology area. The *Multilayer Perceptron* (MLP) classifier is implemented based on the feedforward artificial neural network. The structure of MLP contains different input/output units, which are connected together, and each connection has a weight. Moreover, each unit has several nodes, which are called *neurons* or *processing elements*. Back-propagation is used as a supervised learning algorithm to train the network [7]. In the initial step, the weights of connected links are assigned. After processing inputs, the produced outputs are compared to the anticipated results. Then the weights are changed based on the amount of errors. This process continues until the classifier obtains the expected results. NN classifiers need much more time to create a predication model from a dataset compared to most of the classifiers. In other words, the classifier spends much more time for training because it needs to create the structure that allows it to generalize to new instances. However, it obtains the acceptable accuracy when it classifies a noisy dataset. The approach proposed by McKay and Fujinaga [34], Shawe-Taylor and Meng [50] resulted in more precise outputs.

Bayes Net

Bayes Net (BN) is another type of Bayesian classifier that uses dependency among attributes. In BN classifiers, each attribute is investigated separately, but it uses “class conditional independences” on the subset of attributes. Different algorithms are used to train the networks, such as Adaptive Probabilistic Networks. The trained BN is utilized for music classification [18].

2.3.2 Issues Related to Classifiers in Music

Some of the factors, such as similar styles and genres, and unbalanced datasets, complicate a classifier's ability to group music pieces correctly. In other words, some music pieces are not recognized correctly, i.e., they are tagged as other classes. The problem becomes worse when the dataset is not balanced, and the number of music pieces from some classes is more than the ones in others. Therefore, the training dataset may not contain enough information about all music pieces from all classes.

For example, a dataset contains three classes. These classes are *Pop*, *Jazz*, and *Rock*, which are 1000, 850, and 50 music pieces, respectively. It is important to note that the numbers of samples from each of these classes are not the same in the training set. Thus, the classifier trains more according to *Pop* and *Jazz*, which are more available. The trained classifier cannot identify music pieces from the *Rock* class. The confusion matrix indicates that most of the music pieces of *Rock* class are tagged as *Pop* or *Jazz*. In this situation, the accuracy of the classifier is high, because the number of music pieces that are classified incorrectly is low (the number of music pieces from *Rock* class). Thus, they do not have considerable effects on the accuracy. The work in this thesis attempts to discover the proper method to address these issues that confuse classifiers.

2.3.3 Genre and Style Classification in Music

Various musical genres and styles are used to categorize music in MIR. However, there is a difference between *genre* and *style*. *Genre* is a kind of comprehensive subject, when compared to *style*, which is a content-based aspect. Regarding these terms, music genre and music style have different definitions. Musicologists, researchers, and common people tag pieces of music with different labels as musical genres (or styles) that characterize the music. A *genre* is used to categorize similar pieces of music based on different aspects, such as music composition, rhythm patterns, and similar pitch distributions. *Style* refers more to the rhythm, harmony, melody, arrangement and production of tracks that might be

associated with music of a particular type from a specific area or of a specific genre. *Genre* and *style* are used as descriptors in managing, searching, and also comparing huge amount of music in classification that are usually related to different properties of music, such as instrumentation, the structure of rhythm, and also harmonic content. They can provide valuable information to model the relation between acoustic features and music content. A genre should be defined as pieces of music that share a certain style [10]. Typically, music pieces that are categorized into a type of genre (or style) have some similarities. Moreover, "the degree of commonality is stronger within a type of style than within its enclosing genre" [58].

Common people or musical experts annotate musical genres and styles manually. As examined from earlier studies, culture has significant effects on choosing a genre and style for a music piece. People, who have a similar cultural background, tag music with similar genres or styles. These characteristics have created a large area of study in MIR: genre classification and style classification. In other words, one of the purposes of MIR is to discover them automatically that are helpful for designing content-based MIR systems. The main purpose of many previous studies was to develop a proper setup. It is not easy for computers to analyze and classify composites by considering the relation between musical features, including pitch and rhythm. However, humans can identify different styles and genres in real time. Thus, some studies investigated the manual music classification to identify approaches that have been used by human to classify the music. There are two schemes to design the music classification techniques. The first scheme is based on the *flat music classification*. All music pieces from several classes are classified at the same level. On the other hand, some studies have used a top-down approach to *hierarchical music classification*. In these studies, the subsets of classes are discriminated at the top level of the hierarchy, and then each set of classes are classified separately [13]. In this study, we use both schemes in our experiments.

Flat Music Classification

There have been a few studies on evaluating human genre-tagging behavior. Some studies deal with human-involved evaluation of music genre classification [3, 15]. Aucouturier and Pampalk [3] used short-term features to show audio pattern recognition based on human pattern recognitions. Thus, the proposed approach for selecting the segment of music was based on two aspects. The first aspect of the proposed approach was to select the proper segments that use to identify the genre of music manually. The second one was to select segments of music, which had sufficient amount of information and used as a ground truth to validate technical result. However, some researchers believed that it was not acceptable, because human and algorithms were not tested in the same database. The results showed that tagging by users is an unstructured, un-moderated process and annotated manually (no automatic analysis). The main reason was that tagging music was subjective, which was based on the users' knowledge. Users therefore may have different opinions about the same music [3].

In another study, Shawe-Taylor and Meng [50] investigated human and automatic music classification. The selected genres (*Alternative, Country, Easy Listening Electronica, Jazz, Latin, Pop&Dance, Rap&Hip-Hop, R&B and Soul, Reggae* and *Rock*) were classified by *Support Vector Classifier* (SVC) and *Linear Neural Network (LNN)* classifier, as well as humans. The results of experiments showed that the human music classification performance was better than the automatic music genre classification [50].

Additionally, the main purpose of the study in Gjerdingen and Perrott [15] was to investigate music genre recognition by human. Different lengths of music were studied to consider which size of music segments could represent enough information. Thus, the selected tracks were divided to five segments (250ms, 325ms, 400ms, 475ms, and 3000ms.). They chose ten main genres (*Blues, Classical, Country, Dance, Jazz, Latin, Pop, R&B, Rap*, and *Rock*). In the experiment, 52 participates participated to recognize selected tracks. Some of them did not have experience in music. The results indicated that

short segments did not have enough information. Moreover, relying on small parts of music was not reliable, because it did not contain important content for recognizing the music pieces. Possible effects of this factor were that gender has an influence on music selection. However, it did not have an effect on recognition. Another possibility was that gender affects sub-genre selection but it does not have an effect on choosing the main genre. Furthermore, the participants' ages had an effect in selecting genre. The last experiment showed that the older participants recognized music genres better than the younger ones [15]. Their results showed that there is a certain degree of subjectivity in genre annotation by human and automatic classification.

Various experiments in this area stated that small pieces of music could not have enough information to recognize the type of music. In addition, humans could classify the similar genres (or styles) better than automatic music classification techniques. Therefore researchers attempt to find a proper approach for designing the automatic music classification to solve all of the issues. It is obvious that various aspects have effects on automatic music classification. Thus, different technologies have been used to obtain the best accuracy. The work in Aryafar et al. [2] studied the effects of different parameters, such as the number of Euclidean dimensions, the filter size, and distance weights, on the classification accuracy.

Norowi et al. [43] investigated the effects of different parameters, such as dataset size, dataset track length, dataset starting point, number of cross-validation folds, and classifiers on the performance of music genre classification. They used a collection of *Malay* songs that was collected from the Internet and Audio Compact Discs. Moreover, the Marsyas framework software was used to extract three features, timbal texture, rhythmic content, and pitch content for training and testing the DT (J48) as a classifier. Five different experiments used modified datasets based on each parameter. The first experiment used 10 songs per genre and 30 songs per genre for comparing the effects of different dataset sizes on the performance. Some genres did not have enough number of songs, and so were

removed. The results indicated that 30 songs per genres had a better accuracy. Thus, the suitable number of songs per genre improved a performance. In the second experiment, three different lengths of tracks of 10 second, 30 second, and 60 second, were tested. Interestingly, 10 second tracks have a better performance compared to others. In addition, using the middle part of each track in the third experiment increased the accuracy compared to using the first part of the track. The fourth experiment represented the results of modifying cross-validation rate. Different numbers of cross-validation folds were used, which were between 3 and 6. The results showed that increasing the number of cross-validation folds more than 10 did not improve the performance. Furthermore, the best number of cross-validation folds was 6 folds in this experiment, but relying on 10 folds was more acceptable in all classification problems. In the last experiment, the performance of DT classifier was more acceptable compared to the OneR classifier performance. As a result, choosing a proper classifier was another important issue to increase the performance of music genre classification.

Kofod and Ortiz-Arroyo [24] studied the preprocessing techniques to improve the performance of classification and to reduce runtime complexity. The proposed approach consists of 5 main phases: Feature Extraction, Multi-dimensional Feature Conversion, Feature Selection, Feature Discretization, and classification. The first three phases were preprocessing steps. They provided an improvement by selecting proper features. The datasets chosen included 9 genres: *Bebop, jazz-soul, swing, rap, punk, country, baroque, modern-classical*, and *romantic-classical*. They used different base classifiers and found that the NB base classifier gave better results compared to others. Moreover, the proposed approach has considerable effects to decrease the runtime complexity. Their experiments show that the approach is “capable of outperforming other more example ensembles of classifiers”.

Hierarchical Music Classification

Some music classification techniques classify music data in several steps. These techniques have been designed based on the hierarchical structure. Hierarchy is an approach for creating structure and managing genres and styles. Some musicologists and researchers used the existing structures, which were created from the previous studies. On the other hand, some of them attempted to build a new structure based on their results. However, there is no approach for creating a hierarchical tree [51]. Many of the past works indicate that the hierarchical music classification improves the classification accuracy compared to the flat music classification. For example, one approach proposed to choose the proper classifier from the classifier set for each subset of genres or styles [51]. In another study, the multi-layer classifiers were used that contained three layers to discriminate musical genres. The first layer used the features of beat spectrum and LPC-derived cepstrum to recognize *Pop/Classic* and *Rock/Jazz* in the trained music samples. In the second and the third layer, not only *Pop/Classic* music was classified as *Pop* and *Classic* music, but also *Rock/Jazz* music was classified as *Rock* and *Jazz* classes separately. The features of LPC-derived spectrum, spectrum power and MFCCs were used in the second layer. In the last layer, the feature set consisted of zero crossing rates and MFCCs. For this approach, multi-layer SVMs were more accurate compared to traditional Euclidean distance based method and other statistic learning methods. However, multi-layer SVMs took a long time to train datasets [61].

Tzanetakis and Cook [57] compared the performance of music classification among different classifiers and musical genre datasets. They explored automatic music classification based on the hierarchical genres. Individuals, who participated in the experiment, identified the hierarchical genres. Their opinions had important effects on the hierarchical structures. The *Standard Pattern Recognition* classifiers were chosen to evaluate the proposed feature sets and to classify a real world datasets into genres, which were *classical, country, disco, hip-hop, jazz, rock, blues, reggae, pop*, and also *metal*. The

confusion matrices obtained from the first experiment with *jazz* and *classical* genres showed the misclassification, which had an effect on the classification performance. In the second experiment, the impact of changing the size of texture window in the classification parameters was considered. The best results were obtained when 40 windows was the best size of the texture window size. In the last experiment, They investigated the performance of using of the feature sets combined together against specific feature sets for each genre. The results showed that using specific feature sets gave better results than the set of the combined ones. The results of the method and the previous study, which were based on classification by humans, showed that the accuracy was not significantly different.
McKay and Fujinaga [34] used some techniques to classify music genres based on various levels of feature hierarchy. They chose several digital formats to extract high-level features. After feature extraction, they selected 109 features to characterize and classify recordings. *Genetic Algorithms* were used to select features, which were categorized based on feature dimension, one-dimensional features and multi-dimensional features. Furthermore, the two classifiers, the NN and k-NN, were chosen as classifiers of the subjects. To evaluate the performance, the experiment was repeated three times and was used to study the effects of randomly selecting a subset of extracted features. Using more features increased the classification performance. In addition, a second experiment showed that the selected features, the selected genres and the selected classifiers had significant effects on automatic music genre classification. Moreover, the number of features and the number of genres affected the classification accuracy significantly [34].
It is obvious that various means are available to make enhancement in music classification. In this thesis, some of these ways are studied in order to obtain a better accuracy, using different feature selection techniques, various classifiers, and datasets. Furthermore, designing hierarchical structures for each selected dataset and classifying based on the new structure is also another approach.

Chapter 3

Analyzing Genre and Style Classification in Music

Previous studies on music classification aimed to improve its accuracy. A variety of issues, such as the size of dataset and the number of features, have considerable effects on music classification accuracy. However, there are no particular collective discussions on these issues. Moreover, while different techniques are presented, there is no specific way to guarantee which technologies are better than the others. In this chapter, different empirical experiments are conducted to show the effects of various aspects on music classification. These experiments were implemented in Java and used WEKA Machine Learning Toolkit methods (Version 3.6.10) [16], as discussed in Chapter 2. WEKA is one of the best data-mining platforms and contains various techniques that are useful in MIR.

This chapter analyzes and discusses different experiments, including the effects of using balanced and unbalanced datasets, individual feature set and multiple ones, supervised and unsupervised discretization techniques, different feature selection techniques, and also some selected classifiers. Furthermore, dataset size is another parameter that has an effect on the performance of music classification. In addition, some classifiers were investigated to indicate their music classification performance. Some feature selection techniques were also chosen to evaluate datasets. The main focus of these experiments is to evaluate the proposed approaches by classifying music genre or style. We also create separate datasets based on genre and style to compare the results of classification.

In this chapter, different experiments were investigated to find the appropriate setup for the

main proposed approaches that will be represented in Chapter 4. Therefore, the low accuracy of genre and style classification of some experiments is not important in this chapter.

The chapter is organized as follows. Sections 3.1 and 3.2 introduce how to create a dataset for evaluating music classification, including dataset preparation and feature benchmarks. Section 3.3 presents several experiments that are designed to improve the initial experiment setup (These will be used to evaluate the main approaches in Chapter 4). This section also contains four subsections, which explain the results of each experiment completely. Finally, Section 3.4 discusses some problems of the experiments. Generally, after each section, the experiments are analyzed to find a way to refine the results.

3.1 The DataSet

One of the important issues in *Music Information Retrieval* (MIR) is to choose and design a suitable dataset, which has the appropriate size and also contains well-organized music pieces. Various datasets, such as *Million Song Dataset* (MSD) and the *Latin* dataset, are investigated in the perceptual studies in MIR [6, 21, 30, 49]. MSD datasets have been created based on one or more of the Million Song benchmarks, which used different methods from signal processing to model semantic information extracted from music [49]. There are various feature sets that were used in the various tasks. As mentioned in Chapter 2, many studies have been conducted using a combination of feature sets, which provide more useful information than when using individual feature sets. Therefore, five benchmarks of MSD, *Low-Level features*, *MFCC features*, *Rhythm Histogram feature*, *MSD Allmusic Top Genre Dataset (Top-MAGD)*, and *MSD Allmusic Style Dataset (MASD)*, were chosen to create subsets for evaluating experiments. However, most datasets, except MSD, do not have enough music pieces to represent the situation of the real world.

Several benchmarks of MSD and *Last.fm* were used to perform the validation experiments

of our proposed approaches. To investigate each track, *Track-ID*, features, such as *MFCC, Spectral Centoid, Spectral Rolloff Point, Zero Crossings*, and also genre (or style) were the essential information. Features are used to describe the musical content of audio files. Thus, an algorithm is able to recognize the content of a music track without using annotated tags, such as genre and style. Therefore, the appropriate benchmarks are essential to extract features and the class (genre or style) for each track. For example, MSD not only contains audio analysis for one million popular pieces of music, but also contains various music labels, such as genres, styles, song names, and artists.
The proposed music classification approaches were based on genres and styles. In MSD, the AllMusic web page was used to choose genre and style benchmarks, which were called the *AllMusic Genre* and the *AllMusic Style* datasets[10]. It has two partitions. The first one contains 13 genres and classes for 433,714 tracks. The second part is named the *MSD AllMusic Top Genre Dataset (Top-MAGD)* and consists of the 10 large-scale genres of the *AllMusic* web page (see Table 3.1).

Table 3.1: AllMusic Genre Dataset, Top-MAGD.

Genre Name	Num Of Tracks
Pop/Rock	238786
Electronic	41075
Rap	20939
Jazz	17836
Latin	17590
R&B	14335
International	14242
Country	11772
Reggae	6946
Blues	6836
Vocal	6195
Folk	5865
New Age	4010

Another dataset, *AllMusic Style Dataset* [11], was used to compare the results of music

[10]http://www.ifs.tuwien.ac.at/mir/msd/download.html.
[11]http://www.ifs.tuwien.ac.at/mir/msd/MASD.html.

classification based on genre and style. It was created based on grouping tracks into some sub-genres. Thus, several styles could match to one or more genres [49]. In our study, we used all (25) styles and also the 10 major styles in *AllMusic Style Dataset* (see Table 3.2).

