ANALYZING AND ENHANCING MUSIC MOOD CLASSIFICATION: AN EMPIRICAL STUDY

HOUMAN SHAHMANSOURI
Master of Computer Science, University of Yazd, Yazd, Iran, 2014

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 ENHANCING MUSIC MOOD CLASSIFICATION: AN EMPIRICAL STUDY

HOUMAN SHAHMANSOURI

Date of Defense: December 20, 2016

Dr. John Zhang Supervisor	Associate Professor	Ph.D.
Dr. Gongbing Shan Co-Supervisor	Professor	Ph.D.
Dr. Howard Cheng Committee Member	Associate Professor	Ph.D.
Dr. Wendy Osborn Committee Member	Associate Professor	Ph.D.
Dr. Amir Akbary Chair, Thesis Examination Committee	Professor	Ph.D.

Dedication

To my beloved family.

Abstract

In the computer age, managing large data repositories is one of the common challenges, especially for music data. Categorizing, manipulating, and refining music tracks are among the most complex tasks in Music Information Retrieval (MIR). Classification is one of the core functions in MIR, which classifies music data from different perspectives, from genre to instrument to mood. The primary focus of this study is on music mood classification. Mood is a subjective phenomenon in MIR, which involves different considerations, such as psychology, musicology, culture, and social behavior. One of the most significant prerequisitions in music mood classification is answering these questions: what combination of acoustic features helps us to improve the accuracy of classification in this area? What type of classifiers is appropriate in music mood classification? How can we increase the accuracy of music mood classification using several classifiers?

To find the answers to these questions, we empirically explored different acoustic features and classification schemes on the mood classification in music data. Also, we found the two approaches to use several classifiers simultaneously to classify music tracks using mood labels automatically. These methods contain two voting procedures; namely, Plurality Voting and Borda Count. These approaches are categorized into ensemble techniques, which combine a group of classifiers to reach better accuracy. The proposed ensemble methods are implemented and verified through empirical experiments. The results of the experiments have shown that these proposed approaches could improve the accuracy of music mood classification.

Acknowledgments

I would like to thank everyone who contributed to different parts of this thesis. First and foremost, I thank my supervisor Dr. John Zhang for his patience, the continuous support of my M.Sc. study, and all of the chances I was given to guide my research and further my thesis. I must express my profound acknowledgment to my parents for their quick pieces of advice. You are always there for me. Thank you for your encouragements and supports. I would also like to thank my best friends, Zahra for her persuasions, Farshad, Hossein, and Tom for their help. You should know that your supports were worth more than I can express on paper.

Contents

List of Tables

Chapter 1

Introduction

Advanced technologies are used to facilitate many types of a person′s activity. There are a great number of significant benefits that new technologies are providing for us. In the information age, data plays an essential role in different types of everyday life. To manage a huge amount of data, various types of tools are needed to search, retrieve, and store them efficiently [4].

Music data, including albums and individual music tracks, are now stored in digital formats on the Internet. Music data are not exempted from the issues just mentioned. For example, one of the significant challenges is to categorize music tracks into important groups that help provide a faster identification of each music piece. To achieve this, Music Information Retrieval (MIR) was introduced by Stephen Downie [14] to approach some music related problems, such as classifying audio tracks via different concepts, providing new methods for music recommender systems, and transcripting music automatically.

MIR is an interdisciplinary subject which involves many fields, such as musicology, psychology, audio signal processing, and machine learning [38]. A considerable amount of research has been conducted in MIR during the past few decades; however, many problems remain. To solve the music related problems in MIR, many machine learning techniques, data mining tools, and artificial intelligence algorithms have been used [21].

One of the major challenges in MIR is related to music classification, which is to classify music tracks according to their types, such as mood, genre, style, and instrument. From the viewpoint of machine learning, a wide range of classifiers, feature selection techniques, and

also dimensionality reduction techniques are adopted to improve the classification accuracy in music classification.

Furthermore, extracting and refining information from music repositories are the responsibilities of data mining [21]. Different tools are employed to make this process easier and more precise. Some of these tools will be introduced in the next chapters. In addition, many concepts, algorithms, and formulas borrowed from artificial intelligence to improve the performance of methods in music data mining will be explored.

This study focuses on music classification, especially music mood classification. From now on we are using MMC as the abbreviation for Music Mood Classification.

1.1 Music Information Retrieval

Music Information Retrieval (MIR) consists of several facets, including pitch, temporal, harmonic, and timbral. The primary goal of MIR is to provide quicker access to extracted music pieces among a large number of tracks in an extended-sized audio repository or on the Internet. To achieve this goal, various approaches were presented, such as audio recommended systems and automatic tag annotation systems [38].

An important task in MIR is mood classification, which is more complex compared to other classification tasks [38]. The reason is that mood is a personal expression, containing psychological, behavioral, cultural, and other concepts.

1.2 Problems in MIR

MIR encompasses many areas related to music, such as musicology, cultural dependencies, computation of music, and music analysis [14]. Music classification is one of them, which is the focus of this study. Each piece of music may be categorized into different groups from various viewpoints. Moreover, each may contain information from which many acoustic features derive by audio feature extractors. Some of these features help classifiers predict an appropriate group for each piece of music. As an example, classifiers can

classify a piece of music based on its genre, mood, instrument, and lyrics.

In addition to the complexity of acoustic features, other elements like cultural, geographical, and psychological issues add to the complication of classification. An example of a cultural issue is that a specific musical style of a country may have a history in the style of that country. So the style does not have a standard or known property in the whole world [14]. The mood is another good example, which depends directly on the culture. Cultural influences play a prominent role in the mood classification. Geographical issues make the process of predicting instruments tough, because there are thousands of musical instruments in the world, and many of them are unknown and produce similar sounds. Therefore, distinguishing them is impossible in some cases.

In music classification, the primary concern is how to predict the correct tag or label for a piece of music. In genre classification, there is no single definition for each musical genre [30]. Musicologists have no consensus on the descriptions of each distinct type of genre. Therefore, classifying music tracks to different groups of genres is a hard task. One solution to handling this problem is defining a genre taxonomy, where different types of genre divide into general genres and sub-genres [4].

As the mood depends on some other intuitive phenomena, MMC becomes a more complicated task in the MIR. Suppose in an experiment that researchers asked listeners to listen to a piece of music and then choose a mood tag from among the categories *happy*, *sad*, *angry*, and *relax*, for that piece. If they have any memories of that piece of music, it definitely has effects on the choice of the tag. Also, the current mood of the listeners when listening to the music can have an effect such that each audience selects a different tag. These issues happen to casual listeners. Now assume that listeners have some music knowledge. They will decide to choose a label based on their music knowledge and also on their mood simultaneously. They will try to find the most relevant tag for that piece. In this situation, the combination of musical knowledge and psychological considerations have been used together to choose a proper label.

From the other side, using different acoustic features needs to be done carefully, since they can either improve the efficiency of classifiers or confuse them if irrelevant features have been selected. Among a large number of extracted acoustic features from each piece of music, some are adequate for classification. Feature selection techniques and dimensionality reduction methods may be adopted to determine the best features and eliminate redundant ones in a dataset.

1.3 Contributions

Improving the accuracy of MMC through an empirical study is one of the main objectives in this study. Several machine learning methods, classification algorithms, and pre-processing techniques are used. In pre-processing steps, some feature selection techniques and dimensionality reduction algorithms are used to select the most efficient acoustic features and reduce the complexity of the features$'$ space using removing redundancy and noisy data. As mentioned above, one of the biggest challenges for MMC in MIR is that moods are subjective and detecting mood labels correctly for songs is related to many factors. In other words, there are many items which have considerable effects on the performance of MMC. Therefore, the primary goal of this thesis is to find new ways to deal with automatic tag annotation for MMC. To this end, we empirically found what sort of acoustic features, classification algorithms, and feature selection (and dimensionality reduction) techniques can work better in MMC.

In the next step, in order to increase the accuracy of MMC, we focused on ensemble techniques. The heart of ensemble techniques is how to combine several classifiers and create a new one to obtain higher accuracy. This thesis focuses on two combination methods, both related to ensemble techniques, to increase the final classification accuracy in MMC:

- Combining several classifiers using Plurality Voting.
- Using Borda Count for combination of different classifiers.

1.4 Outline

The structure of the thesis is as follows. In Chapter 2, we review some previous works on acoustic features, feature selection and dimensionality reduction techniques, various types of classifiers, and genre and mood classification on music data. We also consider some evaluation techniques in machine learning to measure the efficiency of our proposed methods. In Chapter 3, we navigate through experiments to find out what combination of classifiers and feature sets work better in mood classification. Some feature selection techniques are implemented to choose the best features. In some cases, dimensionality reduction methods are performed to remove inappropriate features to improve the performance of classification. In Chapter 4, we present new approaches to improve the accuracy of mood classification. Moreover, different acoustic features are analyzed to show which ones are better in the mood classification. Some experiments are conducted based on the proposed methods to evaluate on different datasets. Chapter 5 includes the conclusion of the proposed approaches used in this study. Moreover, some tasks for future work will be presented.

Chapter 2

Background

The improvement of large music repositories allowed users to store billions of music tracks, although managing, accessing, and categorizing these tracks are still immense challenges for millions of listeners who have varying musical tastes. Music Information Retrieval (MIR) has become one of the critical research areas to solve the issues. Although many studies have been conducted in this area to find relevant models for music classification, they cannot solve all problems because there are more aspects to this area. One of the most important research fields in MIR is related to automatic tag annotation, which assigns a label for music tracks from varying perspectives, including emotion, instrument, genre, and style.

This study investigates MMC in MIR. In this inquiry, various data mining and machine learning methods are used to develop and evaluate the proposed approaches. The primary focus of this chapter is to discuss some previous work in music classification from different aspects, including acoustic features, classification techniques, music data mining tools, and machine learning algorithms. In addition, acoustic features, some well-known music repositories, feature selection techniques, classification algorithms, and WEKA -a data mining tool which implemented many machine learning algorithms- are introduced.