Table 3.2: AllMusic Style Dataset.

Name	Num Of Tracks	Name	Num Of Tracks
Big Band	3,115	Metal Heavy	10,784
Blues Contemporary	6,874	Pop Contemporary	13,624
Country Traditional	11,164	Pop Indie	18,138
Dance	15,114	Pop Latin	7,699
Electronica	10,987	Punk	9,610
Experimental	12,139	Reggae	5,232
Folk International	9,849	RnB Soul	6,238
Gospel	6,974	Rock Alternative	12,717
Grunge Emo	6,256	Rock College	16,575
Hip Hop Rap	16,100	Rock Contemporary	16,530
Jazz Classic	10,024	Rock Hard	13,276
Metal Alternative	14,009	Rock Neo Psychedelia	11,057
Metal Death	9,851		

The dataset *Last.fm* includes two types of data: song tags and similar songs. Each track is tagged with different labels by one or more users. Some of the tags represent conflicting or incomplete information. The most important feature of the *Last.fm* dataset is that its 943,347 tracks are matched with the MSD dataset by using their *Track-IDs*. Therefore, it helps us link any subset of *Last.fm* to various MSD resources, including musical features. Several preprocessing techniques were employed to remove redundancy and extract a useful genre set [18]. The genre set consists of 10 major genres that are listed in Table 3.3. Several subsets of these datasets were used to evaluate and analyze the results of each experiment. These experiments are discussed in detail in the following sections.

3.2 Feature Set

Identifying proper features is a non-trivial task of automatic music classification. Furthermore, choosing a list of features is a complex task for common people when they

Table 3.3: The List of Extracted Genres from Last.fm Dataset.

Genre Name	Number Of Tracks
Pop	242,756
Rock	242,756
Electronic	242,706
Rap	242,689
Jazz	242,756
R&B	242,755
Country	242,747
Reggae	242,704
Blues	242,755
Vocal	242,756
Folk	242,753
Newage	242,736
Total	2,670,123

are asked how to discriminate among music genres or styles, even if they can classify those tracks correctly [49]. To select feature sets, Low-level features, MFCC features, and also Rhythm Histogram benchmark were chosen from MSD to make the appropriate datasets. In this study, the Low-level features benchmark was used, which included *Spectral Centroid, Spectral Rolloff Point*, *Spectral Flux*, *Compactness*, *Spectral Variability*, *Root Mean Square*, *Zero Crossings*, and *Fraction of Low Energy Windows*. The MFCC feature set contains 26 features, which are based on the general mean and standard deviations of musical frames. Another feature set, Rhythm Histogram, describes the general rhythmic features in an audio file. The main goal to choose several feature sets is to investigate the musical features from more than one aspect of music content.

3.3 Experiments

In order to prepare the data for experiments, some filtering techniques, discretization techniques, and feature selection methods were performed. For each experiment, some preprocessing methods were applied to prepare the given dataset for the classification task. To control the validity of results, all experiments used a 10-fold cross-validation approach.

In all experiments, the cross-validation tests were performed on different subsets of the selected datasets by one of the given classifiers. Music pieces were randomly grouped into training and testing sets for each fold during the cross-validations. Each music piece was for testing for only one fold and was used for training for other folds. For each fold, the training set was used to train the chosen classifier, which then was tested on the testing set. The accuracy was calculated per fold and used to prepare a report outlining the success of the overall classification rate. Some problems will be discussed regarding the setup of experiments in the following sections.

In our experiments, the accuracy of classification was used as a performance measure. Accuracy measure is the relationship between each classifier and the ground truth class assignments. Most of the time, greater accuracy shows better performance. However, in some situations, it is not reliable for us to just use accuracy to analyze the performance of music classification techniques. It means that in some situations the classifier results in the high accuracy after classifying genres or styles, but it does not mean that all music pieces were classified correctly. This situation will be examined more in the following sections.

Table 3.4: Different experimental sets used when conducting the study.

Set	Factors
1	Supervised and Unsupervised Discretization
2	Balanced and Unbalanced Dataset
3	Comparison Using One Feature Set or Multiple Feature Sets
4	Comparison of the Performance of Feature Selection Techniques and Classifiers
5	Comparison of Selecting Different Numbers of the Highest Ranked Features

The experiments in our study are based on music genre and style. Experiment 1 used styles of music pieces to compare supervised and unsupervised discretization techniques. Moreover, Experiment 4 investigated the effects of several feature selection techniques on music genre classification accuracy. The datasets were created based on classes (genre and

style), which were used in other experiments. Other two experiments used both datasets based on classes, genre and style, to compare the performance of classifying them. Table 3.4 summarizes the five different experimental sets carried out during this study. These experiments, which will be discussed in the following sections, play major roles in analyzing the classification results.

3.3.1 Experiment 1: Supervised and Unsupervised Discretization

Several studies have been investigated using different techniques to discretize nominal numbers. In previous studies, discretization methods are divided into two categories: supervised and unsupervised discretization techniques [12, 31, 60].

When discretization (supervised or unsupervised) is applied to a subset of the nominal attributes, the instances of the subset are categorized into a number of distinct ranges. The purpose of the experiments was to indicate which technique (supervised or unsupervised discretization) makes more improvements in the performance of music classification. In the first set of experiments, MFCC & Style and Low-level & Style datasets were classified. Some datasets were used to compare the accuracy of all classifiers. After comparing the performance of them, *Naive Bayes* (NB), *Sequential Minimal Optimization* (SMO), and *Decision Tree* (DT) with the default settings were chosen to conduct further studies. Table 3.5 shows the datasets that were applied to evaluate the classifiers.

Table 3.5: The subsets of using Low-level features and MFCC features.

Name	Num Of Tracks	Num of Classes	Name of Benchmarks
Low-level & Style	272,561	25	Low-level-MSD&MASD
MFCC & Style	16,834	25	MFCC-MSD&MASD

First, DT was employed without applying preprocessing techniques in the dataset MFCC & Style. The accuracy of music style classification was the lowest one, approximately 7.81%. In the next experiments, some supervised discretization method was applied as a preprocessing step. The results were better, with approximately 11.80%.

Another experiment indicated that unsupervised discretization method was not better than the supervised method for MFCC & Style. NB and SMO were chosen to compare the effects of using discretization techniques. The significant observation that can be made is that supervised discretization method is an effective technique for the dataset MFCC&Style.

In the second part of the experiment, the dataset Low-level & Style was used to analyze the effects of using discretization techniques on the music style classification, and then to compare the results with the previous ones. It is interesting to note that the results showed clear improvements when using some supervised discretization methods. The classification accuracies are presented in Table 3.6.

Table 3.6: The accuracy achieved by using Low-level&Style dataset.

Name of Filter	Decision Tree	Naive Bayes	SMO
Without filter	10.99%	13.94%	16.62%
Supervised Discretization	12.97%	15.27%	17.66%
Unsupervised Discretization	12.13%	14.43%	16.28%

The results of the classification techniques show that using supervised discretization method improves the performance. Thus, the supervised discretization method was chosen for the following experiments that were based on the WEKA data-mining package, and the default parameters were applied. For most experiments of this study, discretization method was added as a supervised filter in the preprocessing step.

3.3.2 Experiment 2: Balanced and Unbalanced Datasets

Some previous work has demonstrated that not only using the appropriate size of datasets can improve the classification quality, but also applying a balanced data is another effective approach in MIR. Since not all genres have similar numbers of tracks, some tracks have to be removed. For example, Norowi et al. [43] investigated different factors, such as dataset size and track length, to improve the classification of performance in the dataset *Traditional Malay Music*. The results showed that the music classification

performance can be improved by providing an appropriate dataset.

In our study, several experiments were conducted to investigate various classifiers. As mentioned above, different subsets of MSD and Last.fm were used to create new datasets, which were used to evaluate the classification performance. The size of each dataset was huge, and also the numbers of music pieces for each genre were not consistent. Thus, huge portions of the dataset were made up of some particular classes. Table 3.7 includes the number of tracks, the number of classes, and the names of benchmarks, which were used to create each dataset. These datasets were unbalanced.

Table 3.7: The experimental datasets

Name	Num Of Tracks	Num of Classes	Name of Benchmarks
Low-level&Style	1,566	25	Low-level-MSD& MASD
MFCC&Style	247,000	25	MFCC-MSD& MASD
Rhythm&Style	272,415	25	Rhythm-Histogram& MASD
Low&MFCC&Rhy&St	1,570	25	Low-level-MSD& MFCC-MSD& Rhythm-Histogram& MASD
MFCC&Genre	23,682	13	MFCC-MSD&Top-MAGD
MFCC&Genre-Balance	68,000	10	MFCC-MSD&Top-MAGD

For experimental purposes, the five classifiers and five feature selection methods were applied to the sample of dataset. The results indicated that NB, SMO and DT Classifiers with two feature selection techniques (*Information Gain (IG)* and *Gain Ratio (GR)*) and also supervised discretization as a filter resulted in better accuracies compared to others. So, these methods were used to evaluate following experiments in this study (see Table 3.8).

As shown in Table 3.8, several unbalanced datasets were used to evaluate the performance of the selected classifiers (NB, SMO and DT). The results showed that classifying

Table 3.8: The highest accuracy achieved by the given classifiers in each experiment.

Dataset	Filter	Feature Selection	Features	Classifier	Accuracy (%)
Low-level&Style	Sup-Dis	IG	16	DT	8.81
MFCC&Style	Sup-Dis	IG	26	DT	10.78
Rhythm&Style	Sup-Dis	IG	60	NB	11.59
Rhythm&Style	Sup-Dis	GR	5	NB	11.035
Rhythm&Style	Sup-Dis	IG	60	DT	9.47
Rhythm&Style	Sup-Dis	IG	5	DT	10.81
Low&MFCC&Rhy&St	Sup-Dis	IG	102	DT	13.37
Low&MFCC&Rhy&St	Sup-Dis	IG	102	SMO	17.2
MFCC&Genre	Sup-Dis	-	26	DT	32.46
MFCC&Genre	Sup-Dis	IG	26	DT	35.35
MFCC&Genre	Sup-Dis	GI	26	DT	34.33

unbalanced datasets is a difficult task, because a classifier is trained based on unbalanced samples of datasets. After the training step, the classifier could identify some classes, which were more frequently available in the given dataset compared to the other classes. The main reason is that the number of samples from some classes is insufficient to report complete details of all music pieces of those classes.

Another important note is that the number of classes has a significant effect on the classifier‘s performance, because a high number of music pieces creates a high probability of error (see Table 3.8). This problem will be discussed in Chapter 4. The first part of Table 3.8 indicates the classification accuracies on music styles. These unbalanced datasets consisted of 25 musical styles. It is interesting to see that the large number of classes has a negative effect on the classifier performance. Furthermore, the low accuracy of each experiment shows that the selected feature set could not represent essential details of each class.

In the second part of Table 3.8, the highest accuracy of music classification is approximately 35.35%. On the other hand, the lowest accuracy is approximately 32.46%, when no feature selection technique was used to rank the MFCC features. The accuracies indicate a passable performance of the DT classifier for different investigations. However,

these results were evaluated in terms of incorrect classification, because the majority of the music pieces in the dataset were classified into incorrect classes. Table 3.9 shows the confusion matrix, where the columns correspond to the actual genre and the rows to the predicated genre. In this confusion matrix, each tag corresponds to a specific genre. The first column of the confusion matrix indicates that the large numbers of music pieces were tagged as *Pop & Rock* class. The reason for this is that the classifier is trained based on the training set, which is a small sample of the main dataset. Therefore, if some classes have more music pieces in the main dataset, the numbers of these classes are more than other classes in the training set. Moreover, the trained classifier does not have enough information on some classes to identify their music pieces correctly. The name of each label can be referred to Table 3.7, along with the corresponding number of pieces.
These calculated results suffer from the major problem of having uneven numbers of music pieces in the selected classes. For example, the number of music pieces tagged with *Pop & Rock, Electronic, Folk*, and *New Age* were 13398, 1559, 302, and 188 respectively. It can be clearly seen that some genres were classified as other ones. Most of the music pieces were classified as *Pop & Rock* (see the first column of the matrix). Thus, the confusion matrix indicates that the subsets are not appropriate for conducting classification. The chosen classifier was confused by the overwhelming number of music pieces of some genres.
The classifier (DT) was trained based on the incomplete sample of datasets. Thus, one of the important tasks is to provide a proper balanced dataset. In order to understand the effects of balanced datasets, experiments were conducted on the newly constructed MFCC-MSD & Top-MAGD. MFCC & Genre-Balance contains the 10 highest genres of Top-MAGD. Different classifiers were used to classify the dataset. Tables 3.10 and 3.11 respectively show the classification performance using SMO (which had the highest correctness, compared to others) and the confusion matrix of the classification. It is clear that the highest accuracy is higher than the one in the previous experiments. Furthermore,

Table 3.9: The Confusion Matrix, Decision Tree (J48).
A: Pop&Rock, B: Electronic, C: Rap, D: Jazz, E: Latin, F: R&B, G: International, H: Country, I: Reggae, J: Blue, K: Vocal, L: Folk, M: New Age

Genre	A	B	C	D	E	F	G	H	I	J	K	L	M
A	13222	21	18	15	16	25	7	17	18	13	22	3	1
B	774	726	13	11	3	5	3	0	14	2	5	0	3
C	129	29	626	1	7	9	1	1	24	2	1	1	0
D	348	29	8	532	1	5	1	3	2	3	13	1	3
E	462	34	34	9	630	16	0	1	10	4	6	1	0
F	409	24	25	17	35	851	2	3	8	5	14	0	1
G	235	14	10	11	19	26	218	4	2	2	4	1	2
H	342	10	5	12	21	23	9	353	2	2	12	0	0
I	130	26	36	10	26	21	12	7	591	2	1	0	1
J	308	24	12	20	24	24	10	15	8	337	2	1	0
K	198	13	14	20	8	27	10	17	9	16	530	1	4
L	109	13	3	14	8	12	6	6	4	7	10	110	0
M	79	9	1	8	2	2	1	2	1	2	14	2	65

though some pieces were classified into incorrect genres, the confusion matrix indicates that the number of correct classified music pieces is higher than the one in the previous experiments.

Table 3.10: The highest accuracy achieved by classifiers in each experiment.

Dataset	Filter	Feature Selection	Num-Of Features	Classifier	Accuracy(%)
MFCC&Genre-Balance	Sup-Dis	IG	10	SMO	42.66

To further validate the experiment, different balanced datasets were created to classify by the selected classifiers. In all experiments, balanced datasets show a considerable improvement on the performance. In the following experiments, the balanced datasets will be used to evaluate the accuracy and reliability of our results.

3.3.3 Experiment 3: Using One Feature Set or Multiple Feature Sets

Much of the past work on music analysis methods was based on audio content, and various feature sets had been tested. There has been some work exploring the effectiveness of joint

Table 3.11: The Confusion Matrix, SMO.
A: Pop&Rock, B: Electronic, C: Rap, D: Jazz, E: Latin, F: R&B, G: International, H: Country, I: Reggae, J: Blue

Genre	A	B	C	D	E	F	G	H	I	J
A	2968	633	246	391	424	385	359	658	201	535
B	616	3275	690	505	205	274	264	205	462	304
C	198	634	3554	122	298	387	89	146	1211	161
D	228	388	100	3929	231	414	359	313	120	718
E	664	253	436	454	1854	879	505	855	385	515
F	270	383	532	433	705	2365	291	714	541	566
G	558	506	416	786	850	601	996	1044	309	734
H	463	101	132	329	572	523	452	3239	126	863
I	112	462	1153	105	262	506	91	188	3634	287
J	508	205	165	737	324	319	353	720	291	3178

use of the two or more types of information sources that attempt to integrate audio contents and non-audio contents for supervised learning, including music classification. For example, Ness et al. [42] presented a study on music classification using short-time features, such as *Spectral Centroid*, *Roll-Off Flux*, *MFCC Mean* and *Standard Deviation*. In addition, Gjerdingen and Perrott [15] and Zhou et al. [68] used Low-level features, such as MFCC, and also some high-level features, such as rhythm and melody, that represented time-varying behavior of music to combine different content sources. In this section, the effects of applying audio information sources and non-audio sources, such as genre and style, are considered in order to improve music classification. The datasets in Table 3.12 were used.

Table 3.13 shows the best performance of different experiments. The main purpose of these experiments was to compare the results of using different feature sets separately and also in a combined way. The balanced datasets were discretized. Five feature selection techniques were applied in order to rank the proper features, including *Correlation-based Feature Selection (CFS)*, *Information Gain (IG)* and *Gain Ratio (GR)*, *Symmetrical Uncertainty (Sym)*, and *Chi-Squared (Chi)*. Furthermore, five classifiers were used, including DT, SMO, NB, BN, and NN.

Table 3.12: The datasets used in the following experiments.

Name	Num Of Tracks	Num of Classes	Name of Benchmarks
Low-level&Style	6,000	10	Low-level-MSD& MASD
MFCC&Style	6,000	10	MFCC-MSD& MASD
Rhythm&Style	6,000	10	Rhythm-Histogram& MASD
Low&MFCC& Rhy&St	6,000	10	Low-level-MSD& MFCC-MSD& Rhythm-Histogram& MASD
Low-level&Genre-Balance	6,000	10	Low-level-MSD&Top-MAGD
MFCC&Genre-Balance	68,000	10	MFCC-MSD&Top-MAGD
Rhythm&Genre-Balance	6,000	10	Rhythm-Histogram& Top-MAGD
Low&MFCC&Rhy& Genre-Balance	6,000	10	Low-level-MSD& MFCC-MSD& Rhythm-Histogram& Top-MAGD

The results in Table 3.13 indicate that using the combination of three feature sets improved the performance more than using the individual feature sets or a combination of two of them. Moreover, the ability of the combined feature sets was improved by applying the discretization method and feature selection techniques. Therefore, the following experiments use the combination of all selected feature sets and the feature selection techniques. In each set of the experiments, the performance of the classifiers is compared as well as the effects of the feature selection techniques.

3.3.4 Experiment 4: Performance of Feature Selection Techniques and Classifiers

Feature selection techniques are applied for dimensionality reduction of music datasets. This section presents five methods to evaluate the importance of features, in which three available feature sets (102 features) are compared. These five feature selection methods (CFS, IG, GR, Sym, and Chi) evaluated the feature importance from different perspectives.

Table 3.13: The highest accuracy achieved by the selected classifiers in each experiment.

Dataset	Filter	Feature Selection	Num Of Features	Classifier	Accuracy(%)
Low-level&Style	Sup-Dis	GR	16	SMO	27.2
MFCC&Style	Sup-Dis	GR	16	SMO	27.2
Rhythm&Style	Sup-Dis	Sym	60	SMO	24.22
Low&MFCC&Rhy&St	-	CFS		SMO	28.71
Low&MFCC&Rhy&St	Sup-Dis	-	102	SMO	27.56
Low&MFCC&Rhy&St	Sup-Dis	CFS		SMO	31.75
Low-level&Genre-Balance	Sup-Dis	Sym	16	SMO	32.22
MFCC&Genre-Balance	Sup-Dis	Sym	24	SMO	36.42
Rhythm&Genre-Balance	Sup-Dis	CFS	26	SMO	27.2
Low&MFCC&Rhy&Genre-Balance	-	GR	45	SMO	31.14
Low&MFCC&Rhy&Genre-Balance	Sup-Dis	-	102	SMO	32.36
Low&MFCC&Rhy&Genre-Balance	Sup-Dis	GR	45	SMO	37.33

To assess the effectiveness of the feature selection methods, five different classifiers were used, including SMO, DT, NN, NB, and BN. Moreover, 10 genres of Last.FM and MSD were chosen to connect with the combined features. Some of the tracks IDs are in overlapping with MSD. These tracks were removed from both datasets. In the next step, each subset of Last.fm and MSD was randomly divided into two balanced datasets. Relatively large datasets were used in these experiments to represent the situation in the real world. The dataset comparison is summarized in Table 3.14.

In the first experiments, Last.fm-1 and Last.fm-2 were separately used as inputs.

Table 3.14: The SubSets Of MSD and Last.fm.

Name	Num Of Tracks	Num of Genres	Name of Benchmarks
Last.fm-1&Genre	6,400	10	Low-level-MSD& MFCC-MSD& Rhythm-Histogram& Last.fm-labels
Last.fm-2&Genre	6,400	10	Low-level-MSD& MFCC-MSD& Rhythm-Histogram& Last.fm-labels
AllMusic-1&Genre	24,260	10	Low-level-MSD& MFCC-MSD& Rhythm-Histogram& MAGD
AllMusic-2&Genre	24,260	10	Low-level-MSD& MFCC-MSD& Rhythm-Histogram& MAGD

Table 3.15: A Comparison of Accuracy Using Different Feature Selection and Classifiers by Using Last.fm-1.

Classifier	CFS	IG	GR	Sym	Chi
SMO	34.95%	33.97%	33.98%	33.97%	33.94
DT	22.31%	21.65%	21.59%	21.69%	21.51%
NB	30.47%	26.59%	26.59%	29.45 %	26.60%
NN	28.14%	32.23%	32.16%	31.36%	32.19%
BN	30.45%	26.61%	26.61%	26.61%	26.61%

Tables 3.15 and 3.16 list the classifiers' accuracies by using each feature selection method. Table 3.15 shows that the highest accuracy of each experiment that was obtained with SMO using the CFS, at approximately 34.95%. In comparison, SMO and NN were proved to be superior to the other classifiers. Moreover, the combination of all classifiers and CFS obtained the better accuracies. However, DT achieved the lowest accuracy with each feature selection.

The results of the second subset of Last.fm dataset were the same as those of the first subset. Table 3.16 indicates that the combination of SMO and CFS have achieved the best

Table 3.16: A Comparison of Accuracy Using Different Feature Selection and Classifiers by Using Last.fm-2.

Classifier	CFS	IG	GR	Sym	Chi
SMO	34.78%	34.39%	34.44%	34.5%	34.42%
DT	23.25%	21.36%	21.45%	21.33%	21.2%
NB	30.73%	27.80%	27.80%	27.80%	27.80%
NN	27.91%	32.39%	32.25%	31.48%	31.76%
BN	30.67%	27.81%	27.81%	27.81%	27.81%

accuracy of 34.78%. DT obtained the lowest accuracy by using Chi as a feature selection technique. Although none of the evaluated classifiers and feature selection methods provided a considerable high accuracy, insightful observation was gained regarding their performance. By comparing Tables 3.15 and 3.16, it is clear that the features selected by IG and GR are not much different. Another significant observation is that CFS is an appropriate feature selection method, which could provide enough details regarding music pieces. To classify the two subsets of AllMusic, CFS and IG were chosen to combine with the selected classifiers. In other words, the results of the selected classifiers and feature selection techniques in the both tables show that the experiments are stable. The descriptive details of datasets are shown in Table 3.14.

Table 3.17: A Comparison of Accuracy Using Different Feature Selection and Classifiers by Using AllMusic-1 & AllMusic-2.

Dataset	AllMusic-1		AllMusic-2	
Classifier	CFS	IG	CFS	IG
SMO	46.14%	46.80%	46.02%	46.32%
DT	27.60%	26.03%	27.49%	27.09%
NB	37.02%	31.60%	31.62%	31.14%
NN	37.52%	32.12%	38.54%	32.80%
BN	37.04%	31.84%	36.94%	32.23%

Table 3.17 indicates the accuracy of each experiment. It is interesting to note that the accuracies of classifiers are not significantly different from the accuracies gained from classifying the Last.fm subsets. SMO with CFS achieved the best accuracies for both

subsets. However, DT could not obtain better accuracies compared to other classifiers. Overall, the results of the AllMusic subsets were better than those produced by the Last.fm subsets. To compare all experiments, the same feature sets were used to create the datasets. In addition, the same classifiers and feature selection techniques were applied to classify the subsets from both datasets. It is clear that the accuracy of the Last.fm subsets was lower than the other datasets. Thus, it is concluded that the music pieces in Last.fm were not tagged with appropriate labels.

3.3.5 Experiment 5: Selecting Different Numbers of Highest Ranked Features

Several feature selection techniques were used to rank features. These techniques rank features based on different perspetives, and then sort the features based on their ranks. The features with high ranks have higher priority, compared to those with lower ranks. As mentioned, some features are not essential in classification. Thus, removing these features does not have any effect on the accuracy of the classifier. In our experiments, different numbers of the highest rank features were used to compare their effect on music classification accuracy. By comparing the results, the accuracy of the classifiers was changed by using different numbers of the highest-ranked features. In this section, some results of classifying several datasets (see Table 3.18) with different numbers of features are shown in Tables 3.19 and 3.20. Several classifiers and feature selection methods were used in this experiment, but only the results of SMO and IG were chosen to compare them, since their accuracies were higher than the other classifiers'.

These tables indicate that the classifier has a specific behavior for different numbers of features. The accuracies were increased gradually by increasing the numbers of features. In classifying some datasets, such as DS7 (see Table 3.19), the accuracies with different numbers of features have fluctuated. It is clear that for classifying 10 classes, a high number of features resulted in better results. However, some datasets with two classes had the highest accuracy after applying some low number of features. Their accuracies did not

Table 3.18: The Descriptive Details of Dataset for Music Genre Classification, AllGenre: {Pop Rock, Electronic, Rap, Jazz, Latin, RnB, International, Country, Reggae, Blues}, AllStyles: {Rock College, Rock Contemporary, Hip hop Rap, Dance, Pop Indie, Rock Hard, Metal Alternative, Pop Contemporary, Rock Alternative, Experimental}, Num Feat: Number of Features.

Name	Instances	Features	Classes	Description
DS1	6000	26	All	MFCC-MSD& AllMusic-Genres
DS2	6000	16	All	Low-level-MSD& AllMusic-Genres
DS3	6000	26	All	MFCC-MSD& AllMusic-Styles
DS4	6000	16	All	Low-level-MSD& AllMusic-Styles
DS5	6000	26	Country, Blues	MFCC-MSD& AllMusic-Genres
DS6	6000	16	Electronic, Rap	Low-level-MSD& AllMusic-Genres
DS7	6000	26	Rock College, Experimental	MFCC-MSD& AllMusic-Styles
DS8	6000	16	Rock Contemporary, Experimental	Low-level-MSD& AllMusic-Styles
DS9	6000	60	All	Rhythm-Histogram& AllMusic-Genres
DS10	6000	60	All	Rhythm-Histogram& AllMusic-Styles
DS11	6000	60	Rap, Reggae	Rhythm-Histogram& AllMusic-Genres
DS12	6000	60	Hip Hop Rap, Dance	Rhythm-Histogram& AllMusic-Styles

change when increasing the number of features.

Tables 3.19 and 3.20 show that the best accuracy of classifying DS6 was shown when using the four highest ranked features. Moreover, DS5 involved two classes that resulted in high accuracy after using the 14 highest ranked features.

The results show that in music classification, if the number of classes is large, we need more features. In most experiments, it is obvious that different features are essential to describe a large number of classes.

Table 3.19: A Comparison of Selecting Different Numbers of the Highest Ranked Features for MFCC & Low-level Datasets, Num Feat : Number of Features.

Num Feat Name	4	6	8	10	12	14	16	18	20	22	24	26
DS1	21.9	22.6	25.3	28.8	30.8	31.5	33.8	34.3	34.7	35.6	36.4	36.3
DS2	22.9	22.9	26.1	25.7	28.2	29.9	32.1	-	-	-	-	-
DS3	20.6	20.3	21.5	23.2	23.2	26.4	28.2	29.2	29.3	29.2	29.4	29.5
DS4	21.6	23.2	24	24.6	26.6	26.8	26.9	-	-	-	-	-
DS5	64.6	65.8	69	70	69.9	71.2	71.2	71.2	71.2	71.2	71.2	71.2
DS6	64.6	64.6	64.6	64.6	64.6	64.6	64.6	64.6	64.6	64.6	64.6	64.6
DS7	64.5	65.2	65.2	64.7	64.7	65.4	66.4	-	-	-	-	-
DS8	64.2	66.2	65.8	66.1	66.9	66.9	67.2	-	-	-	-	-

Table 3.20: A Comparison of Selecting Different Number of the Highest Ranked Features for Rhythm-Histogram Dataset.

Num Feat Name	5	10	15	20	25	30	35	40	45	50	55	60
DS9	20.3	22.4	23	23.3	23.5	24.8	24.9	25.4	25.5	25.8	26	26.8
DS10	18.7	20.8	21.3	22.3	22.6	22.8	23	23.4	23.2	23.3	23.5	24.1
DS11	65	65.9	68.5	68	68.2	68.4	68.2	69.8	70.2	69.6	72.7	72.7
DS12	69.2	70.2	70.6	70.2	70.5	72.3	71.2	71.4	71.2	71.7	72.2	72.2

3.4 Issues in Music Classification

We have designed a set of different experiments to analyze classification in music genre and style. The results of each experiment are compared separately. Improvements are achieved through a series of experiments.

The number of genres or styles is an important issue. If it is higher, as concluded in these experiments, the classifier could not achieve a high accuracy. The problem is that large number of genres or styles confuse the classifiers. In addition, another issue is that the unsuitable features can have effects on the music classification performance, because they confuse the classifier when classifying classes correctly. In Chapter 4, several approaches will be discussed to solve these issues. Furthermore, we will discuss some approaches to improve the music classification performance. The results of the experiments in this chapter will be used there.

Chapter 4

Improving Genre and Style Classification in Music

4.1 Introduction

Music classification based on genre and style is perceptual and subjective. Furthermore, there is no evidence that using one particular approach that improves the performance of music classification will be applicable to other situations, such as different music datasets, different classifiers, etc. The goal of this chapter is to attempt several approaches that will make improvements on music classification performance. A series of experiments were used to discover the relationships between the number of classes and the performance of the given classifiers. From these experiments, many interesting results were obtained to improve music classification accuracy considerably.

In the previous chapter, different aspects of music classification were investigated. The key components of music classification techniques are features and classifiers. The results of the chapter showed that not only was the involved music dataset an important element, but also appropriate feature sets, which have considerable effects on classification accuracy. In this chapter, different balanced datasets were created to evaluate our proposed approaches. Moreover, some classifiers and feature selection techniques, which were outlined in the previous chapter, were selected to compare the outcomes.

This chapter is organized as follows. Section 4.2 explains the effect of using low number of classes on the performance of music classification. It presents two approaches to indicate the direct impact of the number of classes on outputs. The main solution of

improving music classification accuracy is presented in Section 4.3. Several experiments were discussed and analyzed in this section to validate the considerable effects of the proposed approach. Section 4.4 concludes the chapter with some discussions.

4.2 The Effects of Using a Lower Number of Classes

There have been different studies that used various criteria, such as genre, style, mood and also instrument, to classify music. A set of genres and styles were chosen to evaluate music classification accuracy in our study. The intuition is that a better class set will lead to higher music classification accuracy. Thus, when the dataset contains a low number of classes, the classifier is confused less than when a large number of classes are present.

4.2.1 Approach 1: Classifying Each Pair of AllMusic Styles Separately

To evaluate our intuition, two experiments were implemented based on a different number of classes. The first experiment used the *AllMusic Style* dataset (see Table 3.2). Each pair of styles were used to make a dataset for evaluating music classification based on those styles separately. Moreover, three feature sets (Low-level features, MFCC features, and also Rhythm Histogram) were used to provide details regarding each track. As discussed in Chapter 3, SMO with CFS resulted in good accuracies for several classifiers. In the preprocessing step, a dataset was discretized, and also CFS, one of the good feature selection techniques, was used to choose practical features. Furthermore, SMO was used to classify the given dataset. The results of this experiment will be discussed in the following.

Results and Analysis

Tables 4.1 and 4.2 show the accuracy of classifying all possible pairs of styles. It is clear that some styles are much more similar to other; thus, classifying them is not a simple task. On the other hand, whenever styles are more dissimilar, the classifier can recognize them better. Thus, the selected features presented enough details to differentiate each

style. Classifying these instances is not a complicated work. It is interesting that some pairs of styles were classified with high accuracies. For example, some pairs of styles were classified with 100% accuracy when they contained *Metal Death* with another style, such as *Blues Contemporary*, *Country Traditional*, *Hip Hop Rap*, *Reggae*, or *RnB*.

Table 4.1: The Accuracy of Classifying Each Pair (First Part).
A: Big Band, B: Blues Contemporary, C: Country Traditional, D: Dance, E: Electronica, F: Experimental, G: Folk International, H:Gospel, I: Grunge Emo, J: Hip Hop Rap, K: Jazz Classic, L: Metal Alternative, M: Metal Death, N: Metal Heavy, O: Pop Contemporary, P: Pop Indie, Q: Pop Latin, R: Punk, S: Reggae, T: RnB Soul, U: Rock Alternative, V: Rock College, W: Rock Contemporary, X: Rock Hard, Y: Rock Neo Psychedelia

Name	B	C	D	E	F	G	H	I	J	K	L	M
A	78.6	82.2	95.8	96.2	91.2	82.8	88.8	98.4	92	77.3	97	99.2
B		51.4	96.7	93.3	86.2	77.1	74.4	93.7	94.4	86.1	90.9	100
C			94.4	96.6	88.3	79.2	69.5	97.2	93.4	89.3	94.1	100
D				68.8	89.6	90.8	88.7	91.7	83.6	95.6	92.9	98.6
E					84.4	90.9	93.2	91	86.1	89.8	91	96.1
F						79	83.8	80.9	94.3	80	82.3	93.2
G							87.7	92.8	93.6	74.1	90.2	97.5
H								85.5	92.7	85.9	88.8	97.7
I									85.5	92.7	85.9	80.8
J										95.4	96.4	100
K											96.9	98.4
L												81.7

The results of our previous experiments indicated that one of the important aspects, which has a significant effect on the music classification accuracy, is the selected styles that are to be classified. Table 4.3 shows the three appropriate styles for which we created. For example, the first row of the table indicates that if the selected classes of the dataset contained *Big Band* with the another style, including *Metal Death*, *Metal Heavy* or *Grunge*, they were classified with a higher accuracy. Another important note is that these styles with the selected feature sets gave better accuracy. However, it should be noted, if other feature benchmarks were chosen to classify the music pieces, the results might not be the same.