2.1 Acoustic Features

An audio feature extraction is a set of methods to extract meaningful information from audio signals [8]. These methods are the foundation of audio classification, which are used

in MIR. There are several types of audio feature extractors, such as Music Analysis, Retrieval and Synthesis for Audio Signals (MARSYAS) [1] and jAudio [2], that derive a broad range of acoustic features from pieces of music. These features are related to different properties of a music track, and some may introduce some basic factors of an audio signal. MARSYAS is an open source audio processing framework, which can extract audio features, especially low-level audio features [46, 53]. It is used in many studies, such as [42] where a new approach is proposed to classify Malay music based on the genre by WEKA classifiers. Also, the MARSYAS framework derives Mel-Frequency Cepstral Coefficients (MFCCs) and pitch, which are two types of acoustic features and will be introduced later in this chapter, in a study [87]. The MIR Toolbox is another kind of audio feature extractor, which extracts low-level and high-level audio features using Matlab [53]. jAudio is a favorite feature extractor framework, since it avoids duplication in deriving audio features from an audio signal [51].

2.1.1 Features based on Perspective

There are different approaches to categorize acoustic features based on their various aspects. One of these approaches classifies features into three groups: low-level features (e.g. spectral and Cepstrum), mid-level features (such as pitch and beats), and high-level features (including the style or genre, mood, and artist).

Low-level Acoustic Features

Low-level features extracted from raw audio signals may contain many noisy and redundant data. Spectral and Cepstrum are among the most common features of this level, which are used in many music classification works. The spectral is related to the brightness of sound and how it looks [68]. The Cepstrum is a transformation function, which shows the short-term power of a sound. It is "the inverse Fourier transform of the logarithm of the

[1] http://marsyas.info/.

[2] http://jmir.sourceforge.net/index_jAudio.html/.

Spectrum" [76]. The main reason for using these features is to analyze the human voice in pieces of music[83].

Mid-Level Acoustic Features

Pitches are regarded as mid-level features. In music terminology, the pitch is defined as the lowness or highness of the sound level in a music piece [27]. One important thing about the pitch is that it is a useful feature when combined with other acoustic features. Several previous studies have used pitch approaches [27, 86]. For example, pitch, timber, and the combination of Cepstral and temporal features were chosen to create a feature set, which was employed to classify music genre tags [86].

High-level Acoustic Features

Mood, genre and style are among some popular high-level features of a music track. Style and genre are two interchangeable descriptions, which are used to cover the same concept in some areas. However, they describe different concepts in music. The style is a combination of several items which bring forward a piece of music, such as melody and rhythm. There are some common features for each group of genre that help us to categorize peases of music in a dataset.

The mood is the expression of listeners$'$ emotions when they are listening to a piece of music. In a broader sense, psychological, social, emotional, logical, behavioral, cultural phenomena, and musical knowledge are the resources used to represent mood based on music aspects [25, 27, 36, 62, 66].

2.1.2 Features based on Acoustic

Another way of categorizing audio is using acoustic features. Acoustic features classify into three sub-groups, which are timbral, tonal and rhythmic [34].

Timbre Features

Timbre or *Tone Color* is the difference in the sound patterns of two or more musical instruments when they play together. In other words, humans distinguish different musical instruments because of timbre, when they are playing the same musical note, even with the same pitch and loudness [27]. Timbral features include MFCCs, pitch, loudness and Spectral.

MFCCs are one of the Spectral features, which are used to model music and audio, as well as speech [41]. Pitch specifies how high or low the frequency of a sound is and facilitates in determining one tone from another. Loudness indicates the dynamic levels of a sound. In music theory, it classifies into six energy groups for a musical piece: *very quiet*, *quiet*, *moderately quiet*, *moderately loud*, *loud*, and *very loud* [27, 34].

Tonal Features

A sound is tonic when it is the main sound, and other sounds rely on it. A tonic sound is the basis of tonality. Mode and key are classified as tonal features. The mode is related to two main groups of concepts in music terminology; namely, melody and scale [8]. Composers use two primary scales; namely, major and minor. Key (or key signature) in music means that all notes in a piece of music should play sharp or flat, depending on the key signature [8].

Rhythm Features

Rhythm is one of the most significant acoustic features which has important effects on representing mood [34]. Rhythm has some essential elements, such as duration, tempo, and danceability [27]. Duration is related to the period a song continues. Tempo is the number of beats that play in one minute. In other words, it shows the speed of playing beats per minute. For example, Blues, Pop, and Classical music have a slower tempo than R&B, Rock, and Rap music. Danceability means how easy a song is for dancing to. All these elements are helpful to predict the mood of a song. Feng et al. used some acoustic features

to determine the mood of a music piece by tempo and some other features [16]. The results showed that the tempo was an effective feature for distinguishing happiness and fear, or anger and sadness.

2.2 Music Benchmarks

One of the most significant challenges in MIR is related to music data. For instance, there are many music datasets on the Internet that contain low-quality data. There are two main issues in audio music repositories:

- Using extracted audio features may not be enough for current researches;
- The size of most music datasets is small, so those datasets are not comparable to real world data.

In addition, having a balanced dataset is important, since it has direct effects on the results of classifications. If a dataset is not balanced, the results may be biased toward those tags more repeated in the dataset. In other words, the training dataset may not contain enough detail about all music pieces of all tags.
With the growth of music science and music technologies, the number of music pieces has increased significantly [47]. Therefore, finding an appropriate dataset is a hard task in MIR. Million Song Dataset [3], Last.fm [4], and Latin Music Dataset [5] are some of the well-known ones, and will be introduced in the following section. Furthermore, as the focus of this study is mood classification, four of the most common mood tags in music-*Happy*, *Sad*, *Angry*, and *Relax*-were chosen to create some datasets.

2.2.1 Million Song Dataset

The Million Song Dataset (MSD) is one of the most popular benchmarks in MIR and contains audio features for approximately one million prominent songs. This large-sized

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

dataset also has metadata, including artist name, album name and song name [47]. The heart of this dataset consists of extracting audio features which are derived by Echo Nest [6] using some audio feature extractors, such as MARSYAS an jAudio [7]. These features are analyzed and have a huge effect on classifiers. The large-scale of MSD makes it a comparative dataset in the real -and commercial- world [7].

The MSD has three main categories of audio features, including pitches, timbre, and loudness [7]. Echo Nest also provides basic acoustic features, such as Mel-Frequency Cepstral Coefficients (MFCCs) and timber [7]. However, MSD suffers from other challenges, such as the lack of tags for each piece of music, which makes it hard to use effectively in wider areas of music data mining.

2.2.2 Last.fm Dataset

The Last.fm dataset has approximately 943,347 music tracks, matched with MSD with the same ID [52]. Music tracks in Last.fm have different types of tags, which include genre and mood [22]. One of the most important issues in Last.fm is related to the irrelevant tags for each piece of music. In fact, users are able to assign several tags to each music track in the dataset. It is a common problem that unacceptable tags, such as unrelated mood labels, exist in the dataset. However, having the same ID in both MSD and Last.fm is a positive viewpoint that helps researchers find more information about each music track. In this study, several datasets are created based on audio features of the MSD and mood-related tags of Last.fm.

2.2.3 Latin Music Database

The Latin Music Database (LMD), with more than 3,000 music songs, is one of the most accurate music databases in the area of music data mining. The main reason for its reputation is that some professional music instructors have worked on it, so the accuracy of

[6] http://the.echonest.com/.

this database is very high [71, 72]. Based on the Chorus website [7], which is a collaborative effort, there are three different datasets from LMD, including 30 seconds of the beginning of music tracks, 30 seconds of the middle of music tracks, and 30 seconds of the end of music tracks. Also, there are both training and testing datasets for each of the aforementioned time durations. All parts are derived by MARSYAS. In our study, the dataset which was created based on the middle part of music track was used in the experiments.

2.2.4 Computer Audition Lab 500 Dataset

Computer Audition Lab 500 (Cal500) is one of the most popular datasets in music tag annotation-related research [87]. The dataset involves 500 music tracks from 500 artists in the last 50 years. It has 1,708 annotations, including genre and mood, for each piece of music. Moreover, time series acoustic features, including MFCCs, are among the audio features in this dataset [78].

2.3 Data Pre-processing

After introducing various types of acoustic features, we find that not all of them are efficient for classifiers to find a label prediction model for a piece of music. The intuition is that they confuse the classifiers to pick a class correctly, because some features cannot represent enough characteristics of a music class in a music dataset. Therefore, feature selection techniques are applied to remove inappropriate, irrelevant, and redundant features.

2.3.1 Feature Selection Techniques

In this study, different feature selection techniques are chosen for music classification and these techniques are compared to find the best features. These techniques include Correlation-Based Feature Selection, Information Gain, Gain Ratio, Symmetric Uncertainty and Principal Component.

[7] http://www.avmediasearch.eu/wiki/index.php/Latin_Music_Database.

Correlation-Based Feature Selection

Correlation-Based Feature Selection (CFS) is one of the most popular feature selection techniques, which is used in many studies. This technique creates all possible subsets of features using a search method, such as Best-First or Greedy-Step Wise, and then attempts to find the best subset [20]. It evaluates the ability of prediction for each attribute and selects sets of attributes, which have the highest correlation with the class. Among the highly correlated sets of attributes, it chooses the ones with low intercorrelation [83]. In two studies [12, 31] CFS was used as the feature selection technique during the pre-processing steps. The results of both studies showed that the CFS technique could improve the performance of classifiers.

Greedy-Step Wise (GSW) selects features based on "Forward Selection and Backward Elimination" as mentioned in [83]. In this study, the GSW method is used as the search method for CFS.

Information Gain

Information Gain (IG) is another favorite feature selection technique. IG assesses how an individual attribute may have an effect on reducing the entropy [83]. It calculates the information gain by counting the number of bits of information for each acoustic feature in a music dataset to reduce the size of the data [86].

Gain Ratio

Gain Ratio (GR) is a modified version of IG, which is a ratio of IG to the essential and internal information [26]. GR computes the worth of each attribute using a normalized score of the IG of the attribute [83].

Symmetrical Uncertainty

Symmetrical Uncertainty (SYM) is a feature selection method which uses the correlation between features and classes. This approach investigates the relevance between fea-

tures and instance classes. Thus, those features which are not related to each class are not chosen by this approach [83].

Chi-Squared

The Chi-Squared method chooses the best features based on the statistic of a Chi-squared distribution with respect to its class. Features with higher Chi value are more important in sampling than the ones with a lower Chi value [88].

Some Previous Works on Feature Selection Techniques

Since the 1960s, Marill and Green [83] focused on a new subject that includes a selection of the proper features. After that, many researchers investigated different feature selection methods to choose the best features. Music datasets contain various features since audio feature extractors derive many acoustic features for each piece of music. However, some of them are irrelevant and/or redundant and could have a negative effect on the performance of music classification. Thus, choosing the proper feature selection method should have positive effects on improving the accuracy of classification. A criterion of a feature selection method is that relevant features should be selected to provide enough information for classifiers to recognize music pieces [83].

Feature selection techniques are categorized based on the search method they use to find the proper features. These search methods group into two principal categories: Wrapper and Filter methods. The filter methods, such as Information Gain and Gain Ratio, use a ranker as a search method [83]. The ranker method uses some statistical measures to assign a score to each feature and then choose a number of the highest ranked features [40]. The wrapper technique uses a learning method to evaluate the combination of features and then selects those combinations of features, which obtains the best outcomes [21]. By comparing wrapper and filter methods, the results showed that the filter method achieved better results compared to the wrapper one. However, a filter method is more expensive and is also slower than the wrapper method [38, 40, 83].

Various studies apply feature selection to improve the performance of their approaches. For example, Lopes et al. [42] used two methods, including choosing the best features and selecting features randomly to pick the proper features using LMD. In another study [12], CFS with Generic Search Strategy, as a feature selection method, obtained the best accuracies, where MLP was used to classify the instances. Ariyaratne et al. [2] investigated two approaches, which are named One-Vs-One and One-Vs-all. In the first approach, every two classes were grouped into one, and then the feature selection technique was applied to classify each group separately. In the second method, one class was classified against all classes. The first approach obtained better results compared to the second one. In another study [87], two feature selection techniques were applied using forward and backward methods separately. By using these techniques, there was consideration increase in the classification accuracy.

2.3.2 Dimensionality Reduction Technique

From a general viewpoint, dimensionality reduction techniques are used to decrease the amount of insufficient data, such as random data, redundant data, and noisy data [38]. These techniques transform the data into a principal component space to filter the less-relevant data by using a ranking search method. Partial Least Squares Regression and Principal Component are among the most well-known techniques in this area, which are implemented in WEKA.

Partial Least Squares Regression

The Partial Least Squares Regression (PLS) technique is a statistical one that uses a linear regression model to calculate the variance among available variables and predicated variables. It calculates the maximum variance of multidimensional direction in the $Space1$ space using multidimensional direction in the $Space2$ space [83]. It applies to the pre-processing step, before applying any machine learning methods. Moreover, PLS is a supervised method that improves the ability of predication by reducing the dimensions [3].

Principal Component

Principal Component (PC) is a linear statistical technique that uses principal component analysis to transform the input into a new coordinate space. For the features that are missing a value, PC fills in the lost value by calculating means. It also converts the nominal features with multiple values into binary attributes [21, 83]. The results of a study [20] show that the PC method could increase the accuracy considerably on high dimensional data.

2.3.3 Discretization Technique

Existing tags and labels in a dataset can create several issues for some machine learning algorithms. Thus, the discretization technique is used to convert nominal features, tags, and labels, to some metric ranges. In other words, the discretization technique is a method to decompose data and label features with some discrete ranges. This technique could decrease the processing time to obtain the accuracy [83].

2.4 Classification Algorithms

Every track of music is represented via some acoustic features, which are derived by different audio feature extractors. In general, the audio features categorize into various groups: continuous features, categorical features, and binary features. Moreover, machine learning methods group into two main sections, which are the supervised and unsupervised methods. The supervised techniques use the set of input and output data, which is used to create a predictive model. The predictive model is used to tag the new inputs, which are unlabeled. The unsupervised method measures unlabeled data based on objective similarity [32, 19, 21].

It is common to measure and evaluate the classifiers′ accuracy in order to classify instances of a dataset. One of the most popular evaluation methods in machine learning is called cross-validation. In this approach, a dataset is separated into two parts, training and testing, which is done several times randomly. In each iteration, a specific proportion of the data

is used to train the classifier, and the remainder is employed to test the trained classifier. The final error rate for the whole training process is calculated using the error rates of all iterations [83].

In this study, some learning methods are adopted to classify the dataset based on different specifications. For easy reference, the complete names and the corresponding abbreviations of used classifiers in this study are found in Table 2.1. All of these classifiers are also implemented in WEKA.

2.4.1 Naive Bayes

Naive Bayes (NB) is one of the supervised machine learning techniques that uses some statistical analysis of the training set. It creates maximum likelihood estimators and conditional probability, by using the details of tracks, including features and classes. Each feature has independent effects on prediction [6]. The mechanism of NB is as follows:

- Calculate the probability value for each attribute;
- Compute a joint conditional probability for all attributes using the product rule;
- Apply Bayes rule to extract conditional probabilities of class variables;
- Choose the highest probability value.

NB is used for music genre classification on a dataset which contains 417 of Malay music tracks in eight different genres [58]. The results of the paper depict some important factors which can improve the accuracy of classification including the size of the dataset, the length of music tracks, and cross-validation.

2.4.2 Hidden Naive Bayes

Hidden Naive Bayes (HNB) is another method which is based on the NB's logic. Each feature has a parent and also a hidden parent, which are created based on the effects of all other features. In addition, the hidden node is created based on "the average of weighted one-dependence estimators" [83].

Table 2.1: Classifiers′ name and corresponding abbreviations.

No.	Classifier Name	Abbreviation
1	Naive Bayes	NB
2	Hidden Naive Bayes	HNB
3	Naive Bayes Updatable	NBU
4	Bayes Network	BN
5	Decision Tree	DT
6	Decision Table	DTable
7	Classification Via Regression	CVR
8	Logistic Regression	Logistic
9	Multi-Leyer Perceptron	MLP
10	Sequential Minimal Optimization	SMO
11	Random Forest	RandF
12	Random Tree	RandT
13	Repeated Incremental Pruning	JRip
14	Multi-Class Classifier	MCC
15	Averaged One-Dependence Estimators	AODE
16	Weighted Averaged One-Dependence Estimators	WAODE
17	Random Sub-Space	RS
18	Ripple-Down Rule Learner	Ridor
19	LogitBoost	LogitBoost
20	Learns Alternating Decision Trees	LADTree
21	Decision Table Naive Bayes	DTNB
22	Bagging	Bagging
23	Boosting	Boosting
24	Stacking	Stacking
25	Dagging	Dagging
26	Rotation Forest	RF
27	Ensembles of Nested Dichotomies	END

2.4.3 Naive Bayes Updatable

Another type of Naive Bayes classifier is Naive Bayes Updatable (NBU), which is an incremental type. This classifier learns about each feature in every iteration using kernel estimators [83].

2.4.4 Bayes Network

Another machine learning method is Bayes Network (BN). BN has two main parts: a function to evaluate the quality of data and a search method to explore the data. BN contains random variables and shows the relationship between these variables via a probabilistic model [38]. For each instance of the data, a probability value is calculated. The overall result is accounted for multiplying the probabilities of individual instances. BN uses different learning methods, such as Adaptive Probabilistic Networks, at the training level [83]. Nasridinov et al. [56] used BN and some other algorithms to classify musical genre automatically. The output of the study shows BN could be a good option for genre classification of music data. The reason is that it integrates several acoustic features, such as key, chord, and bass, and creates a high-level model to increase the accuracy of music genre classification.

2.4.5 Decision Tree

The Decision Tree (DT) is a supervised predictive machine learning classifier, which obtains the results by considering the observation of items in a tree. Leaves are class labels and branches are the way to achieve to labels regarding passing features. One of the popular DT methods is J48 [6]. J48 creates a decision tree based on training data using information theory. Patra et al. [60] used several classifiers, such as J48 to classify music pieces in a dataset, including Hindi tracks with mood tags. The given results depict that J48 tagged instances correctly more often than other selected classifiers.

However, this learning method has some disadvantages, as the others do. For example, J48 usually needs too much time to obtain results [31, 58, 71]. In addition, various parameters

could change the performance of J48, which include the length of tracks and the given dataset size [58].

2.4.6 Decision Table

To choose appropriate attributes, Decision Table (DTable) finds all subsets of attributes, and then calculates the table's cross-validation performance for the subsets. Finally, it selects the subset that obtained the highest performance. DTable is also a classifier which is categorized as a DT method. In this classifier, the best-first search method is used to evaluate features [83]. In [12], the authors used a combination of some feature selection techniques and also some classifiers, such as DTable and J48, to classify the genre on music data. The study was focused on which combination of features reaches the best result in genre classification.

2.4.7 Classification via Regression

In Classification via Regression (CVR), each class of features converts to a group of binary numbers, the classifier creates discrete values via a regression model, and a regression model makes for each class label [83]. Barbosa et al. [5] used several classifiers, including CVR, SMO, and MLP, to classify music genre on 75 Brazilian music tracks. The results showed that it is important to know what types of acoustic features are used to classify music tracks, since different classifiers may work better with some specific audio features.

2.4.8 Logistic Regression

Logistic Regression (Logistic) is a type of linear classification which uses numeric attributes. It assigns some weights to input values and combines them linearly. Every contribution in this method has the chance to belong to any classes. The main difference between logistic and linear regression is that the output of Logistic is a binary value [83]. The proposed method in the study [11] used NB, Logistic and MLP to recognize genres and artists on the music data and detect key. By comparing the results, Logistic obtained the highest

accuracy for the key detection task. For two other recognition methods, MLP performs better than NB and Logistic. In another study [10], the proposed approach attempted to improve the performance of multi-label classification. Three different learning methods were used: Logistic, C4.5, and KNN. The results showed Logistic could obtain acceptable performance on estimating the optimal regression by using the combination of attributes, rather than using them separately.