On the other hand, classifying some of these styles with the selected features was

Table 4.2: The Accuracy of Classifying Each Pair (Second Part).

Name	N	O	P	Q	R	S	T	U	V	W	X	Y
A	98.5	85.9	89.8	91.7	95.8	93.3	84.4	92.3	92.1	78.7	93.9	93.8
B	96.6	81.7	75.4	87.8	87.8	88.8	74.3	86.3	82.5	71	87.4	86.2
C	98.3	83.9	82	83.6	93	96.1	80.1	88.6	82.9	79.3	86.2	90.3
D	95.7	80.1	92.6	87.5	94.7	80.5	86.9	90.6	85.7	93.3	96.7	91.2
E	92.4	81.1	94.4	91.6	95.1	79	85.9	89.1	92.7	92.8	94.6	90.6
F	86.6	92.3	67.4	92.6	79.4	90.2	87.1	62.9	63.1	89.1	84.3	80.1
G	97.5	83.6	86.1	81.5	90.9	93.5	80	78.8	76.4	80.9	92.5	88.9
H	95.4	80.5	79.5	73.3	86.8	89.2	68.8	80.8	80.6	66.1	90.9	84.5
I	73.8	93.2	80.3	94.5	69.4	91.8	96.4	75.8	70.5	94	71.3	52
J	97.8	90.1	96.2	91.1	95.9	79.8	90.5	94.8	92.3	96.2	94.8	94
K	98.4	84.4	88.4	93.8	95.6	96.5	83.6	84.7	83.4	78.5	95.2	96.7
L	50.7	92.2	77.9	95.8	70.8	89.2	94.3	78.5	77.7	90.5	72.7	51.9
M	74.4	99.2	96.8	99.1	90.1	100	100	92.1	87.5	99.2	90.7	82.5
N		100	88.8	98.3	70.4	94.9	99.2	81.2	84.7	99.2	73.1	68.5
O			88.4	64	63.4	81.6	52.6	87.1	88	57.8	89.7	85.4
P				92	74.4	92	88.7	58.5	54.5	89.2	79.2	79.5
Q					94.6	84.9	71.3	93.1	85.6	62	94.1	92.2
R						97.7	93.9	71.4	76.5	89.8	67.6	55.4
S							81.5	96.5	93.6	90.3	91.5	91.3
T								89.8	85.8	66.9	90	91.4
U									62.2	84.5	74.2	72
V										86.4	75.2	76.9
W											84.8	92.6
X												93

non-trivial, because these features are not the appropriate ones to represent enough information for the selected classes (genre or style). For example, the accuracies of classifying *Pop Contemporary* or *Pop Indie* with most other styles were low. The accuracy of classifying *Pop Contemporary & RnB* was approximately 52.6%, which meant that they were classified almost randomly. It is interesting to note that several features of some classes, such as *Pop Indie* and *Rock College*, were discretized into the same ranges; thus, classifying them is a complex task for most of the classifiers. Table 4.4 shows the results of classifying this pair.

The confusion matrices of these experiments indicate that music tracks with the *Pop Indie*

Table 4.3: The Appropriate Style Sets.

Style	Three Proper Styles	Style	Three Proper Styles
A	M, N, I	N	O, T, W
B	M, D, N	O	N, M, W
C	M, N, I	P	M, J, E
D	M, X, N	Q	M, N, L
E	M, R, X	R	S, J, A
F	J, Q, O	S	M, R, K
G	M, N, J	T	M, N, I
H	M, N, J	U	S, J, Q
I	T, Q, W	V	U, J, A
J	M, N, L	W	M, N, J
K	M, N, L	X	R, D, K
L	Q, T, O	Y	K, J, X
M	S, T, O		

Table 4.4: The Accuracy of Classifying Pop Indie and Rock College.

Name Dataset	Filter	Feature Selection	Classifier	Accuracy (%)
MFCC+Style	Dis	-	SMO	48.37
MFCC+Style	-	CFS	SMO	50.65
MFCC+Style	Dis	IG	SMO	55.16

were classified as *Rock College*. The same results were observed by using low-level features, such as *Spectral Centroid*, *Spectral Rolloff Point*, *Spectral Flux*, *Compactness*, *Spectral Variability*, *Root Mean Square*, *Zero Crossings*, and *Fraction of Low Energy Windows*. Furthermore, the accuracy of using all selected feature sets was not better than using the individual one (see the accuracy of *Pop Indie and Rock College* in Table 4.2). Overall, using appropriate classes is an important task in style classification. Thus, one approach is to conduct classification on each pair of classes separately for a dataset with a large number of classes. Furthermore, this approach indicates which classes are similar, and also which ones are different. To improve the accuracy of similar classes, using other feature sets might be helpful. Different feature sets could represent different aspects of music contents. Thus, using various feature sets could provide more details for each pieces of music.

4.2.2 Approach 2: Removing Some Classes

As mentioned in the previous sections, some classes were not recognized correctly. Different issues could cause this problem. For example, the feature sets were not suitable to present the differences among those classes and others. Moreover, the large numbers of classes confused the classifier. Thus, our proposed approach is to remove instances that are tagged with those classes. In this section, different experiments have been done to indicate the effects of using dissimilar styles on the classifier performance.

The first experiments used the dataset that was created based on 25 AllMusic Styles[12], and the three selected feature sets. The supervised discretization method was used to discretize the dataset. Firstly, DT with IG was chosen to classify them. The accuracy of the classification was low. The confusion matrix (Tables 4.1 and 4.2) indicated which styles were more different from others. To compare the results of the high and low number of styles, six different styles, including *Big Band, Experimental*, *Hip hop Rap*, *Jazz Classic*, *Metal Death*, and *Punk*, were chosen to continue the investigation of the approach.

In the following experiments, CFS, IG, GR, and Sym were used to select the appropriate features. Moreover, SMO, NB, and DT were chosen to classify the dataset. The new dataset was created based on the six styles and the three feature sets, and was discretized by the supervised discretization method. After getting all the results, the confusion matrix was analyzed to find which styles were frequently classified as others. These styles were removed. After removing similar styles, three remaining styles, including *Big Band*, *Hip Hop Rap*, and *Metal Death*, were used to create the new dataset. This dataset was used to evaluate the performance of style classification based on the new classes in the third experiment.

Results and Analysis.

Several combinations of classifiers and feature selection methods were used to compare the classification results. Table 4.5 shows which classifier had the highest accuracy with

[12]http://www.ifs.tuwien.ac.at/mir/msd/download.html.

different feature selection techniques.

Table 4.5: The Best Accuracy of Classifier by Using Different Feature Selection Techniques.

Feature Selection	Classifier	Number of Style	Accuracy (%)
CFS	SMO	6	95.43
CFS	SMO	3	97.11
IG	SMO	6	90.62
IG	SMO	3	95.25
GR	SMO	6	90.68
GR	SMO	3	100
Sym	SMO	6	90.71
Sym	SMO	3	100

In these experiments, it is clear that SMO appears to be the best option. On the other hand, NB could not work effectively, and the results were lower than others in all classifications (around 45% after using one of the feature selection techniques). It is obvious that the number of classes affected music classification. Furthermore, the similarity of classes had considerable effects to prevent the classifier from recognizing them correctly. It is interesting to note that SMO is one of the best classifiers in this study. Different datasets were classified with high accuracy by SMO. The main purpose of this experiment is to find the acceptable accuracy with the suitable classes, but not with all of them. However, if we want to classify the exact number of classes (i.e., removing no any classes), this approach would not be the best one.

4.3 Grouping The Similar Classes

As mentioned, the main purpose of music classification techniques is to find the best way to classify all classes. The previous approach removed similar classes. Thus, classifiers performed better based on the remaining classes. On the other hand, in most situations, the purpose is to classify all classes, not only a small number of them. Therefore, We proposed another approach toward this end.

The approach is based on the similarities among classes. It means that classes, which are more similar, are categorized into a group. Then, our proposed approach classifies those groups as new classes. In the second step, each group is classified separately. If a group has large numbers of classes, the similar classes could be categorized into a new group and we create subgroups for each main group. This step will be done recursively until two classes are remained in all subgroups. To calculate the accuracy of music classification, the average of accuracies has been calculated for subgroups of each group. The final accuracy is the average of the main groups' accuracies. These steps are defined in the following algorithm. In the following sections, some examples are presented to analyze our proposed approach.

In each experiment, we used the confusion matrix o identify the similar classes. The scenario of grouping the similar classes is as follows:

n: Number of classes in the dataset

k: Number of top maximum classes, which are confused the classifier more than others, $k = 1, ..., n-1$

C_j: The classes used in the dataset, $j = 2, ..., n$

L_i: The list of classes, which are more different from other unselected classes. Each class has a list, $i = 2, ..., n$

G_m: The groups of classes, $m = 2, ..., w$, w depends on the number of n

L_{temp}: The temporary list

Algorithm 1 (Grouping Similar Classes Algorithm).

1. for all classes of the dataset

    Add k numbers of C_j in L_i

2. for all L_i

    if L_i is empty,

        Create G_i, add C_i in L_i

    elseif there is C_j in L_i, and C_i is in L_j

if there exists G_m and it includes C_i

Add C_j to G_m

elseif there exists G_m and it including C_j

Add C_i to G_m

else

Create G_m, add C_i and C_j

elseif there is C_j in L_i, and C_i is not in L_j

Add to L_{temp}

3. Check the members of L_{temp}

if L_{temp} is empty

stop

else

Find the appropriate group for each member of L_{temp} based on the number of incorrect classified instances of each class in the Confusion Matrix.

4. Stop

The selected classifiers (SMO, NB, NN, BN, and DT) are used to classify all classes. The results of the classifier include a confusion matrix that shows how each class is tagged. For example, the row i shows how class i is tagged. In other words, the values indicate that how many of instances are tagged incorrectly as other classes in columns j $(j = 1, ...n)$. For each class, the k highest values are added to the list, which shows which classes are tagged as class i more than others. When the lists of all classes are created, these lists are used to group those classes, which are tagged as each other more than other classes. That is, if the instances of class C_i are labeled as class C_j, and the instances of C_j are tagged as C_i, these classes are so similar. Thus, these classes are grouped together. The temporary list (L_{temp}) is used for those classes that are tagged as different classes frequently. In this situation, each class in L_{temp} is considered with those classes that are tagged with it

frequently. The sum of the values (i.e., those values in the confusion matrix) of each pair of those classes is used to find which group is more suitable for the selected class in L_{temp} (the group that contains more similar class(es)).

For example, Table 4.6 shows the confusion matrix of classifying five classes. The first row shows that 148 pieces of music are tagged correctly. It means that this class does not confuse the classifier. The third and fifth rows of the confusion matrix indicate that pieces of music of class B and D are tagged as each other frequently. These two classes are similar, and the classifier does not tag them with high accuracy. Therefore, Group B is created to add music pieces of the both classes. The third row shows that 40 music pieces of class C are labeled as class A. Thus, it means that the classifier is confused. Therefore, these music pieces are similar with the ones from class A. In this situation, all pieces of class C are added to Group A. The last row indicates that most of the music pieces of class E are tagged correctly, and also the last column shows that music pieces of other classes are not tagged frequently as class E. Thus, the instances of this class form a separate group, Group C. After grouping instances of all classes, each group is classified correctly. If a group contains the instances from more than two classes, the instances could be grouped based on similar classes again.

Table 4.6: Confusion Matrix.

Genre	A	B	C	D	E
A	148	0	1	1	0
B	2	101	6	40	1
C	40	2	99	1	0
D	3	50	0	92	5
E	0	2	1	4	143

Several experiments were conducted to investigate this approach on different balanced datasets. The datasets contained a restricted set of classes that had 10 top classes. All experiments used WEKA, and also were implemented in Java. As a preprocessing step, all datasets were discretized by the supervised discretization technique.

To compare the results and choose the best one, five selected classifiers (SMO, DT, NN, NB, and BN) and also five feature selection techniques (CFS, IG, GR, Chi, and Sym) were chosen [28, 50, 54, 62]. After discretization, some features were categorized into one range for most of the instances [12, 31]. In our study, each experiment was repeated several times while choosing different numbers of the highest-ranked features. The feature selection techniques with different number of the highest ranked features were used, in order to compare the accuracy of each classification and choose the best possible feature set. All experiments are presented with more details below.

4.3.1 Music Genre Classification

Experiments involving music genre classification are discussed in this section. Several datasets were used to evaluate and analyze the suggested approach. Table 4.7 lists the 10 top genres that were selected to create the practical datasets. In addition, the descriptive details of the datasets are shown in Table 4.8.

Table 4.7: The Descriptive Genres for Music Genre Classification.

Name	List of Classes
AllMusic-Genres	Pop Rock, Electronic, Rap, Jazz, Latin, RnB, International, Country, Reggae, Blues
Last.fm-Genres	Pop, Rock, Electronic, Rap, Jazz, RnB, Country, Reggae, Blues, Folk
LatinDataset	Tango, Bolero, Batchata, Salsa, Merengue, AxE, ForrU, Sertaneja, Gaucha, Pagode

Classifying the Latin Dataset.

One of the good datasets in MIR is the *Latin Dataset* that includes 10 genres (see Table 4.7). However, it does not have a large number of tracks for each genre. The *MARSYAS* framework was used to extract 30 features of the Latin dataset. Table 4.9 shows that the features were grouped into three categories: *Beat Related* (features 1 to 6), *Timbral Texture* (features 7 to 25), and *Pitch Related* (features 26 to 30) [53].

Table 4.8: The Descriptive Details of Datasets for Music Genre Classification.

Name	Num of Instances	Num of Features	Description
AllMusic-MFCC-Gen	6000	26	MFCC-MSD& AllMusic-Genres
AllMusic-Low-Gen	6000	16	Low-level-MSD& AllMusic-Genres
AllMusic-Rhy-Gen	6000	60	Rhythm-Histogram& AllMusic-Genres
AllMusic-All-Gen	6000	102	Low-level-MSD& MFCC-MSD& Rhythm-Histogram& AllMusic-Genres
Last.fm	6000	102	Low-level-MSD& MFCC-MSD& Rhythm-Histogram& Last.fm-Genres
Latin Dataset	3000	30	middle-LatinDataset

The Latin dataset has three subsets, which are based on 30 seconds of first (D_f), middle (D_m), and end parts of each track (D_e). The selected features were extracted from those subsets. Some studies had previously been investigated on different parts of music (see Table 4.8) [52, 53, 55].

Several studies concluded that D_m and D_e of music represent more information, compared to D_f (see Table 4.8) [52, 53, 54, 55]. In our study, they were used to investigate which subset was better by using the selected classifiers and feature selection methods.

Tables 4.10 and 4.11 show the classifying performance using (D_m) and (D_e) respectively. The results of the experiments show that the accuracies of using D_m were higher than using D_e and that NN appeared to be the best option for classifying them. However, DT showed the worst accuracies compared to others in both datasets.

It is clear that the results of using D_m is better than D_e. By considering the confusion matrix of D_m, it is obvious that the error rate of classification is low. But, some genres were classified as other genres. Table 4.12 shows the confusion matrix of the best accuracy that was carried out by NN and Sym.

Table 4.9: The List of Feature Vector Description of the Latin Dataset.

#	Description	#	Description
1	Relative amplitude of the 1st histogram peak	16	1st. MFCC mean
2	Relative amplitude of the 2nd histogram peak	17	2nd. MFCC mean
3	Ratio between the amplitudes of the 2nd peak and the first peak	18	3rd. MFCC mean
4	Period of the 1st peak in bpm	19	4th. MFCC mean
5	Period of the 2nd peak in bpm	20	5th. MFCC mean
6	Overall histogram sum (beat strength)	21	Standard deviation for 1st. MFCC
7	Spectral centroid mean	22	Standard deviation for 2nd. MFCC
8	Spectral rolloff mean	23	Standard deviation for 3rd. MFCC
9	Spectral flow mean	24	Standard deviation for 4th. MFCC
10	Zero crossing rate mean	25	Standard deviation for 5th. MFCC
11	Standard deviation for spectral centroid	26	The overall sum of the histogram (pitch strength)
12	Standard deviation for spectral rolloff	27	Period of the maximum peak of the unfolded histogram
13	Standard deviation for spectral flow	28	Amplitude of maximum peak of the folded histogram
14	Standard deviation for zero crossing rate	29	Period of the maximum peak of the folded histogram
15	Low energy	30	Pitch interval

Table 4.10: Comparison Accuracy of Using Different Feature Selection and Classifiers by Using Middle-Latin Dataset (D_m).

Classifier	CFS	IG	GR	Sym	Chi
SMO	73.63%	73.97 %	73.93%	73.97%	73.97%
DT	54.3%	53.4%	53.17%	53.37%	53.4%
NB	69.57%	69.13%	69.13%	69.13%	69.13%
NN	72.33%	75.2%	75.17%	75.23%	75.23%
BN	69.83%	69.27%	69.27%	69.27%	69.27%

In our experiment, the selected classifiers and features classified all classes. After this step, the confusion matrix of the best accuracy was considered to determine the classes, which were classified as other classes frequently. Those genres were then chosen and categorized

Table 4.11: Comparison Accuracy of Using Different Feature Selection and Classifiers by Using End-Latin Dataset (D_e).

Classifier	CFS	IG	GR	Sym	Chi
SMO	65.7%	64.83%	64.87%	64.93%	64.87%
DT	48.5%	48.57%	48.5%	48.53%	48.53%
NB	60.7%	60.6%	60.6%	60.6%	60.6%
NN	64.23%	65.77%	66.3%	65.87%	63.9%
BN	60.63%	60.7%	60.7%	60.7%	60.7%

Table 4.12: The Confusion Matrix: NN & Sym.
Tan: Tango, Bol: Bolero, Bat: Batchata, Sal: Salsa, Mer: Merengue, For: ForrU, Ser: Sertaneja, Gau: Gaucha, Pag: Pagode

Genre	Tan	Bol	Bat	Sal	Mer	AxE	For	Ser	Gau	Pag
Tan	290	0	0	0	0	7	0	3	0	0
Bol	0	232	15	8	4	8	5	14	5	9
Bat	1	13	178	19	4	10	8	26	20	21
Sal	1	8	14	201	4	5	8	25	22	12
Mer	0	7	3	1	283	2	1	2	0	1
AxE	11	12	14	8	1	218	0	8	25	3
For	0	7	6	6	10	0	262	6	1	2
Ser	3	8	28	24	2	19	6	181	18	11
Gau	0	9	22	27	3	27	2	11	183	16
Pag	1	4	25	14	1	10	1	5	10	229

into one group. The new datasets were created with new labels. For example, the main dataset contained 10 genres. After classifying the main dataset, the similar genres in D_m were grouped as follows: Group A: *Salsa*, *ForrU*, *AxE*, *Sertaneja*, *Gaucha*, *Pagode*, Group B: *Tango, Bolero*, and Group C: *Batchata, Merengue*. The instances of these groups were tagged by new labels that were Group A, Group B, and Group C, when creating a new dataset. Then, IG and NN classified the dataset again, and the best accuracy was approximately 90.94%. In the next step, a new balanced dataset was made for each group, and each of them was classified separately. The accuracy of Group A, Group B, and Group C were 82.67%, 96.83%, and 90.94% respectively. To improve the accuracy of Group A, the confusion matrix of the best accuracy was again used to select the similar genres. The instances of the genres were separated into two new subgroups, Group AA (*ForrU* and

Sertaneja) and Group AB (*AxE and Gaucha*). Then, each group was classified separately. This process continued until each main group or sub-group contained two genres. To estimate the accuracy of classifying all genres, the average of accuracies was calculated for sub-genres of each group. Then, the average of the main groups was computed to gain the final accuracy.

Analyzing Results of Classifying the Latin Dataset

Figures 4.1 and 4.2 show the hierarchical tree of the grouped genres and the averages of D_m accuracies respectively.

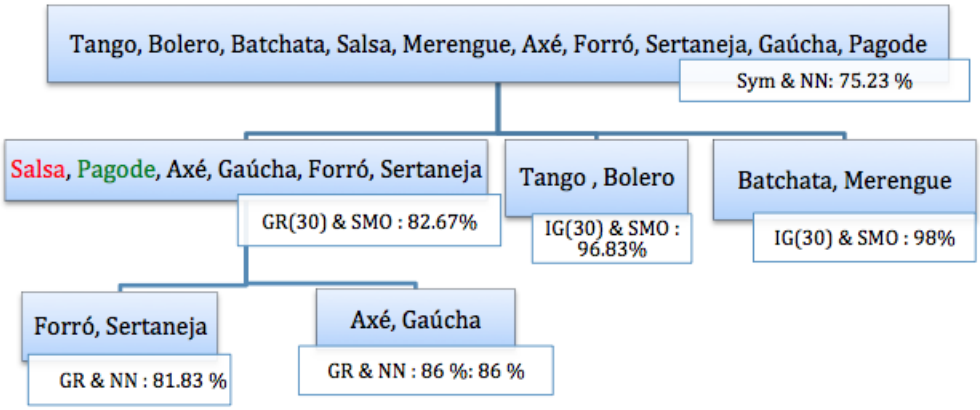

Figure 4.1: The Hierarchical Tree of Latin Dataset Accuracies.

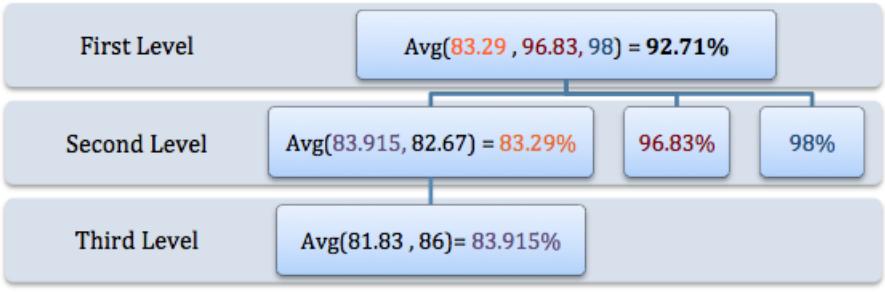

Figure 4.2: The Averages of Latin Dataset Accuracies.

As mentioned before, the chance of choosing wrong classes decreases considerably by using a low number of classes. By comparing the final accuracy of the proposed approach and the previous approach (approximately 92.71% and 75.23%, respectively), there is an obvious improvement in genre classification. In addition, SMO and NN appeared to be

better than the other classifiers. However, there were no specific methods that were chosen to select the suitable features for all groups. In other words, the best accuracies were obtained by using different feature selection methods and by experiments for each different group of genres.

As discussed, several groups used various numbers of features, and each feature might have different ranks for classifying each group. Table!4.13 shows the highest ranked features, which were used to classify each group in this dataset.

Table 4.13: Comparison Different Highest Ranked Feature for Each Group of Middle-Latin Dataset.
Group A: {Salsa, Pagode, Ax, Gacha, Forr, Sertaneja}, Group B: {Tango , Bolero}, Group C: {Batchata, Merengue}, Group AA: {Forr, Sertaneja}, and Group AB: {Ax, Gacha}

Name of Group	List of Features
AllGenre	11, 9, 10, 15, 13, 4, 14, 20, 23, 12, 19, 16, 21, 26, 5, 6, 7, 2, 22, 8, 25, 24, 18, 17, 3, 0, 1, 28, 27
A, B, C	9, 12, 11, 20, 4, 13, 15, 10, 26, 19, 6, 21, 5, 23, 24, 2, 0, 8, 25, 17, 16, 22, 18, 7, 3, 1, 28, 27
A	16, 7, 26, 10, 2, 19, 24, 25, 13, 3, 14, 4, 20, 23, 11, 18, 8, 15, 0, 21, 22, 6, 9, 28, 5, 27, 1, 12, 17
B	13, 4, 10, 11, 9, 23, 19, 16, 25, 22, 18, 7, 15, 17, 8, 24, 26, 20, 2, 14, 21, 3, 28, 12, 0, 5, 1, 6, 27
C	14, 23, 5, 7, 6, 8, 22, 15 ,16, 20, 19, 21, 2, 10, 25, 3, 24, 18, 26, 4, 9, 1, 0, 11, 13, 12, 17, 27, 28
AA	10, 6, 8, 12, 15, 19, 21, 13, 26, 23, 5, 14, 7, 11, 2, 3, 18, 9, 16, 22, 20, 17, 25, 24, 1, 28, 4, 27, 0
AB	14, 10, 26, 24, 19, 15, 2, 20, 9, 13, 7, 23, 16, 5, 22, 25, 6, 17, 18, 1, 4, 3, 28, 11, 21, 27, 12, 0

It is clear that not only the numbers of effective features were changed for each group, but also each feature could get different ranks. On the other hand, while some features could be useful to classify some genres, it does not mean they are practical for classifying others. For example, Feature 11 had the highest ranked accuracy in classifying all genres. This means that it represents much more information about those classes when compared to others, and it has more effects on classifying them (see Table 4.13). However, the rank of this feature were decreased in sub-groups. When classifying Group BBB, it was one of the

lowest ranked features. In the following, we present several experiments that have been conducted to investigate the proposed method.

Classifying AllMusic Datasets

In this section, different subsets of AllMusic dataset were created to evaluate music classification based on genres (see Table 4.8). In the previous discussions, some experiments concluded that the performance of using several feature sets was better than using an individual feature set in music classification. Our proposed approach was used to compare the performance of music genre classification based on an individual feature set or a multiple feature set. In these experiments, the selected classifiers and feature selection techniques were used to compare and analyze the results.

Analyzing Results of Classifying AllMusic Datasets

In the first experiment, AllMusic-MFCC-Gen dataset was used to classify instances several times with different combinations of classifiers and feature selection techniques. The experiments were repeated several times using different numbers of the highest ranked features. The first experiment used even numbers of features from 2 until 26 for each classifier and feature selection technique, except CFS.

Figure 4.3 shows the accuracy of grouping similar genres. In the first level, all genres were classified, and the accuracy was approximately 36.42%. In the next level, each group was classified separately. The best possible result of each group is shown in Figure 4.3. If each group had more than two genres, some subgroups of them ware created. In this figure, the numbers that are between parentheses show how many features were used to obtain the best accuracy.

There are different numbers of features that were used to obtain the best accuracy. It is obvious that different genres were represented by various numbers and types of features. Figure 4.4 indicates the accuracy of getting the average among subgroups. The final accuracy is approximately 72.19%, which is much higher than 36.42% (see Figures 4.3

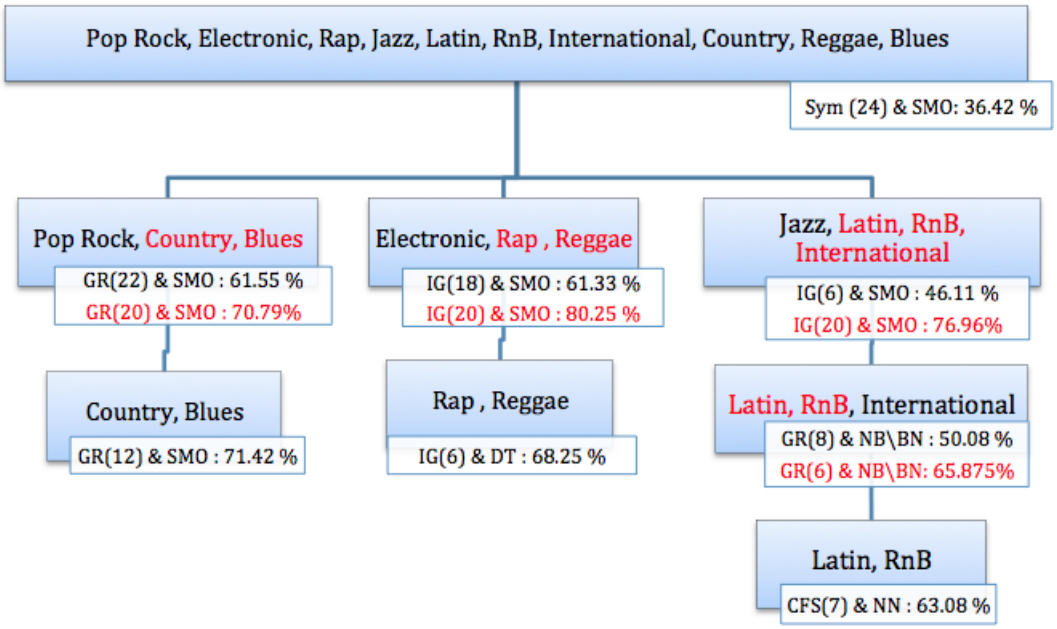

Figure 4.3: The Hierarchical Tree of AllMusic-MFCC-Gen Dataset Accuracies.

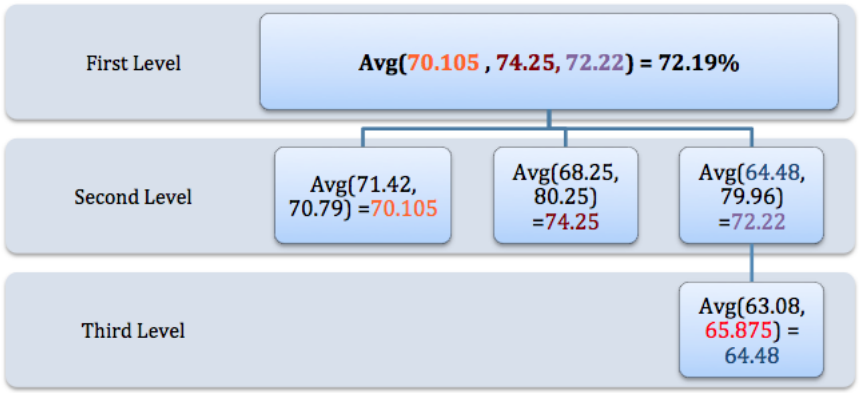

Figure 4.4: The Averages of AllMusic-MFCC-Gen Dataset Accuracies.

and 4.4).

In the next experiment, AllMusic-Low-Gen dataset was used. This dataset was made by using Low-level features and 10 selected genres from AllMusic benchmarks. Figures 4.4 and 4.5 indicate the results of the experiments.

Figure 4.5 indicates that the best accuracy of classifying all selected genres is around 32.22%. By grouping the similar genres, the final accuracy is around 67.98%. As mentioned in the previous discussions, using fewer number of genres could reduce the

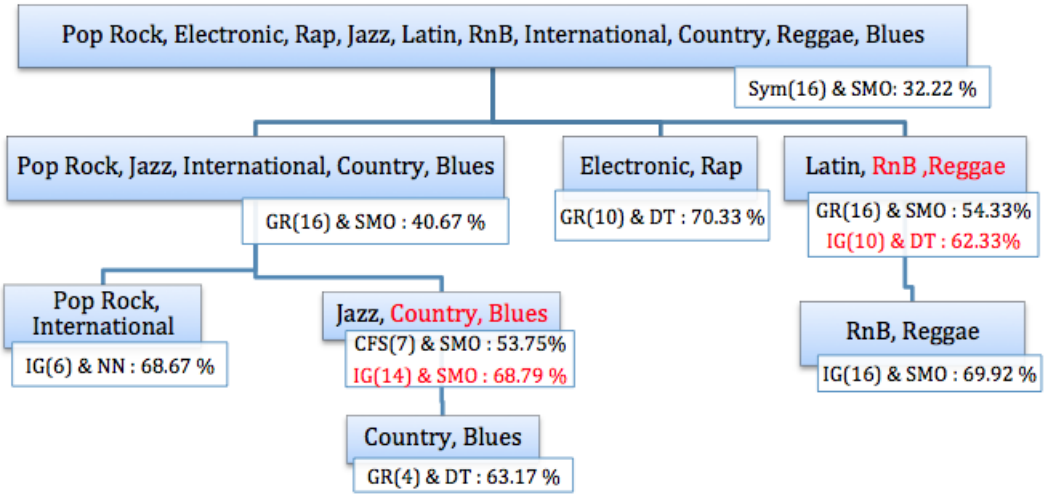

Figure 4.5: The Hierarchical Tree of AllMusic-Low-Gen Dataset Accuracies.

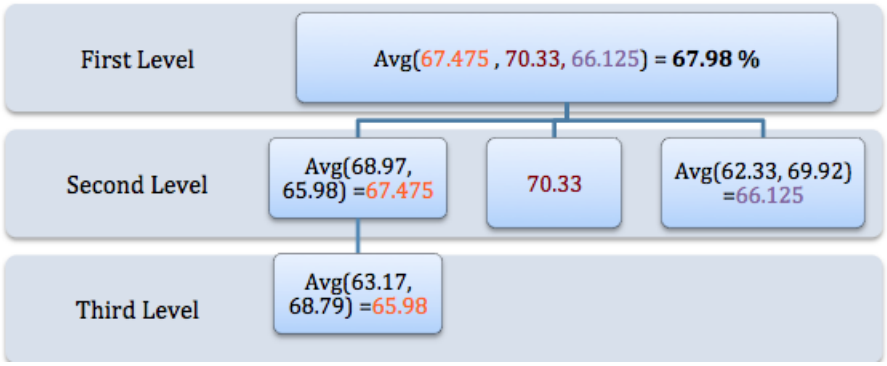

Figure 4.6: The Averages of AllMusic-Low-Gen Dataset Accuracies.

error rate. Figure 4.8 shows the accuracy of each group, and also the final accuracy.

The dataset AllMusic-Rhy-Gen is the next one that was used to evaluate our proposed approach. To create this dataset, 60 features and 10 genres were used. This experiment was repeated several times to investigate how many features of the selected feature set were useful and to provide sufficient information for the classifiers.