2.4.9 Multi-Layer Perception

Multi-Layer Perception (MLP) maps a set of input values to a set of outputs. It includes several layers of nodes, and there is a connection between each layer and the next layer. Input values create the first layer, and each node at this level is called a neuron. MLP is a supervised learning technique and utilizes back-propagation neural networks [83]. Different parameters are set before applying this classifier. For example, NominalToBinaryFilter is a filter that is used to improve the performance of MLP [83]. The proposed method in study [11] focuses on creating a network to train several classifiers and classify the data based on three tasks: artist recognition, genre recognition and key defection. Key definition is used to show sharp and flat notes in a piece of music [11]. MLP without training had better results than other classifiers in genre recognition, and MLP with training reached the best result for artist recognition.

MLP is one of the most popular versions of neural networks, which is used in different areas of MIR, such as tag annotation, genre classification, musical instrument detection and missing rate of audio segmentation reduction [9, 15, 39, 69, 86, 87]. MLP is used in music genre classification in [93], where the combination of a neural network with a back-propagation algorithm applied on 353 music pieces. The results of the study show that MLP with back-propagation has achieved higher accuracy in mood classification.

2.4.10 Sequential Minimal Optimization

Support Vector Machines (SVM) is a learning classifier, which has used labeled data in training stage to produce a hyperplane, and that hyperplane to classify unlabeled new instances [83]. Sequential Minimal Optimization (SMO) is initially introduced to resolve the quadratic programming problems that exist during SVM training process.To this end, SMO provides the minimal optimization algorithm sequentially using different kernel functions, such as polynomial or Gaussian methods. It also transforms the nominal attributes to binary ones and normalizes the attributes using coefficients [83]. Both SMO and SVM classify the dataset base on some statistical learning theories. SMO divides the given data into different groups based on their features. It is used in many studies in MIR and improved the accuracy of music classification [32, 40, 69, 70, 85]. In music genre classification, it increased the accuracy [54]. Also, it is used in different areas of MIR, such as discovering the semantics of music by Lopes and his colleagues [42].

2.4.11 Random Tree

Random Tree (RandT) is a classifier that creates a tree of random features by a stochastic process. RandT does not use any pruning algorithm [83]. The proposed approach in the study [90] uses a RandT algorithm to handle some problems, such as low accuracies and efficiency. The results show that RandT could improve multi-label classification accuracy. Moreover, RandT achieves a lower running time when a dataset has many tracks and also increases the accuracy of classification in a multi-label tag annotation context.

2.4.12 JRip

JRip is a type of rule classifier, which is considerably fast. The core strength of this classifier is using “heuristic global optimization of the rule set” [83]. Authors in another study [6] used several classifiers, such as NB, Voting Feature Intervals (VFI), J48, NN, and JRip to evaluate the proposed approach. They trained based on 5 fold, 10 fold, and 20 fold cross validation on 90%, 75% and 60% of the dataset. The results of JRip for single

classification were better than multi-genre classification.

2.4.13 Multi-Class Classifier

One of the oldest problems in machine learning is multi-labelling, which means that some instances have more than one class. Multi-Class Classifier (MCC) assigns an individual classifier for each class; then the classifier sees only that class. The classifier has been trained based on the features of that class [38]. In other words, MCC uses two classifiers on a dataset, which contains instances of multi-classes. Moreover, it applies *error correction* techniques to improve the classification accuracy [83].

2.4.14 Averaged One-Dependence Estimators

The Averaged One-Dependence Estimators (AODE) classifier is a type of Tree-Augmented Naive Bayes (TAN) method. It obtains great classification results by computing the average result of Naive Bayes on different parts of the data [83]. Moreover, there is another type of AODE called Weighted Averaged One-Dependence Estimators (WAODE). The weight is calculated based on the mean of each ensemble for a super parent in each class.

2.4.15 Random Sub-Space

It is important for some classifiers to choose which parts of the data will be used as samples for performing calculations on them. To produce more reasonable results, selecting random subspaces of the data could be a good idea. Random Sub-Space (RS) is capable of choosing some random parts of the data. It also uses some ensemble techniques, such as bagging [83].

2.4.16 Ripple-Down Rule Learner

Ripple-Down Rule Learner (Ridor) is a kind of M5Rules classifier, which is a learner algorithm. The M5Rules method creates a tree model in each repetition. The model attempts to choose the appropriate leaf as a rule. The Ridor uses an error pruning mechanism

to decrease the amount of the error situation for regression problems [83].

2.4.17 Decision Table Naive Bayes

Decision Table Naive Bayes (DTNB) has been created by the combination of DT and NB. This classifier performs computations on the dataset by both methods, where DT applies to one-half of the features and NB applies to the other half [83].

2.5 Ensemble Techniques, Combination Methods, and Classifiers

Ensemble techniques are supervised learning algorithms. The primary purpose of these techniques is to apply different learning techniques to the same problem, which help improve the accuracy of prediction. In other words, they improve the classification results by combining the results of more than one classifier. There are some machine learning techniques, such as bagging, boosting and stacking, which have been categorized as ensemble techniques. They have better performance, compared to using a single machine learning method. In addition, one of the most important benefits of using several classifiers is that the combination of various models which have more similar behaviors increases the performance better than when some models of the same type are applied to a dataset [83]. However, similar to all other techniques, this machine learning technique has some disadvantages. For example, it is extremely complex to analyze the results when different models are applied. Furthermore, sometimes it is hard to understand why the combination model improves the performance [83].

2.5.1 Bagging

As an ensemble technique, Bagging combines the decisions of several classification algorithms. In this technique, all decisions have equal weight. Bagging produces new subsets of data using the original data to improve the training process and decrease the variance of predictions. The final accuracy is calculated employing the average or majority voting of all classifiers′ results [83].

2.5.2 Boosting

As with bagging, this method uses voting to calculate the final output of the results. Boosting uses confidence instead of equal weights of the models. In addition, the boosting technique learns from incorrect instances by giving higher weights to them in each new build [83].

Silla et al. [70] used the ensemble learner AdaBoost to classify music tracks based on the *weighted votes* of incorrect instances. The results indicated that the proposed supervised learning method obtained a higher accuracy when it applied to three different datasets.

Another study [48] stated that one of the advantages of using an ensemble learning method is using different classifiers of a category, such as Bayse and trees. Based on the outcomes of this study, boosting improved the performance of classification considerably compared to bagging, when the classifiers are trained on a dataset and applied on a test dataset which is different from the training one. In addition, the size of the training dataset has an effect on the boosting technique. In other words, more training can make boosting stronger. Furthermore, it can be seen that AdaBoost obtained higher classification accuracy on different datasets when compared to NB, NN and J48.

The proposed approach in the study [70] compared the performance of individual classifiers and applied ensemble techniques on three segments of music tracks. The results show that different classifiers obtained better accuracies on the middle segments of the selected tracks. The ensemble approach increased the accuracy slightly. The results of another study [75] indicate that the ensemble method achieves higher accuracy, compared to the results of each individual classifier.

2.5.3 LogitBoost

LogitBoost is a type of logistic tree model, which has the potential ability to describe linear-logistic regression models. It is a kind of Boosting method, but there are some differences among them. LogitBoost improves the likelihood of features, while Boosting im-

proves loss exponentially [83].

2.5.4 Learns Alternating Decision Trees

Learns Alternating Decision Trees (LADTree) is a modified version of Alternating Decision Trees (ADTree). Both focus on building "an alternating decision tree for two-class problems," while the first one uses LogitBoost and the second one adopts Boosting [83].

2.5.5 Stacking

Stacking (or stacked generalization) is another approach to combining different machine learning models. Boosting and bagging methods use the same types of machine learning algorithms, such as applying decision tree classifiers. However, a stacking technique applies various models, such as DT and NB [83]. In other words, we can set several classifiers from different categories as the base classifier in Stacking, while Bagging and Boosting accept just one classifier as the base classifier. By comparing the results of stacking versus individual classifiers, the results show that stacking obtained better accuracy [38].

2.5.6 Dagging

Dagging method makes different subsets of training data from a dataset and arranges data for each instance of the datset. This method is useful when the dataset contains an enormous amount of instances, which take too much time to generate results by applying base classifiers. In addition, the number of classifiers in the ensemble and the size of the training set for the base classifier are set based on the number of folds. Thus, the performance improves by choosing the proper folds [83]. Like Bagging and Boosting, Dagging uses just one type of classifier as the base classifier.

2.5.7 Rotation Forest

A Rotation Forest (RF) is another ensemble technique, which uses the combination of two classifiers, namely RF and Bagging. It also uses the principal components feature

selection technique to create some decision trees for facilitating the decision process and obtaining more precise performance [83].

2.5.8 Ensembles of Nested Dichotomies

In some cases, when the data include two classes, Ensembles of Nested Dichotomies (END) is a suitable classifier to create a sample tree. The final prediction will be calculated separately for each class by the average amount [83].

2.5.9 Random Forest

Random Forest (RandF) is one of the strongest meta classifiers, which is an ensemble learning method.It creates several decision trees from the input data at the training level and calculates the output using classification or regression of each tree. RandF is created based on various ensemble classifiers and trained by different methods, such as subspace [83]. The proposed approach in a study [1] is applied to the combination of low-level features, which are Spectral, temporal, energy, and pitch characters, to classify musical genre. The authors used RandF with different trees instead of using one tree to handle classification problems. The first order tree of RandF was created to train subsets of the dataset randomly. The results show that a combination of several low-level features increased the accuracy of RandF in genre classification on music data.