Figure 4.7 shows that the accuracy of classifying genres *Latin* and *RnB* is 50%. By considering the confusion matrix, the classifier could not identify any instance as *RnB*, and all music pieces were tagged as the other genre. It is interesting that any feature selection methods and classifiers with different numbers of features could not improve this issue. Thus, it is clear that this feature set could not provide enough information for all genres.

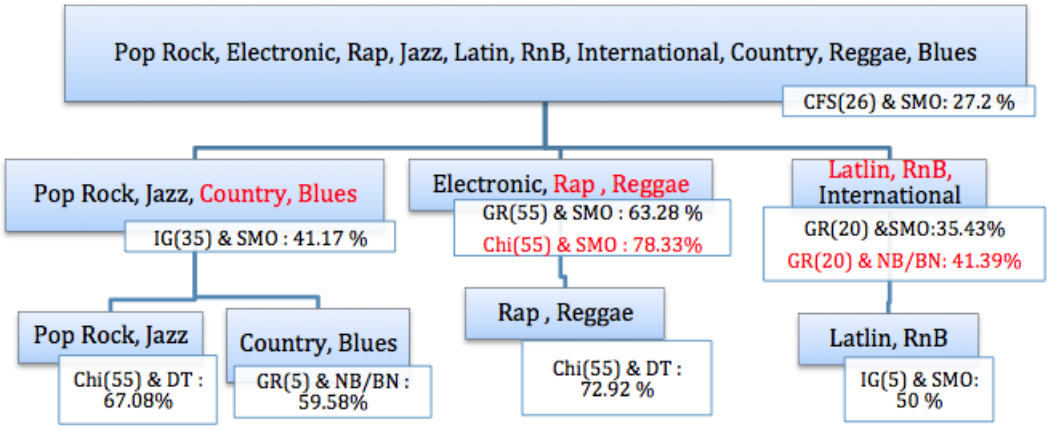

Figure 4.7: The Hierarchical Tree of AllMusic-Rhy-Gen Dataset Accuracies.

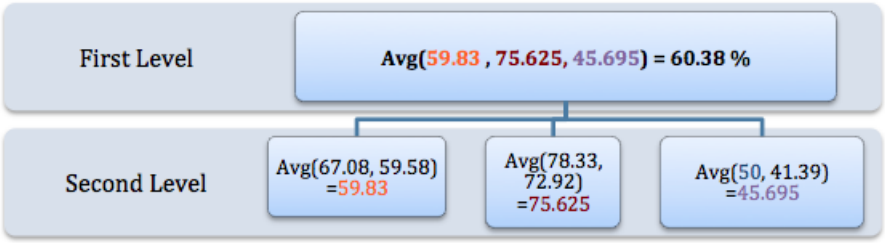

Figure 4.8: The Averages of AllMusic-Rhy-Gen Dataset Accuracies.

However, the final accuracy (see Figure 4.8) was considerably improved if it compared to the previous result.

The feature sets were evaluated by the selected classifiers. It is noted that the accuracy of classifying the dataset based on those feature sets increased considerably. To compare the results of individual feature sets, we find the dataset that provided MFCC features had higher accuracy. Moreover, some genres, such as *Country* and *Blues*, were grouped together in all experiments. The accuracy of classifying the dataset that was created by using MFCC features is higher than using other feature sets. The classifiers performed better when the dataset used MFCC features.

In the next experiments, all feature sets were used to create the dataset AllMusic-All-Gen. Figures 4.9 and 4.10 show the hierarchical trees of the grouped genres and the averages of the accuracies respectively.

The performance of classifying all genres by using this dataset was better than the scenario

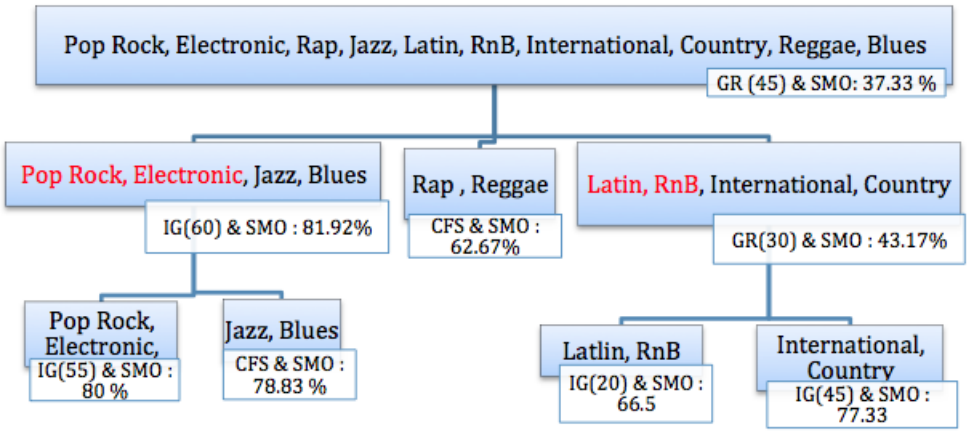

Figure 4.9: The Hierarchical Tree of AllMusic-All-Gen Dataset Accuracies.

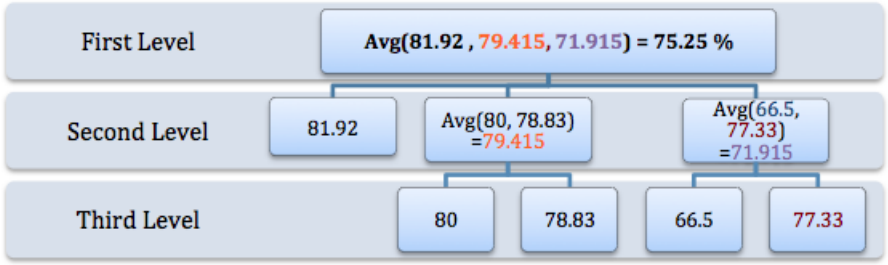

Figure 4.10: The Averages of AllMusic-All-Gen Dataset Accuracies.

where the previous AllMusic datasets were used. Using this new approach increased the accuracy of classification from around 37.33 % to 75.25%. To compare the performance of several classifiers, SMO was chosen more than the other classifiers as it resulted in the best accuracy. Moreover, IG, GR, and CFS performed better than the other methods. However, there was no exact number of features for each feature set that was selected as the best one. Another significant observation is that the accuracy of using all feature sets for classifying *Rap* and *Reggae* was lower than using the dataset AllMusic-Rhy-Gen. However, its hierarchical tree shows higher accuracies when datasets were classified with a low number of classes.

Summary of Classifying AllMusic Datasets

All experiments show that various features were chosen to use in music classification. The feature sets do not have same number of features, i.e., Low-level features, MFCC, and

Rhythm Histogram have 16, 26, and 60 features respectively. Not all of these features are useful for all genres. To investigate which features of each feature set were used more than others, the selected features in each step were considered. The percentage of each feature set was calculated to make them more comparable (see Table 4.14).

Table 4.14: Comparison Percentage of Different Feature Sets for Each Group of AllMusic Datasets.
Group A: {Pop Rock, Electronic, Jazz, Blues}, Group B: {Rap , Reggae}, Group C: {Latin, RnB, International, Country}, Group AA: {Pop Rock, Electronic}, Group AB: {Jazz, Blues}, Group CA: {Latlin, RnB}, and Group CB: {International, Country}

Name of Group	Low-level (%)	MFCC (%)	Rhythm Histogram (%)	Total Num
AllGenre	20	17	62	45
A, B, C	20	50	30	15
A	15	6	78	60
B	37.5	42	21	24
C	43	14	40	30
AA	14.5	14.5	71	55
AB	17	46.4	35.7	13
CA	35	20	45	20
CB	24	33	42	45

An observation is that there were a different number of features that were chosen. For example, the first row of the table (see Table 4.14) shows that 62% of features were chosen from the Rhythm-Histogram feature set, and the lowest percentage (17%) involved choosing MFCC feature sets. However, different features were selected for classifying Group A, Group B, and Group C (from the second row until the fifth row of Table 4.14). The highest percentage involved using the MFCC feature set, and the lowest accuracy used the Low-level feature set. For the first dataset, 45 features resulted in the best accuracy; but for another dataset 15 features were enough to gain better results.

Classifying Last.fm Dataset

The dataset Last.fm is another dataset that was chosen to show the performance of our approach. Ten genres were selected, which were combined with the three feature sets of AllMusic[13] (see Tables 4.7 and 4.8). The same selected methods were used in this dataset as in the previous experiments.

Analyzing Results of Classifying the Last.fm Dataset

The results indicate that SMO obtained higher accuracies for most groups. The accuracy of classifying this dataset increased from 34.68% to 72.098% by grouping the similar genres. Furthermore, the accuracy of the same sub-groups in the low level of the hierarchical tree was high. For example, *Rap & RnB* could be classified with 81.17% accuracy. The results are shown in Figures 4.11 and 4.12.

The number of features is shown in Table 4.15. The best accuracies for some groups were obtained from the large number of features. It is clear that it is a complex task to find a feature set that represents all classes of a dataset. Furthermore, another significant observation that can be made is that the choice of the appropriate feature set and also the number of classes have a direct impact on each other. The feature sets could cover the low number of classes much more easily.

By considering the results of all experiments, our proposed approach has a significant effect on the performance of music genre classification. It is obvious that not only some classifiers obtained higher accuracies compared to others, but also some feature selection methods performed better than others.

4.3.2 Music Style Classification

Music style classification was also considered in the past few years with some features and classifiers [10, 39, 59]. Several studies have investigated the difference among the classification tasks, such as genre classification and style classification, and have examined

[13] http://www.ifs.tuwien.ac.at/mir/msd/download.html.

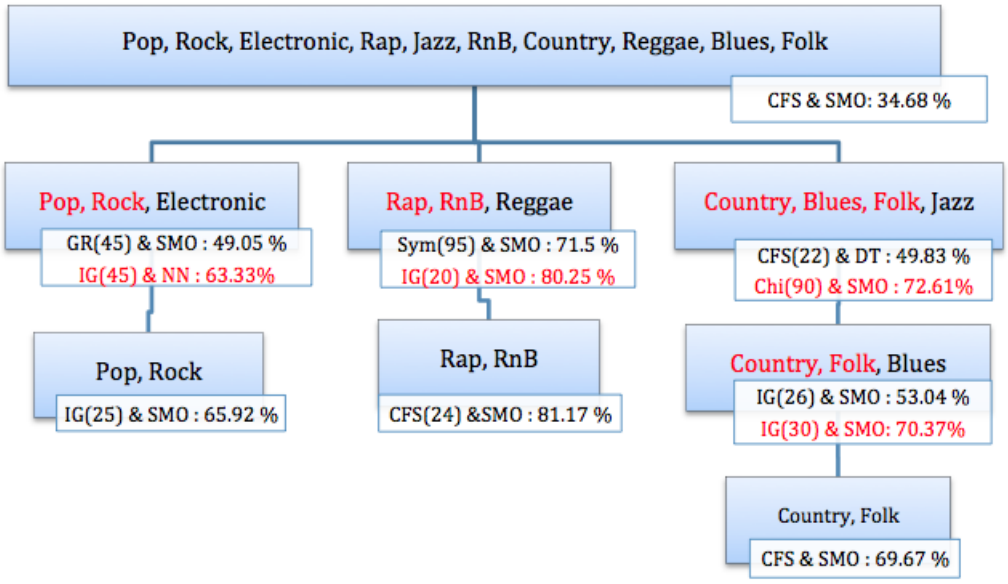

Figure 4.11: The Hierarchical Tree of Last.fm Dataset Accuracies.

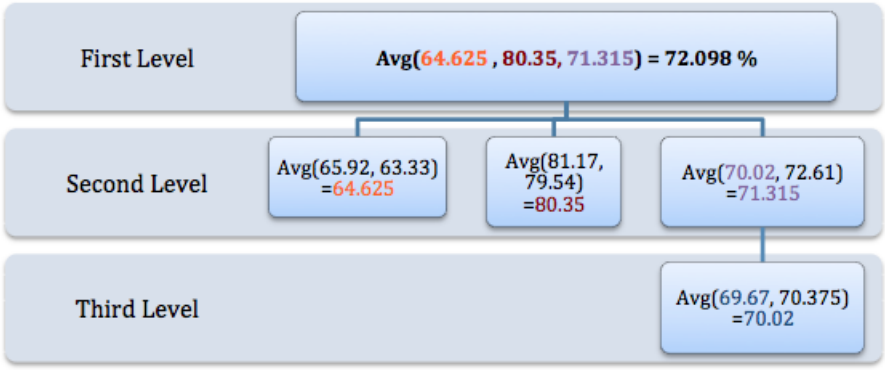

Figure 4.12: The Averages of Last.fm Dataset Accuracies.

the features and classification techniques suitable for each of them.

In this section, music style classification is investigated and discussed from different perspectives using our proposed approach. Several experiments have been conducted to classify the selected AllMusic styles. Some of these experiments were discussed in Chapter 3. In this chapter, the effects of using different classifiers, feature selection techniques, and also using individual feature sets or combining them are investigated. Furthermore, 10 styles in AllMusic were involved in this study (see Table 4.16).

Table 4.15: Comparison Percentage of Different Feature Sets for Each Group of Last.fm Datasets.
Group A: {Pop, Rock, Electronic}, Group B: {Rap, RnB, Reggae}, Group C: {Country, Blues, Folk, Jazz}, Group AA: {Pop, Rock}, Group BA: {Rap, RnB}, Group CA: {Country, Folk, Blues}, and Group CAA: {Country, Folk}

Name of Group	Low-level (%)	MFCC (%)	Rhythm Histogram (%)	Total Num
AllGenre	31	45	7	29
A, B, C	31	45	24	24
A	31	33	35	45
B	17	23	60	95
C	14	64	21	14
AA	24	68	8	25
BA	17	42	37	80
CA	23	43	33	30
CAA	18	22	59	22

Table 4.16: The List of Styles for Music Style Classification.

Name	List of Classes
AllMuisc-Styles	Rock College, Rock Contemporary, Hip Hop Rap, Dance, Pop Indie, Rock Hard, Metal Alternative, Pop Contemporary, Rock Alternative, Experimental

Table 4.17 shows the datasets which were used to experiment music style classification. All of them are balanced datasets, and they were created based on the selected feature sets from an individual feature set or multiple feature set (see the last column of Table 4.17).

Analyzing Results of Music Style Classification

The datasets of this experiments were created using three individual selected feature sets and a combination of them. These datasets were employed to analyze the performance of five feature selection methods and classifiers. Moreover, as in the previous experiments, each experiment was repeated several times to select various numbers of the highest ranked features. In each step of these experiments, the similar styles were grouped, and

Table 4.17: The Descriptive Details of Datasets for Music Style Classification.

Name	Num of Instances	Num of Features	Description
AllMusic-MFCC-Sty	6000	26	MFCC-MSD& AllMuisc-Styles
AllMusic-Low-Sty	6000	16	Low-level-MSD& AllMusic-Styles
AllMusic-Rhy-Sty	6000	60	Rhythm-Histogram& AllMusic-Styles
AllMusic-All-Sty	6000	102	Low-level-MSD& MFCC-MSD& Rhythm-Histogram& AllMusic-Styles

each one was classified separately. After each step, the number of styles in each group was considered. If there were more than two styles, they were split into different categories based on their similarity. To identify similar styles, the confusion matrix was used. This process continued until two styles were contained in all subgroups (see Algorithm 1).
The first experiment employed the dataset AllMusic-MFCC-Sty. Figures 4.13 and 4.14 indicate which styles were categorized into the same group, and also the best accuracy obtained for each group.
The highest accuracies were obtained with SMO and IG. It is remarkable to note that the numbers of features, which gained the highest accuracies for different styles, were different. NB and BN with IG obtained the same accuracies for two groups: *Rock College & Experimental* and *Pop Indie & Rock Alternative*. Thus, both of them were chosen as the best ones. Another important observation is that using our approach had significantly improved the performance accuracy of music style classification. The accuracy was increased from around 27.2% to 67.04%. By comparing the accuracies of using MFCC features for genre set and style set, it is easy to see that the final accuracies of music genre classification were higher than the ones of music style classification.
The AllMusic-Low-Sty Dataset was used in the next experiment to evaluate the performance of using Low-level features. Several classifiers and feature selection

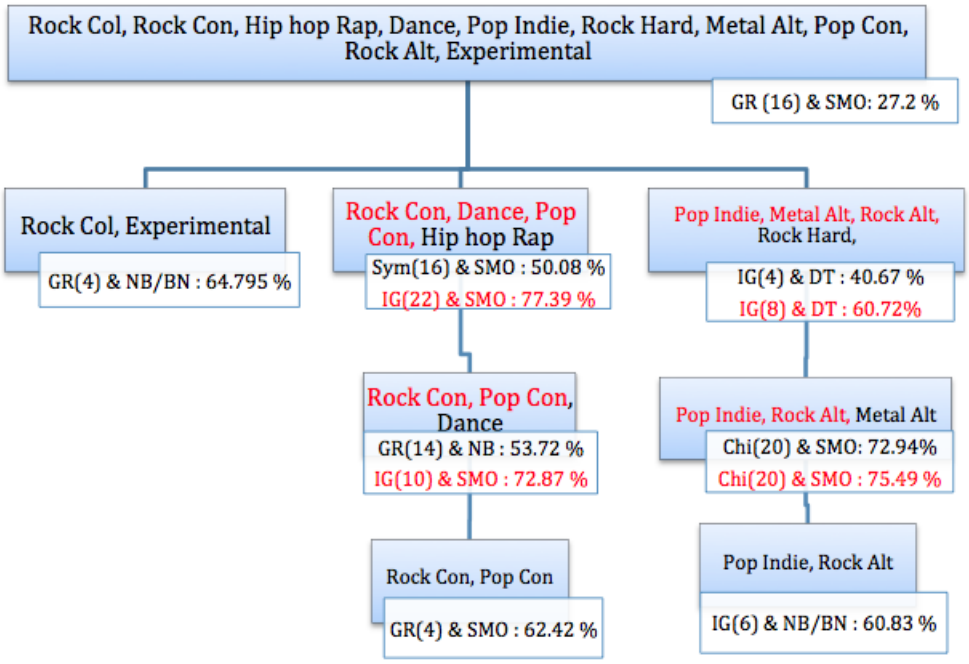

Figure 4.13: The Hierarchical Tree of AllMusic-MFCC-Sty Dataset Accuracies.