2.5.10 Some Existing Combination Methods

There are several combination methods, which are discussed in other studies. Minimum Probability, Maximum Probability, and Majority Voting are chosen to compare with our approaches in this study. These methods were introduced in the Sanden et al. study [63]. The authors introduced two N-dimensional vectors to calculate a score vector and a bipartition vector. The score vector is used to calculate the minimum score and maximum score. However, the majority voting is computed using bipartition vector. A combination method (CME) in an ensemble technique combines t classifiers H_1, H_2, ..., H_t. We define

H_{CME} as a classifier created after combining t classifiers. Suppose x_j is an unseen instance of the dataset and H_s is the classifier, so the score vector is P^j_{CME} = [$p^j_{1,CME}$, $p^j_{2,CME}$,..., $p^j_{N,CME}$] and the bipartition vector is B^j_{CME} = [$b^j_{1,CME}$, $b^j_{2,CME}$,..., $b^j_{N,CME}$]. In order to ease understanding formulas in the next three subsections, some indices will be provided in the following:

- q_i shows a mood label
- x_j indicates an instance of the dataset
- H_s represents a classifier

By considering these indices, we can define $P^j_{i,s}$ as the probability of assigning label q_i to unseen instance x_j by classifier H_s. We define $P^j_{i,k}$ as the probability of assigning label q_i to unseen instance x_j by classifier k.

Minimum Probability Method

Based on the done work by Sanden et al. [63], the Minimum Probability method chooses the tag which has the minimum score among all achieved scores. The logic behind this pessimistic method is to give a second chance to the weakest classifier to obtain a better performance. The formula to calculate the minimum score is:

$$P^j_{i,CME} = min_k(P^j_{i,k}), i = 1,2,...,N.$$

Maximum Probability Method

The Maximum Probability method is the opposite side of the minimum one. In other words, it is an optimistic approach, which finds the highest score and selects its tag as the final emotional tag for a piece of music. To achieve the maximum score, the formula is:

$$P^j_{i,CME} = max_k(P^j_{i,k}), i = 1,2,...,N.$$

Majority Voting Method

Majority Voting is another method proposed by Sanden and Zhang [63]. This method selects the tag where half of the classifiers or more vote for it. For example, if there are 5 classifiers and 3 of them selected sad as the final tag, Majority Voting selects sad as the final label too. However, if any tags could not achieve more than half of the votes for a music piece, the method does not select any label for that piece of music. The formula is:

$$B^j_{i,CME} = \begin{cases} 1 & \text{if } \sum_{k=1}^{s} B^j_{i,k}/s \geq 0.5 \\ 0 & \text{otherwise.} \end{cases}$$

where i = 1,2,...,N, and s is the number of classifiers.

2.6 Previous Works on Classification in MIR

Different factors affect the performance of music classification in MIR. For example, providing the properly balanced dataset is one of the necessary steps in music classification. Furthermore, the size of the dataset can affect the accuracy of music classification. Moreover, choosing the appropriate acoustic features is another important subject in MIR.

2.6.1 Music Classification based on Features

In most previous studies, selecting suitable features has a considerable impact on the accuracy of classification. Moreover, the combination of different features has been used to improve music classification accuracy. The comparison of low-level and high-level features was revealed in many studies. Extraction of low-level features is easier than to achieve in high-level in MIR [80]. The reason is that high-level features contain more perceptual features and a wide range of factors may affect them, such as musical knowledge, psychological elements, cultural and empirical phenomena. Among all defined acoustic features, MFCCs are the most popular ones and are used in many proposed approaches. For example, MFCC is employed in the proposed method by Mandel et al. to classify music tracks

[44].

Li et al. [37] extracted low-level and high-level content-based features by MARSYAS [8] from the corresponding audio signals. The approach works based on the detection of the best tags in a multi-label dataset.

McKay and Fujinaga [50] believe that the combination of low-level and high-level features, *audio data* and *symbolic data*, can efficiently improve the automatic music classification compared to using the individual features.

In addition to song-level classification, acoustic features are used to predict appropriate tags [12, 69]. However, there is a famous conjecture; the combination of several audio features increases the accuracy of classification [91]. Zhang et al. [91] found that a combination of four types of features can achieve better outcomes. In their proposed method, four short-time features are extracted to classify music based on content. These features include short-time energy, short-time zero-crossing rate, short-time fundamental frequency, and harmonious degree.

Pohle et al. [61] selected six machine learning algorithms to classify a music dataset based on the feature sets, including timbral texture, beat histogram, and pitch. They show that musical content and audio description techniques are among suitable sources to gather information.

Yang et al. [86] choose MFCCs and Spectral features, such as temporal description, to predict annotations for unlabeled music tracks. The idea of another investigation is to derive Spectral, rhythmic (tempo) and MFCCs features to classify genres [49]. Besides, MFCCs and Spectral features are extracted from audio signals to reduce the missing rate when audio tracks have been segmented [92].

To tackle music tagging problems, Orio and Piva [59] presented a new approach for semantic music tagging based on the combination of timbric and rhythmic features. The proposed method shows that the combination of features could increase the accuracy of music clas-

[8] http://MARSYAS.info/about/projects.html.

sification considerably.

2.6.2 Music Genre Classification

There are many studies which are focused on finding the proper approach for classifying music pieces based on genre [84, 58, 79, 45, 43]. Xu et al. [84] used MFCCs, Spectral, and Cepstral features to classify music tracks via genre. In another study, timbre and rhythm features were an acceptable combination for automatic genre classification in [58].

In 2002, George Tzanetakis and Perry Cook [79] performed a study on automatic music classification based on hierarchical genres. One of the most significant prerequisites in their approach was feature extraction. They chose timbral texture, rhythmic content, and pitch content. Those timbral textures, which were used to classify speech and general sounds, contain Spectral Centroid, Spectral Rolloff, Spectral Flux, MFCCs, Analysis and Texture Window, and Low-Energy Features in their study. Moreover, rhythm and pitch represented the musical content, such as harmony. Finally, they stated that for each genre of music, there is a specific group of audio features that shows the specifications of the genre better than the others [79].

In addition to acoustic features, the classifiers play important roles in improving of classification performance. Mandel et al. [45] used Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM) [9] to classify music tracks based on content in a dataset with around 1,000 music tracks. Further, those classifiers achieved acceptable accuracy, around 70%, when classifying music genre via explicit semantic analysis in a dataset with around 1,800 pieces [43]. However, another study shows that multi-layer SVM obtained higher accuracies, around 97% within optimal class boundaries, compared to the traditional Euclidean distance between different kinds of genres [57].

Scaringella et al. [64] used music genres to characterize music collections. These contents

[9] Among all implemented classifiers in WEKA, we chose some of the most popular ones, which used more by some other researchers in MIR, in order to use them in different parts of our experiments in this thesis. For easy reference, the complete names and the corresponding abbreviations of those classifiers can be found in Table2.1.

include three distinct categories of the complex interplay of cultures, artists, and market forces. K-Nearest Neighbor (KNN) and SVM are used to find appropriate annotations for music tracks in two datasets, which are made by the authors. The first dataset contained 1,515 songs over ten genres, of which 1,005 tracks were used for training and 510 tracks were used for testing. The second dataset consisted of 1,414 songs over six genres, of which 940 songs were used for training and 474 songs were used for testing. The results of experiments depicted that classifying multi-genre songs are still a difficult problem, since there is no available proposed approach to label a song with more than one genre [64].

In another study, Silla et. al. [71] improved an automatic music genre classification using a machine learning approach. To attain this goal, two frameworks, MARSYAS and WEKA, are employed to extract and classify features, respectively. The MARSYAS extracted features such as Beat-related, Timbral Texture, and Pitch-Related acoustic features from audio signals. Moreover, some WEKA classifiers are chosen to classify music pieces of a dataset with 300 tracks of the Latin Music Dataset. These classifiers include Decision Tree (J48), K-NN, Naive Bayes (NB), MLP, and SVM. Consequently, they stated that reducing the feature sets and using selected features decreased the actual running-time complexity. In addition, using feature selection techniques improved the accuracy of some classifiers, such as J48, k-NN, NB, while it had no extreme effect on the accuracy of SVM and MLP.

2.6.3 Music Mood Classification (MMC)

There are few studies in MMC. In this section, several studies on MMC are considered in order to find more details regarding this area of MIR [38, 24, 23, 61, 77, 29]. There are some previous works along this direction, in which acoustic audio features, feature-selection techniques as well as computational classifiers play an important role to improve the classification accuracy. Li et al. [38] used content-based acoustic features to investigate computational approaches in MMC.

Hu and Downie [24] investigated different factors, which include lyrical tags, audio fea-

tures, and their combination to improve mood classification accuracy in music. They used a dataset of 18 groups of tags, where each group contains one to 25 social tags. Also, in order to model moods in music in the context of psychology, they used the two-dimensional Russell's model. These dimensions are the psychological valence to show moods in a negative and positive Spectrum. Also, they indicate the degree of activeness or inactiveness of different moods [38]. Hu et al. [23] and Pohle et al [61] studied in this area. From the outcomes of their experiments, we see that a combination of audio features could increase the accuracies of classification.

As mood is a perceptual phenomenon, after hearing a musical piece, it is possible to assign more than one tag to it. Some works have attempted to solve this problem and focused on multi-label classification. To deal with this issue, four multi-label classification algorithms are examined by Trohidis et al. [77], namely BR, LP, RAKEL, and MLKNN, of which RAKEL achieved the highest accuracy, with 87%, by SVM and a 10-fold cross-validation evaluation scenario.

In genre classification, there are many studies which focused on how different classifiers and acoustic features may increase the accuracy of music genre classificaion. Their are fewer studies in MMC than in music genre classification. Jamdar et al. [29] used a weighting method for acoustic features, and then classified them by a k-Nearest Neighbor classifier (k-NN). From the results of this study, k-NN can provide Music Recommendation Systems and Automated Playlist Generation Systems using the weighting method for acoustic features.

K-NN, NB, as well as SVM are popular classifiers, and are widely employed in different studies [23, 55, 61, 74] to handle the classification issues in music data mining. Also, a back-propagation neural network classifier is employed to make better decisions by a layered structure in MMC [16, 21].

It should be noted that while there are questions raised by Fiebrink and Fujinaga [17] regarding overstating the effectiveness of using feature-selection or dimensionality-reduction

techniques for classification tasks in music, their work represents their initial efforts toward mood classification. Laurier et al. [33] investigated MMC via SMO, Logistic, and RandF based on the audio and lyrics. The SMO achieves the best accuracy, with 80.7%, compared to others in both audio and lyrics classification. The second best performance is obtained by applying RandF for audio features. Song et al. [74] created a dataset containing 2904 songs of the Last.fm, which are tagged by one of these: *happy*, *sad*, *angry* and *relaxed*. During the preprocessing phase, they removed some noisy data, such as confusing tags and repeated songs, from the dataset manually. Then, some modified versions of SVM were applied to classify the dataset based on a 10-fold cross-validation method. In that study, two popular approaches -categorical and dimensional- were used to classify music based on emotion. In the first one, emotion was categorized based on universal terms, such as happiness, sadness, anger, and fear. However, in the second approach, they used terms in neurophysiological systems, including valence (negative to positive) and arousal (calm to exciting). Finally, they show that the Spectral class with 32 features reached good accuracy around 51.9%, while a combination of Spectral, rhythm, and harmony were the most accurate, with 53.6% [74].

2.7 Waikato Environment for Knowledge Analysis

Data Mining is a complex subject which contains a broad range of algorithms and preprocessing steps, as well as different tools, to process an enormous amount of data and derive the appropriate subsets of the data.

Waikato Environment for Knowledge Analysis (WEKA) [10] is a powerful open-source data mining tool that is implemented in JAVA. It was developed by a group of researchers at the University of Waikato in New Zealand [83]. WEKA provides a group of Graphical User Interfaces (GUI) which helps users work with it. One of the most important parts of this application is called Explorer, which involves different supervised and unsupervised filters,

[10] http://www.cs.waikato.ac.nz/ml/weka/.

classifiers, clustering techniques, and association rule mining algorithms. These parts implement many machine learning algorithms that are used in the real world [28]. WEKA uses a specific data format called Attribute-Relation File Format, which consists of two main sections: header and data [67]. The former contains the name of the attributes, while the latter involves the real attributes which are id, the value of attributes, and the tag.

All used datasets in this study are in $.arff$ format. In our study, the two main parts of WEKA used are filters and classifiers. Among many filters implemented in WEKA, feature selection filters employed in the preprocessing steps trim data before classification. There are references to some previous studies that illustrate the popularity of using filters. For example, the focus of the study [73] is to find a group of optimal features for each classifier in an ensemble method. In another study, to reduce the complexity of the calculations, principal component analysis (PCA) is applied [83]. Silla et al. [71] used five classifiers to create an automatic music genre classification. All of these examples used WEKA to implement and evaluate their approaches.

It can be clearly seen that WEKA is a popular software among MIR researchers. On the one hand, some researchers make their own modified classifiers. To illustrate, Norowi et al. used the J48 classifier to classify non-western music tracks based on genre tags [58]. Fiebrink et al. [18] applied the K-Nearest Neighbor classifier to weight different acoustic features to facilitate the classification process. On the other hand, many researchers work with WEKA, since it contains some of the best feature selection techniques. For example, Wand et al. [81] used the CfsSubsetEval to reduce the dimensionality of features. CfsSubsetEval is one of the feature selection techniques that was implemented by WEKA and discussed earlier in this chapter.

WEKA is used as the main core of many new frameworks. For instance, McKay et al. [48] created a system named ACE, which related to WEKA and provides more classification facilities.

Chapter 3

An Empirical Study on Music Mood Classification through Computational Approaches

In this chapter, we empirically explore different classification schemes on the mood classification problems in music data. Using mood to classify music data is subjective and ambiguous. Through comprehensive empirical experiments, we demonstrate that the current classification schemes are not sufficient to conduct music classification through mood. Various issues such as feature selection and feature discretization are analyzed and discussed.

The primary purposes of this chapter are to find, through empirical experiments, what combinations of classifiers and feature selection techniques work better in order to classify moods in music data and to analyze and discuss various issues related to the music mood classification problem [11].

3.1 Introduction

Music classification, i.e., categorizing music pieces into classes according to some criteria, such that subsequent operations (mainly querying) can be efficiently conducted, is usually conducted as an initial step toward high-level MIR tasks, including automatic tag annotation, recommendation, and playlist generation. Some favorite criteria employed by

[11] The chapter has been accepted by the 3rd International Conference on Systems and Informatics conference for publication.

practitioners include genres, moods, and instruments. While some of these criteria are still considered ambiguous and subjective, musicians and listeners alike use them to categorize music.

Music moods can help listeners choose the proper songs based on their interests [65] and thus narrow down the music list that they are looking through. Compared to musical genres, mood expressions in music are even more ambiguous and subjective. For instance, a musical piece might be labeled with two entirely different moods, depending on the judges involved. Even for the same music piece judged by the same person, it could be assigned a different mood tag at various times. However, due to the growing music datasets, it will be more error-prone and labor-intensive if the practitioners are still engaged in manual mood classification in a music dataset. Therefore, computational approaches are actively sought to automate the mood classification process.

To this end, we will conduct a series of experiments in the following sections to determine empirically the appropriate sets of acoustic features, which present enough information regarding the moods of music tracks, and the proper classifiers that achieve higher accuracies in MMC. There are some previous works that show computational mood classification is still largely a virgin land in the MMC [16, 23, 35]. We believe the MMC needs more cultivation. Our experiments will involve several music datasets that are publicly available for benchmark purposes. There are four types of mood tags involved in our study, including *happy*, *sad*, *angry*, and *relax* in these datasets, which were assigned by music some experts.

This chapter also shows whether or not the current classification schemes can improve the performance of mood classification in music by various sets of current audio features that represent acoustic characteristics of music.

3.2 Experiment Preparation

As mentioned before, mood classification is an ambiguous and subjective process. In order to automate it by devising more intelligent computational schemes, we have conducted a series of empirical experiments to explore different computational classifiers and acoustic features.

During those experiments, a 10-fold cross validation is used to obtain impartial estimations of the error rates of a given dataset. It is hoped that after these experiments, we could have a good understanding of the current practice and therefore will be engaged in more focused efforts in our works on mood classification in music in the next chapter. Before considering the experiment results and analysis, we will describe the datasets, acoustic features, classifiers, and some other related issues.

3.2.1 Experiment Environment

In this study, we employ a set of classification algorithms to obtain the classification results on moods in music data. The programming environment for our experiments is WEKA [12], which is introduced in Chapter 2.

3.2.2 Music Datasets

Three popular music datasets are used for our experiments in this chapter; namely, the MSD[13] and the Last.fm dataset [14], and Cal500 [82]. Marsyas is also used to extract acoustic features, such as low-level spectral features, timber, as well as MFCCs from the MSD dataset in many previous works [7]. Our study concentrates on the four popular mood tags, including *happy*, *sad*, *angry*, and *relax*.

The music pieces in the MSD and Last.fm are adapted to create balanced datasets. We obtained two balanced datasets from these two datasets. The first balanced dataset, denoted as DS_1, has 1280 music pieces per mood class, while the second one, denoted as DS_2, has

[12] http://www.cs.waikato.ac.nz/ml/WEKA/.
[13] http://labrosa.ee.columbia.edu/millionsong/.
[14] http://labrosa.ee.columbia.edu/millionsong/lastfm.

500 musical pieces of each mood. In addition, a small balanced dataset, called DS_3 created from Cal500, which has 30 songs per mood tag is used to get more classification accuracy. However, since there is no relax label in the DS_3, we use $Calming/Soothing$ labels instead of $Relax$ tag.

3.2.3 Feature Sets

In Chapter 2, we found that a combination of two or more feature sets improves the accuracy of music genre classification [50, 91, 61]. In this chapter, our experiments use multiple feature sets, which contain a total of 102 features and grouped into three feature sets, including low-level features, rhythm histogram features, and MFCC features [38].

3.2.4 Feature Selection and Dimension Reduction Techniques

The primary goal of feature selection techniques is to select the best features [38, 83]. These methods grouped into two main categories: filter and wrapper, as discussed in chapter 2. Among a large number of filter techniques, four of them are used in our experiments, namely Information Gain (IG), Gain Ratio (GR), Symmetrical Uncertainty (SYM), and Chi-Squared (Chi). For the latter method, we are used the Correlation Feature Selection (CFS). Moreover, dimensionality reduction techniques are used to eliminate the redundant and irrelevant data. In our experiments, we used the Principal Component Analysis (PCA) to transform the data into a new space with less irrelevant data. We chose these techniques, because they were popular among many researchers in MIR.

3.2.5 Various Classifiers

Finding a predictor model is a significant task that classifiers can accomplish. The predictive model is used to generate predictions for new unlabeled input data [21]. A set of classifiers has been chosen to compare the MMC performance and investigate the impacts of different feature-selection techniques. All of these classifiers are introduced in chapter 2. Here, the 23 classifiers are used in our experiments, which categorized into five main

groups, as follows.

- Bayes: Naive Bayes (NB), Bayes Net (BN), Averaged One- Dependence Estimators (AODE), Weightily Averaged One-Dependence Estimators (WAODE), Hidden Naive Bayes (HNB), and Naive Bayes Updatable (NBU);
- Functions: Sequential Minimal Optimization (SMO), and Multilayer Perceptron (MLP);
- Meta: Bagging, Dagging,Classification Via Regression (CVR), LogitBoost, Multi-Class Classier (MCC), Random Sub-space (RS), Rotation Forest (RF), and Ensembles of Nested Dichotomies (END);
- Rules: Repeated Incremental Pruning (JRip), Decision Table (DTable), Decision Table Naive Bayes (DTNB), and Ridor;
- Trees: Decision Tree (J48), multi-class Alternating Decision Tree (LADTree), and Random Forest (RandF).

3.2.6 Data Cleaning

Some classifiers are not enabled in WEKA unless the data cover a classifier′s capability. One of these capabilities is the compatibility of the classifier to work with nominal or numeric data. Many classifiers in WEKA only work with discrete values, so the data should be transformed into a discretized format. Discretization is a supervised filter that transfers a spectrum of continuous attributes into some distinct ranges [13, 83]. Partial Least Squares (PLS), as a dimensionality reduction technique, are also applied to the dataset to remove the redundant instances. Both filters are applied in the pre-processing step in our experiments to clean and prepare the data [23].