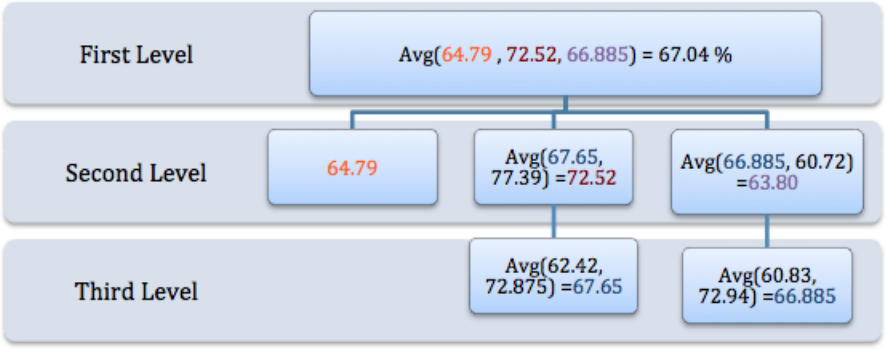

Figure 4.14: The Averages of AllMusic-MFCC-Sty Dataset Accuracies.

techniques resulted in the best accuracy, as shown in Figure 4.15. Moreover, *Rock College & Rock Alternative* were grouped together, and no classifiers and feature selection methods could not classify them correctly. All instances were classified as *Rock College*. It is clear that those Low-level features could not be an appropriate feature set for representing these styles.

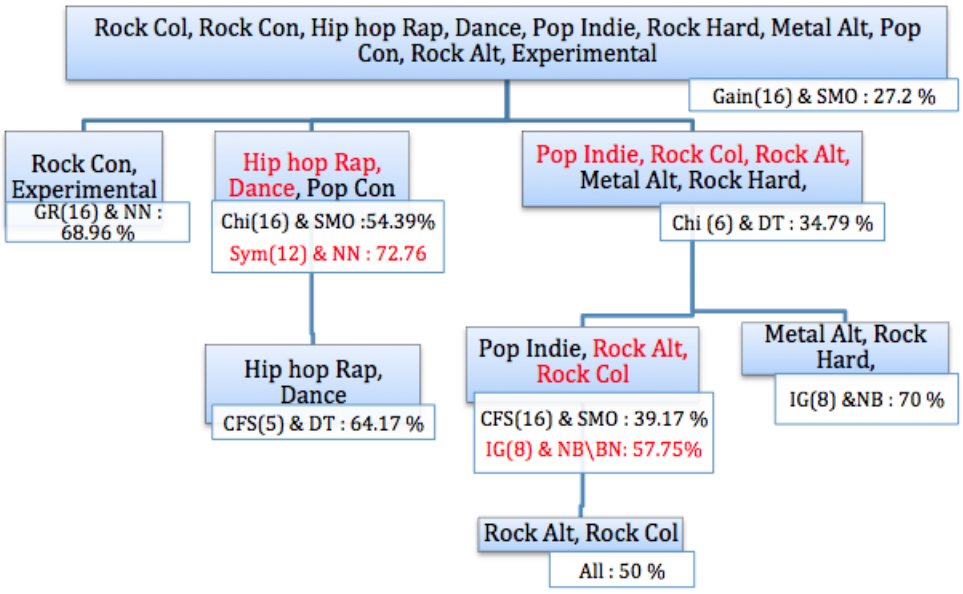

Figure 4.15: The Hierarchical Tree of AllMusic-Low-Sty Dataset Accuracies.

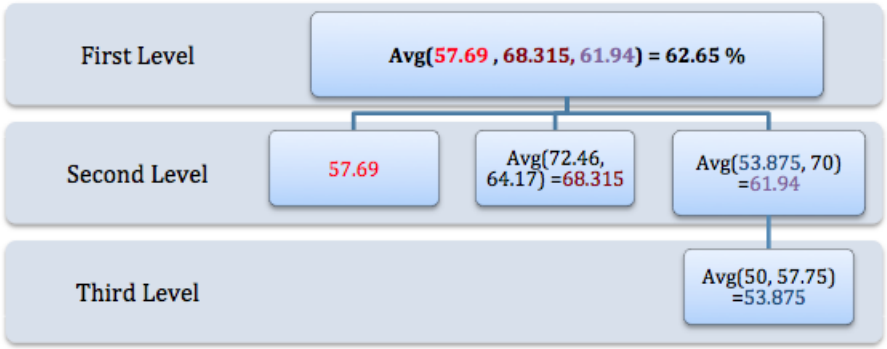

Figure 4.16: The Averages of AllMusic-Low-Sty Dataset Accuracies.

In Figure 4.15, the averages of the accuracies of classifying groups and subgroups are shown. As in the previous experiments, similar style groups were classified with appropriate results, and then they were classified to achieve a better output.

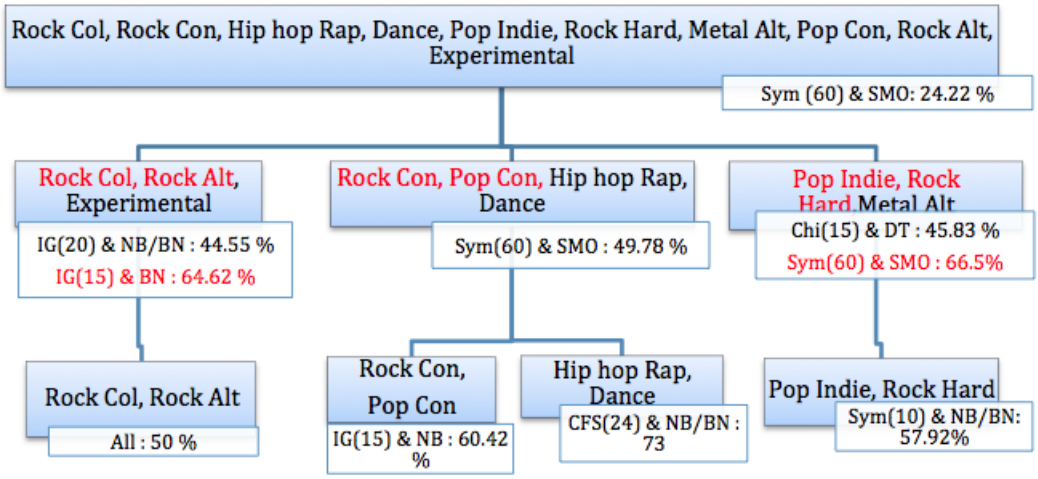

Figure 4.17: The Hierarchical Tree of AllMusic-Rhy-Sty Dataset Accuracies.

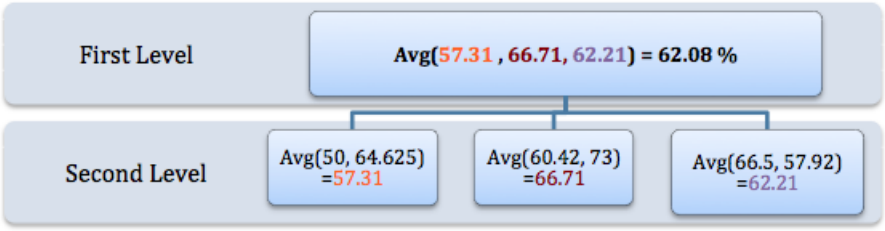

Figure 4.18: The Averages of AllMusic-Rhy-Sty Dataset Accuracies.

Figures 4.17 and 4.18 show the averages of the grouped styles and the hierarchical tree of accuracies of using AllMusic-Rhy-Sty Dataset respectively.

The confusion matrix of all experiments with the selected classifiers and feature selection techniques indicate that any classifier could not identify *Rock College & Rock Alternative*, the same as in the previous experiment. All of them were tagged as *Rock College* (see Table 4.18). The results show that the selected feature sets could not represent enough information; thus, the classifiers could not identify the instances of *Rock Alternative* class correctly. In total, there is not a single classifier with a feature selection method which could produce good results for all groups and subgroups.

Table 4.18: The Confusion Matrix of Rock College & Rock Alternative.

-	Rock College	Rock Alternative
Rock College	1200	0
Rock Alternative	1200	0

The last experiment for evaluating music style classification involved the combination of the three feature sets. Figure 4.20 indicates that the accuracy of classifying all styles was around 31.75%. In the deep level of the hierarchical tree, the classifiers performed better, and the accuracies were more than 60%. The hierarchical tree of this experiment shows that the final accuracy was around 74.55%. By combining the features, both approaches achieved better results compared to using the individual features.

4.4 Summary

In the previous chapters, a series of experiments were conducted to discover the appropriate setups for music classification. The experiments indicate that using those setups is an essential task. Therefore, all experiments in this chapter were implemented based on them. Several experiments were carried out to increase the performance of music classification techniques based on genre and style. To validate the results, different datasets were used. The first approach involved some classes which appeared to be suitable ones. Thus, its purpose was to improve the accuracy by the subset(s) of classes. The outcomes show that not only the members of class sets have effects on the performance, but also the number of them directly affects music classification accuracy. However, this approach was not complete because it worked with some classes, but not all of them.
The second approach has been suggested based on using all classes. The proposed approach was based on the similarity among classes. To the best of our knowledge, this is the first study that aims to categorize similar classes and to evaluate them separately. In our study, different datasets were evaluated, not only to improve music classification

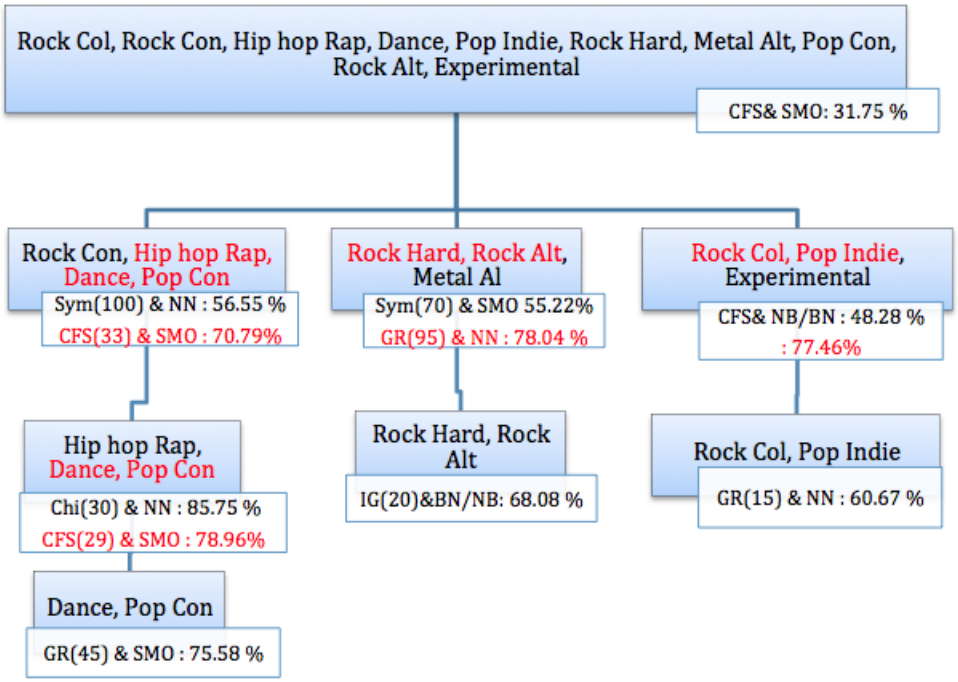

Figure 4.19: The Hierarchical Tree of AllMusic-All-Sty Dataset Accuracies.

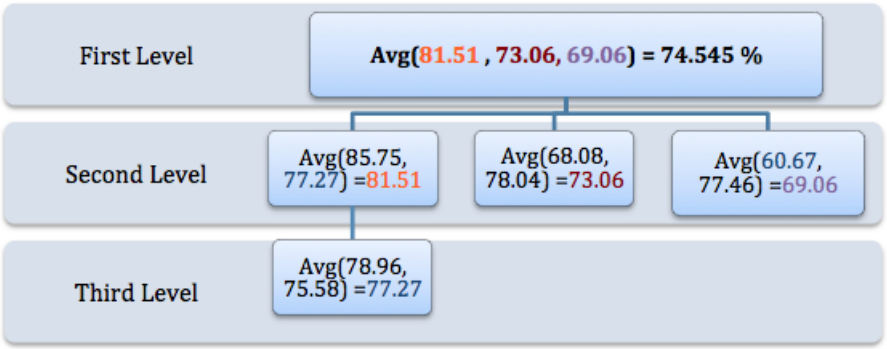

Figure 4.20: The Averages of AllMusic-All-Sty Dataset Accuracies.

performance, but also to overcome the limitation of classifying all classes. Furthermore, the hierarchical tree of each dataset shows how classes were categorized into different groups. In the low level of the tree, it is clear that the classifier could perform better with a low number of classes. In addition, the hierarchical tree can be used for any new input. The classifier is used to tag a new piece of music. After tagging a new piece of music, the place of it is identified by checking the place of that class in the hierarchical tree.

In Chapter 3, the combination of the selected feature sets, which covered different perspectives of musical details to present each track, provided better information for

classifying datasets. The combination of feature sets could improve music classification accuracy. Furthermore, to test the validity of our approach, some classifiers and feature selection techniques were used. Also, it is possible to compare the results and choose the appropriate techniques. Overall, SMO demonstrated a good performance for most classifications. As feature selection techniques, CFS and IG performed better than other selected techniques.

Another important observation is that different classes could be represented by different features, and also by a different number of features. Moreover, all results of the different experiments indicate that using the MFCC features were better than using other features to classify musical genres and styles.

Chapter 5

Conclusion

One task of Music Information Retrieval (MIR) is to classify large musical datasets. Music classification is a complex task in MIR. It has wide applications in managing large music repositories, such as organizing, browsing and categorizing music. Music classification based on genre or style is a subjective task. There is no a specific standard approach to it. Different people might have various opinions regarding the musical genre of a track, due to historical background, cultural diversity, and personal experience. This not only indicates that manual music classification is inadequate, but it also shows that automatic music classification is needed. In this thesis, automatic music classification based on musical genre and style is investigated from perceptual and algorithmic perspectives.

5.1 Our Contribution

The work presented in this thesis focuses on improving the music classification accuracy by using different data mining techniques. In addition, different approaches were investigated to reduce the confusion of classifiers. First, some work was done in order to provide a suitable setup for further experiments. Thus, different techniques, such as different discretization techniques, feature selection techniques, and classifiers learning, were used to evaluate music classification from different perspectives to compare the results and to choose the best techniques. Furthermore, two novel approaches were proposed toward this end. Several experiments explained some datasets based on different musical genres, styles, and also features. The suggested approaches used the confusion

matrix to identify similar and different classes.

5.1.1 Investigating Different Approaches

Different parameters from various perspetives were considered to provide the better setup for experiments. Some experiments were designed to compare the performance of supervised and unsupervised discretization techniques used to classify music. In addition, the dataset is one of the important considerations in music classification. Thus, a suitable dataset has a significant effect on the classifier performance. Another important component in music classification is features. Not only were some feature sets chosen to recognize music classification, but also some feature selection techniques were selected to rank the appropriate features. Thus, a classifier could use the high-rank features to obtain better results, compared to using all features. To evaluate the effectiveness of using individual feature sets or the combination of them, a set of empirical studies had been designed and implemented. In Chapter 3, the results of these experiments were not accurate. Some problems were identified. Moreover, a large number of classes (genre or style) obviously decreases the accuracy of the classifier. Thus, other approaches were suggested to solve the problems and to improve the performance of music classification techniques.