3.2.7 Different Experiments

Three categories of experiments were conducted on the created datasets. In the first category, the datasets were classified by all 23 classifiers without any filter. In the second

and third categories, all six feature selection and dimensionality reduction techniques were used to choose more *appropriate* features. In general, the proposed experiments are:

- **Experiment 1** , denoted EXP_1: classification without using any filter.
- **Experiment 2** , denoted EXP_2: classification via discretization and the selected feature selection and dimensionality reduction techniques.
- **Experiment 3** , denoted EXP_3: classification with performing PLS as a supervised filter and the selected feature selection and dimensionality reduction techniques.

All experiments are conducted in WEKA without manipulating the default value of parameters of any techniques and classifiers, because these experiments are performed to compare the accuracies of classification by applying basic techniques in WEKA. Changing the default value of parameters and comparing the results of classification algorithms are among another type of experiments, which are done by other researchers and we do not focus on these parts in this study. The results are summarized and analyzed below.

3.3 Results and Discussions

In the following discussions, we only present the performance of some of the top classifiers, due to space limitations.

3.3.1 Classification Without Pre-processing

In EXP_1, all 23 classifiers are used to classify the three datasets, DS_1, DS_2, and DS_3, without performing any pre-processing and data cleaning techniques, such as feature selection and discretization techniques. The best accuracies of this experiment are 49.5% for DS_1, 46.1% for DS_2, and 84.85% for DS_3, respectively. Figure 3.1 shows the best five classifiers in EXP_1 on DS_1, while Figure 3.2 shows the best five classifiers in DS_2. We attribute the higher accuracy in DS_3 to the smaller number of musical pieces and high data quality from Cal500.

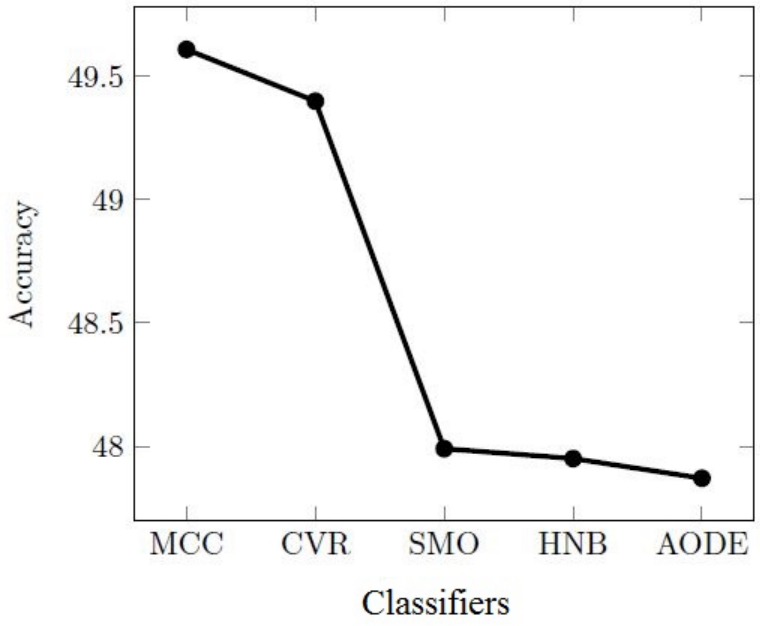

Figure 3.1: The performance of the best five classifiers on DS_1 without pre-processing.

Three experts considered the Cal500 dataset, so the derived acoustic features of it, i.e. emotional tags are more precise than those in the Last.fm datasets. Therefore, we can safely say that the data quality of Cal500 is higher compared to other employed datasets in this chapter. It also appears that the best classifiers on DS_1 are different from those on DS_2. However, for these more practical datasets, the highest accuracy is still below 50%, when the chosen classifiers classify them.

The top five classifiers on EXP_1 on DS_3 showed in Figure 3.3. As we can see, BN obtained the highest accuracy around 81%. The accuracy of RandF is around 80%, which is very close to BN. LogitBoost, RS, and MLP with the accuracy over 75% are among the top five classifiers in EXP_1 on DS_3. As discussed before, the quality of data in DS_3 is better than other datasets in this chapter so that classifiers could achieve higher accuracy on it.

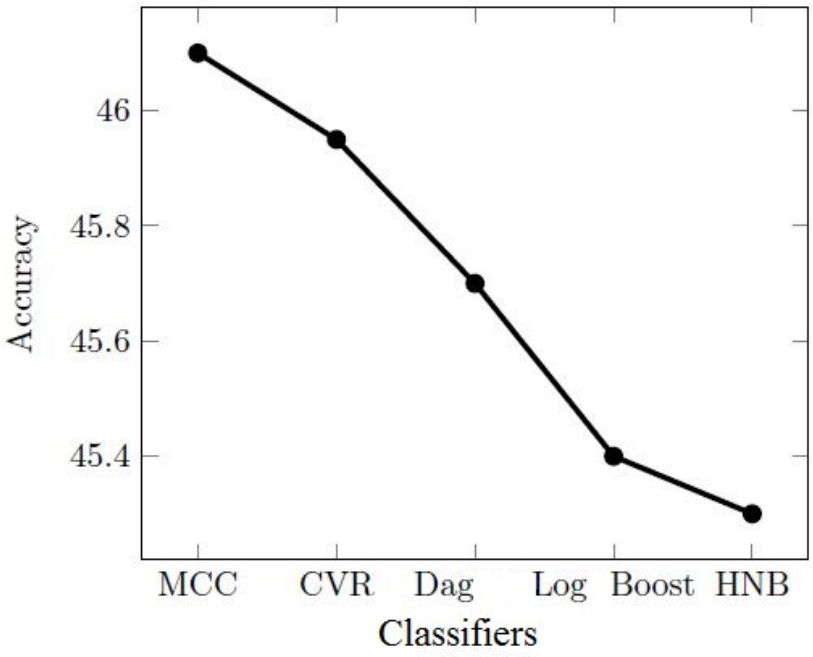

Figure 3.2: The performance of the best five classifiers on DS_2 without pre-processing.

3.3.2 Classification With Pre-processing

For EXP_2, discretization and the selected feature selection techniques are used as pre-processing steps. It means that the datasets discretized at first, and then more appropriate features are extracted.

By comparing the results of EXP_1 and EXP_2, there is a slight increase of accuracies on both DS_1 and DS_2. In general, we have found there are almost no changes in the classification accuracy on DS_3 in both EXP_1 and EXP_2 since the musical pieces in the DS_3 are already pre-processed and of high quality.

In detail, the performance of top six classifiers on the DS_1 is depicted in Figure 3.4. We observe that the general performance of classifier MCC (around 50.41%) is higher than the others, especially when it combined with the dimensionality reduction method PCA. One of the most important observations from this figure is that using PCA improves the performance. Whether or not this is some general trend needs further investigation, since as pointed in [17], involving PCA achieves only comparable results when it comes down using features to conduct classification in music.

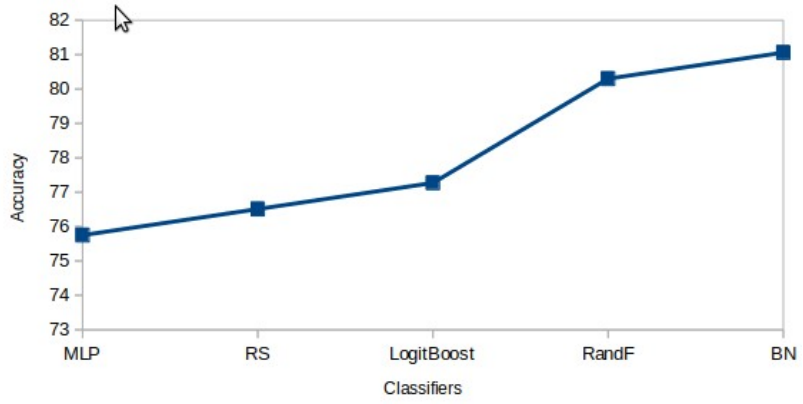

Figure 3.3: The performance of the best five classifiers on DS_3 without pre-processing.

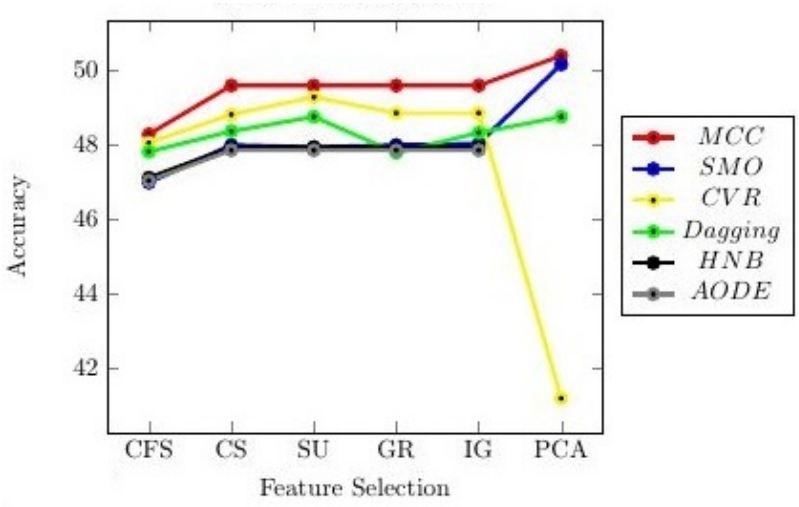

Figure 3.4: Top six classifiers on DS_1 after discretization.

There is an interesting observation for EXP_2. Figure 3.5 indicates the effects of applying feature-selection methods in EXP_2. By comparison, we observe that the classifier NBU is one of the best classifiers, which obtained the highest accuracy in this study. We also find that the dimensionality reduction method PCA does not result in better accuracies, as compared to Figure 3.4. The reason is that PCA could overlap some features in the DS_3, which is already high quality.