5.1.2 Improving the Performance of Music Classification

Two approaches were proposed to improve the accuracy of music classification. The first one was to remove some ineffective classes, which confused the classifier. Thus, the size of the class set was changed. The new class set involved the remaining classes, which were more different from each other. It was obvious that there was a considerable improvement by removing the similar classes. However, this approach was not useful when all classes had to be classified. The second approach grouped similar genres, and then each group was classified separately. The hierarchical tree of each experiment was created to show each similar class based on the selected features. To evaluate the effectiveness of the method, several experiments were designed and investigated.

Moreover, different datasets were used to validate the approach from several perspectives. It is interesting to note that the suggested approach improved accuracy considerably. Furthermore, we have done some experiments to investigate the effects of using the combination of feature sets and individual feature sets, on the performance of using the proposed approach. To test their validity, different datasets based on several feature sets and classes (genre or style) were used. Moreover, each experiment was repeated several times to compare the effect of using different numbers of the highest rank features on the accuracy of the classifier. In all experiments, the performance of the selected classifiers and feature selection techniques was considered.

5.2 Future Work

For our future work, more case studies will be done involving other feature sets. Another idea is not only to use the confusion matrix to identify similar (or different) classes, but also to use musical knowledge for categorizing them. Moreover, different types of music classification problems, such as music mood classification, can be used in these approaches to evaluate their performance. Some feature selection techniques and classifiers were chosen in this study. Other techniques can be investigated to find the best ones. There are several feature benchmarks, which are available to create the dataset. Choosing other feature sets could be useful to compare the accuracy of music classification, and also to evaluate the effect of using other feature sets on the music classification performance. As an example, using association rules can be a suitable approach to finding the relation among features in order to choose the appropriate features for classifying musical instances.

Bibliography

[1] Ariyaratne, Hasitha Bimsara; Zhang, Dengsheng, and Lu, Guojun. A class centric feature and classifier ensemble selection approach for music genre classification. In *Structural, Syntactic, and Statistical Pattern Recognition*, pages 666–674. Springer, 2012.

[2] Aryafar, Kamelia; Jafarpour, Sina, and Shokoufandeh, Ali. Automatic musical genre classification using sparsity-eager support vector machines. In *The Institute of Electrical and Electronics Engineers (IEEE): Proceedings of the 21st International Conference on Pattern Recognition (ICPR)*, pages 1526–1529. IEEE, 2012.

[3] Aucouturier, Jean-Julien and Pampalk, Elias. Introduction–from genres to tags: A little epistemology of music information retrieval research. *Journal of New Music Research*, 37(2):87–92, 2008.

[4] Barbedo, Jayme Garcia Arnal and Lopes, Amauri. Automatic genre classification of musical signals. *EURASIP Journal on Applied Signal Processing*, 2007(1):157–157, 2007.

[5] Basili, Roberto; Serafini, Alfredo, and Stellato, Armando. Classification of musical genre: a machine learning approach. In *International Society for Music Information Retrieval*, 2004.

[6] Bertin-Mahieux, Thierry; Ellis, Daniel PW; Whitman, Brian, and Lamere, Paul. The million song dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference, Miami, Florida*, pages 591–596, 2011.

[7] Bogdanov, Dmitry; Serra, Joan; Wack, Nicolas, and Herrera, Perfecto. From low-level to high-level: Comparative study of music similarity measures. In *The Institute of Electrical and Electronics Engineers (IEEE): de leoProceedings of the 11th International Symposium on Multimedia*, pages 453–458. IEEE, 2009.

[8] Cataltepe, Zehra; Yaslan, Yusuf, and Sonmez, Abdullah. Music genre classification using midi and audio features. *EURASIP Journal on Applied Signal Processing*, 2007(1):150–150, 2007.

[9] Chang, Kaichun K; Jang, Jyh-Shing Roger, and Iliopoulos, Costas S. Music genre classification via compressive sampling. In *International Society for Music Information Retrieval*, pages 387–392, 2010.

[10] Craft, Alastair JD; Wiggins, Geraint A, and Crawford, Tim. How many beans make five? the consensus problem in music-genre classification and a new evaluation

method for single-genre categorisation systems. In *International Society for Music Information Retrieval*, pages 73–76, 2007.

[11] Doraisamy, Shyamala; Golzari, Shahram; Mohd, Noris; Sulaiman, Md Nasir, and Udzir, Nur Izura. A study on feature selection and classification techniques for automatic genre classification of traditional malay music. In *International Society for Music Information Retrieval*, pages 331–336, 2008.

[12] Dougherty, James; Kohavi, Ron; Sahami, Mehran, and others, . Supervised and unsupervised discretization of continuous features. In *The International Conference on Machine Learning*, pages 194–202, 1995.

[13] Essid, Slim; Richard, Gaël, and David, Bertrand. Hierarchical classification of musical instruments on solo recordings. In *The Institute of Electrical and Electronics Engineers (IEEE) International Conference on Acoustics, Speech and Signal Processing*, volume 5, pages 817–820. IEEE, 2006.

[14] Flexer, Arthur; Gouyon, Fabien; Dixon, Simon, and Widmer, Gerhard. Probabilistic combination of features for music classification. In *International Society for Music Information Retrieval*, pages 111–114, 2006.

[15] Gjerdingen, Robert O and Perrott, David. Scanning the dial: The rapid recognition of music genres. *Journal of New Music Research*, 37(2):93–100, 2008.

[16] Hall, Mark; Frank, Eibe; Holmes, Geoffrey; Pfahringer, Bernhard; Reutemann, Peter, and Witten, Ian H. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[17] Hall, Mark A and Smith, Lloyd A. Practical feature subset selection for machine learning. pages 181–191, 1998.

[18] Han, Jiawei; Kamber, Micheline, and Pei, Jian. *Data mining: concepts and techniques*. Morgan kaufmann, 2006.

[19] Herrera, Perfecto; Amatriain, Xavier; Batlle, Eloi, and Serra, Xavier. Towards instrument segmentation for music content description: a critical review of instrument classification techniques. In *International symposium on music information retrieval*, volume 290, pages 23–25, 2000.

[20] Hu, Yajie; Li, Dingding, and Ogihara, Mitsunori. Evaluation on feature importance for favorite song detection. In *Information Society for Music Information Retrieval*, pages 323–328, 2013.

[21] Jun, Sanghoon and Hwang, Eenjun. Music segmentation and summarization based on self-similarity matrix. In *Proceedings of the 7th international conference on ubiquitous information management and communication*, pages 82:1–4. ACM, 2013.

[22] Kaminskas, Marius and Ricci, Francesco. Contextual music information retrieval and recommendation: State of the art and challenges. *Computer Science Review*, 6 (2):89–119, 2012.

[23] Kirthika, P and Chattamvelli, Rajan. A review of raga based music classification and music information retrieval (mir). In *Innovative Practices and Future Trends (AICERA), The Institute of Electrical and Electronics Engineers (IEEE) International Conference on Engineering Education*, pages 1–5. IEEE, 2012.

[24] Kofod, Christian and Ortiz-Arroyo, Daniel. Exploring the design space of symbolic music genre classification using data mining techniques. In *The Institute of Electrical and Electronics Engineers (IEEE) International Conference on Computational Intelligence for Modelling Control & Automation*, pages 43–48. IEEE, 2008.

[25] Li, Tao; Ogihara, Mitsunori, and Tzanetakis, George. *Music data mining*. CRC Press, 2011.

[26] Lidy, Thomas; Rauber, Andreas; Pertusa, Antonio, and Quereda, José Manuel Iñesta. Improving genre classification by combination of audio and symbolic descriptors using a transcription systems. In *Information Society for Music Information Retrieval*, pages 61–66. Citeseer, 2007.

[27] Lidy, Thomas; Silla Jr, Carlos N; Cornelis, Olmo; Gouyon, Fabien; Rauber, Andreas; Kaestner, Celso AA, and Koerich, Alessandro L. 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.

[28] Lim, S-C; Lee, J-S; Jang, S-J; Lee, S-P, and Kim, Moo Young. Music-genre classification system based on spectro-temporal features and feature selection. *The Institute of Electrical and Electronics Engineers (IEEE) Transactions on Consumer Electronics*, 58(4):1262–1268, 2012.

[29] Lippens, Stefaan; Martens, Jean-Pierre, and De Mulder, Tom. A comparison of human and automatic musical genre classification. In *The Institute of Electrical and Electronics Engineers International Conference on Acoustics: Speech, and Signal Processing*, volume 4, pages iv233–iv236. The Institute of Electrical and Electronics Engineers, 2004.

[30] Lopes, Miguel; Gouyon, Fabien; Koerich, Alessandro L, and Oliveira, Luiz ES. Selection of training instances for music genre classification. In *The Institute of Electrical and Electronics Engineers (IEEE): Proceedings of the 20th International Conference on Pattern Recognition (ICPR)*, pages 4569–4572. IEEE, 2010.

[31] Lud, Marcus-Christopher and Widmer, Gerhard. Relative unsupervised discretization for association rule mining. In *Principles of data mining and knowledge discovery*, pages 148–158. 2000.

[32] Mandel, Michael I and Ellis, Daniel PW. Multiple-instance learning for music information retrieval. In *Proceedings of the 9th International Conference of Music Information Retrieval*, pages 577–582. D rexel University, 2008.

[33] McKay, Cory. *Automatic music classification with jMIR*. PhD thesis, 2010.

[34] McKay, Cory and Fujinaga, Ichiro. Automatic genre classification using large high-level musical feature sets. In *Information Society for Music Information Retrieval*, volume 2004, pages 525–530, 2004.

[35] McKay, Cory and Fujinaga, Ichiro. jsymbolic: A feature extractor for midi files. In *Proceedings of the International Computer Music Conference*, pages 302–305, 2006.

[36] McKay, Cory and Fujinaga, Ichiro. Musical genre classification: Is it worth pursuing and how can it be improved? In *International Society for Music Information Retrieval*, pages 101–106, 2006.

[37] McKay, Cory and Fujinaga, Ichiro. Combining features extracted from audio, symbolic and cultural sources. In *Information Society for Music Information Retrieval*, pages 597–602. Citeseer, 2008.

[38] McKay, Cory and Fujinaga, Ichiro. Improving automatic music classification performance by extracting features from different types of data. In *Proceedings of the international conference on Multimedia information retrieval*, pages 257–266. ACM, 2010.

[39] McVicar, Matt and De Bie, Tijl. Cca and a multi-way extension for investigating common components between audio, lyrics and tags. In *The Institute of Electrical and Electronics Engineers (IEEE): Proceedings of the 9th international symposium on computational music modeling and retrieval (CMMR)*, pages 53–68, 2012.

[40] Meng, Anders; Ahrendt, Peter, and Larsen, Jan. Improving music genre classification by short time feature integration. In *The Institute of Electrical and Electronics Engineers(IEEE) International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages v–497. IEEE, 2005.

[41] Nagavi, T.C. and Bhajantri, N.U. Overview of automatic indian music information recognition, classification and retrieval systems. In *International Conference on Recent Trends in Information Systems (ReTIS)*, pages 111–116, Dec 2011.

[42] Ness, Steven R; Theocharis, Anthony; Tzanetakis, George, and Martins, Luis Gustavo. Improving automatic music tag annotation using stacked generalization of probabilistic svm outputs. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 705–708. ACM, 2009.

[43] Norowi, Noris Mohd; Doraisamy, Shyamala, and Wirza, Rahmita. Factors affecting automatic genre classification: an investigation incorporating non-western musical forms. In *Proceedings of the International Conference on Music Information Retrieval*, pages 13–20. Citeseer, 2005.

[44] Priyadarsini, R Praveena; Valarmathi, ML, and Sivakumari, S. Gain ratio based feature selection method for privacy preservation. 1(4):201–205, 2011.

[45] Ren, Jia-Min; Chen, Zhi-Sheng, and Jang, J-SR. On the use of sequential patterns mining as temporal features for music genre classification. In *The Institute of Electrical and Electronics Engineers (IEEE): Proceeding of The International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 2294–2297. IEEE, 2010.

[46] Sanden, Chris and Zhang, John Z. Enhancing multi-label music genre classification through ensemble techniques. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 705–714. ACM, 2011.

[47] Sanden, Chris; Befus, Chad R, and Zhang, John Z. A perceptual study on music segmentation and genre classification. *Journal of New Music Research*, 41(3): 277–293, 2012.

[48] Scaringella, Nicolas; Zoia, Giorgio, and Mlynek, Daniel. Automatic genre classification of music content: a survey. *The Institute of Electrical and Electronics Engineers (IEEE) Signal Processing Magazine*, 23(2):133–141, 2006.

[49] Schindler, Alexander; Mayer, Rudolf, and Rauber, Andreas. Facilitating comprehensive benchmarking experiments on the million song dataset. In *Information Society for Music Information Retrieval*, pages 469–474, 2012.

[50] Shawe-Taylor, JS and Meng, Anders. An investigation of feature models for music genre classification using the support vector classifier. pages 604–609, 2005.

[51] Silla, Carlos N and Freitas, Alex Alves. Novel top-down approaches for hierarchical classification and their application to automatic music genre classification. In *The Institute of Electrical and Electronics Engineers (IEEE) International Conference on Systems, Man and Cybernetics*, pages 3499–3504. IEEE, 2009.

[52] Silla, Carlos N; Kaestner, Celso AA, and Koerich, Alessandro L. Automatic music genre classification using ensemble of classifiers. In *The Institute of Electrical and Electronics Engineers (IEEE) International Conference on Systems, Man and Cybernetics*, pages 1687–1692. IEEE, 2007.

[53] Silla, CN; Koerich, Alessandro L, and Kaestner, Celso AA. Feature selection in automatic music genre classification. In *The Institute of Electrical and Electronics Engineers (IEEE): Proceedings of the 10th International Symposium on Multimedia*, pages 39–44. IEEE, 2008.

[54] Silla Jr, Carlos N; Koerich, Alessandro L, and Kaestner, Celso AA. A machine learning approach to automatic music genre classification. *Journal of the Brazilian Computer Society*, 14(3):7–18, 2008.

[55] Silla Jr, Carlos Nascimento; Koerich, Alessandro L, and Kaestner, Celso AA. The latin music database. In *Information Society for Music Information Retrieval*, pages 451–456, 2008.

[56] Soares, Caio; Williams, Philicity; Gilbert, Juan E, and Dozier, Gerry. A class-specific ensemble feature selection approach for classification problems. In *Proceedings of the 48th Annual Southeast Regional Conference*, page 33. ACM, 2010.

[57] Tzanetakis, George and Cook, Perry. Musical genre classification of audio signals. *The Institute of Electrical and Electronics Engineers (IEEE) Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

[58] Wang, Dingding; Li, Tao, and Ogihara, Mitsunori. Are tags better than audio features? the effect of joint use of tags and audio content features for artistic style clustering. In *The 11th International Society on Music Information Retrieval Conference, number Information Society for Music Information Retrieval*, pages 57–62, 2010.

[59] Wang, Fei; Wang, Xin; Shao, Bo; Li, Tao, and Ogihara, Mitsunori. Tag integrated multi-label music style classification with hypergraph. In *International Society for Music Information Retrieval*, pages 363–368, 2009.

[60] Witten, Ian H and Frank, Eibe. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[61] Xu, Changsheng; Maddage, MC; Shao, Xi; Cao, Fang, and Tian, Qi. Musical genre classification using support vector machines. In *The Institute of Electrical and Electronics Engineers (IEEE) International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages V429 – V432. IEEE, 2003.

[62] Xu, Changsheng; Maddage, MC, and Shao, Xi. Automatic music classification and summarization. *The Institute of Electrical and Electronics Engineers (IEEE)Transactions on Speech and Audio Processing*, 13(3):441–450, 2005.

[63] Yang, Yiming and Pedersen, Jan O. A comparative study on feature selection in text categorization. In *The International Conference on Machine Learning*, volume 97, pages 412–420, 1997.

[64] Yaslan, Yusuf and Cataltepe, Zehra. Audio music genre classification using different classifiers and feature selection methods. In *The Institute of Electrical and Electronics Engineers (IEEE): Proceedings of the 18th International Conference on Pattern Recognition*, volume 2, pages 573–576. IEEE, 2006.

[65] Yu, Lei and Liu, Huan. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *The International Conference on Machine Learning*, volume 3, pages 856–863, 2003.

[66] Zhang, Yibin and Zhou, Jie. A study on content-based music classification. In *The Institute of Electrical and Electronics Engineers (IEEE): Proceedings of the 7th International Symposium on Signal Processing and Its Applications*, volume 2, pages 113–116. IEEE, 2003.

[67] Zhang, Yibin and Zhou, Jie. Audio segmentation based on multi-scale audio classification. In *The Institute of Electrical and Electronics Engineers (IEEE) International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages iv–349 – iv–352. IEEE, 2004.

[68] Zhou, Yatong; Zhang, Taiyi, and Sun, Jiancheng. Music style classification with a novel bayesian model. In *Advanced Data Mining and Applications*, pages 150–156. Springer, 2006.