In EXP_3, the PLS technique is chosen as a filter instead of using the discretization technique. Though classifier NB achieves the highest accuracy on DS_3 in comparison to the other two datasets, the overall accuracy is $5-10\%$ lower than those with the discretization

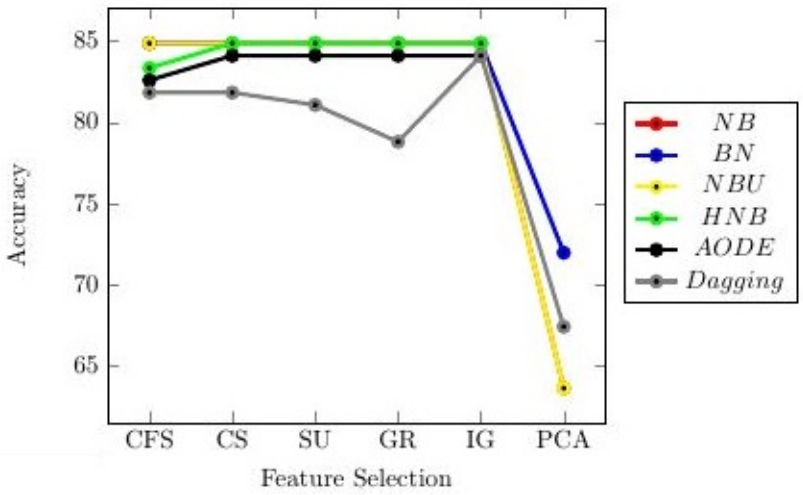

Figure 3.5: Top six classifiers on DS_3 after discretization.

technique. Again, for the relatively higher classification accuracy in the DS_3, we conjecture that it is due to the higher data quality from Cal500. Based on our observation, PLS is inferior to the discretization technique in mood classification. One reason could refer to the amount of irrelevant and redundant data in the datasets since PLS tries to transform data into a new subspace, but it may fail in creating the right correlation of the datasets.

3.3.3 Feature Selections and Dimension Reductions

Table 3.1 shows the five best feature-selection and dimensionality-reduction techniques based on the mood classification accuracy for each dataset in the EXP_2 and EXP_3. We observe that IG and GR are among the best feature selection techniques in those experiments, except when they are used to classify the DS_2 in the EXP_2. Considering the DS_1 and DS_2 are two more realistic datasets in applications, we maintain that the PCA is a good candidate for data preparation when conducting classification in music.

As we see in Table 3.1, IG, GR, SYM, Chi, and CFS could not change the accuracy most of the time. It is because they could not differentiate acoustic features of mood tags for each song. The table just shows the top five feature selection techniques in the second and third experiments on all three datasets. For example, we cannot see the CFS in the EXP_2 on DS_1 because its accuracy is less than 49.61%.

Table 3.1: Order of feature selection techniques and highest accuracies in EXP_2 and EXP_3.

	EXP_2	Highest Accuracy	EXP_3	Highest Accuracy
DS_1	PCA	50.41	IG	45.84
	IG	49.61	GR	45.84
	GR	49.61	SYM	45.84
	SYM	49.61	Chi	45.84
	Chi	49.61	CFS	43.14
DS_2	PCA	47.15	IG	43.3
	CFS	46.45	GR	43.3
	Chi	46.25	SYM	43.3
	IG	46.15	Chi	43.3
	GR	46.1	Cfs	41.35
DS_3	CFS	84.85	CFS	67.42
	IG	84.85	IG	67.42
	GR	84.85	GR	67.42
	SYM	84.85	SYM	67.42
	Chi	84.85	Chi	67.42

3.3.4 Classifiers

Table 3.2 shows those classifiers that obtained the highest accuracy in our experiments. After considering the accuracy of all classifiers in all experiments, the top five classifiers of each experiment are selected and placed in the table based on different datasets. We can see that classifiers MCC, CVR, SMO, and HNB are among the best in the most experiments for the three datasets. Also, MCC achieves the highest accuracy on DS_1 and DS_2 in the first two experiments. However, on DS_3, classifiers NB, HNB, BN, and NBU reach the similar accuracies in the first two experiments.

Table 3.3 shows the combined results on the best classifiers when they work with the best feature-selection or dimensionality-reduction techniques. For example, the highest accuracy of EXP_3 on DS_1 is 45.84%, which is obtained by HNB after applying one of IG, GR, SYM, and Chi. It means that the next accuracy of EXP_3 on DS_1 is less than 45.84%, and the accuracy obtained by either another classifier(s) or feature selection techniques. Based on our results, it is clear that the accuracy is increased slightly when a combination of classifiers and feature selection techniques are used to classify the datasets. Depending

Table 3.2: Best classifiers in all experiments in descending order.

	EXP_1	EXP_2	EXP_3
DS_1	MCC	MCC	HNB
	CVR	SMO	CVR
	SMO	CVR	MCC
	HNB	Dagging	Dagging
	AODE	HNB	SMO
DS_2	MCC	MCC	WAODE
	CVR	SMO	MCC
	Dagging	CVR	NB
	LogitBoot	Dagging	HNB
	HNB	DT	NBU
DS_3	NB	NB	NB
	BN	NBU	NBU
	HNB	HNB	AODE
	NBU	AODE	SMO
	Dagging	Dagging	BN

Table 3.3: General results of classification on all three datasets.

	EXP_1		EXP_2			EXP_3		
	Classifier	Accuracy	Classifier	Feature Selection	Accuracy	Classifier	Feature Selection	Accuracy
DS_1	MCC	49.61	MCC	PCA	50.41	HNB	IG	45.84
							GR	45.84
							SYM	45.84
							Chi	45.84
DS_2	MCC	46.1	MCC	PCA	47.15	WAODE	IG	43.3
							GR	43.3
							SYM	43.3
							Chi	43.3
DS_3	NB	67.42	NB	CFS	84.84	NB	CFS	67.42
	BN	67.42	BN	IG	84.84	NBU	IG	67.42
	HNB	67.42	HNB	GR	84.84	AODE	GR	67.42
	NBU	67.42	NBU	SYM	84.84	SMO	SYM	67.42
	Dagging	67.42		Chi	84.84		Chi	67.42

on the results of EXP_1 and EXP_2, MCC is the best classifier, and PCA is the best data preparation technique on DS_1 and DS_2.

Table 3.4: Highest accuracy of EXP_2 for all 23 classifiers on dataset DS_1.

No.	Classification	Feature selection	Acuracy
1	MCC	PCA	50.41
2	SMO	PCA	49.61
3	CVR	SYM	49.61
4	Dagging	PCA	49.61
5	HNB	IG	49.61
6	AODE	IG	47.87
7	WAODE	IG	47.8
8	LogitBoost	GR	47.55
9	RandF	IG	47.35
10	RS	IG	47.25
11	RF	SYM	46.75
12	LADTree	IG	46.7
13	MLP	SYM	46.6
14	NB	CFS	46.3
15	BN	CFS	46.3
16	NBU	CFS	46.3
17	Bagging	GR	46.25
18	DTNB	Chi	46.2
19	END	GR	46.1
20	Ridor	SYM	43.65
21	DT	CFS	43.15
22	J.48	CFS	43.05
23	JRip	Chi	42.9

Table 3.4 depicts the overall performance of all 23 classifiers on DS_1 after performing the discretization technique. We observe that the combination of MCC and PCA reaches the highest accuracy around 50.41%. SMO, CVR, Dagging, and HNB obtained the same level of accuracy, with 49.61%. This table illustrates the importance of how choosing a classifier, which works better with a specific feature-selection or dimensionality-reduction technique, can improve the performance of MMC. As we see, using JRip and Chi for classifying songs

on DS_1 obtained an accuracy around 8% less than using MCC and PCA on the same dataset.

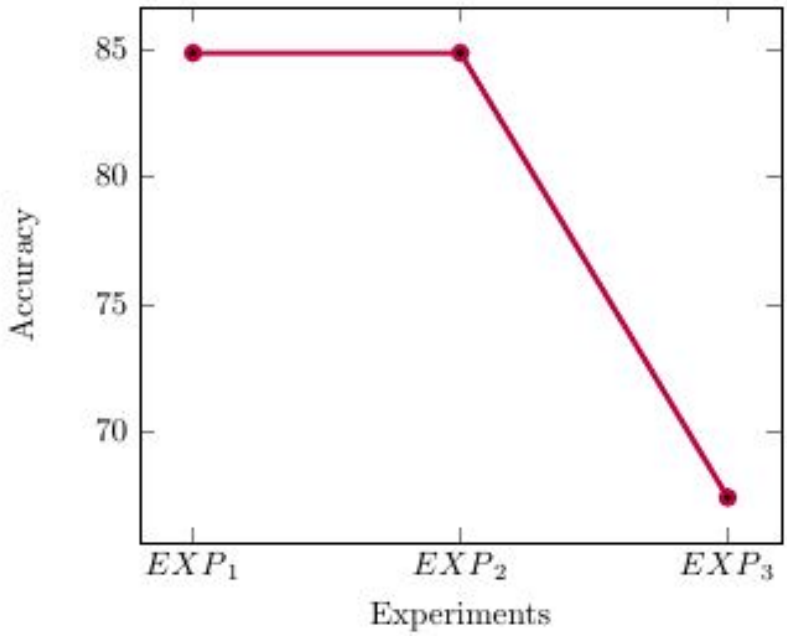

Figure 3.6: General results of classificion on dataset DS_3.

3.3.5 Dataset DS_3

Dataset DS_3 has a better quality than the other two datasets in our experiments, since its originator dataset, Cal500, is generated with a considerable care. Figure 3.6 illustrates the best results of the experiments on DS_3. We observe that the combination of different classifiers and feature-selection or dimensionality-reduction techniques can achieve different accuracies. As an example, NBU with CFS, Chi, SYM, GR, and IG reached the highest accuracy. However, this classifier with PCA results in the lowest accuracy in comparison to other feature selection techniques. Based on our results, it is clear that some feature selection techniques remarkably improve certain classifiers' performance, while others decrease the classification accuracy. Therefore, there is no guarantee that, while one feature-selection or dimensionality-reduction technique enhances the performance of a classifier, it will have the same effects on the others.

3.3.6 Discussions

Through our experiments and results, we have found some interesting observations (though some of them have been reported before in the literature), which is summarized in the following.

We want to emphasize the importance of involving data preparation or data pre-processing before the classification tasks in music, as shown by the improvements in Figures 3.5 and 3.6. Moreover, feature-selection or dimensionality-reduction techniques are crucial to the effectiveness of mood classification in music. In other words, employing all features or a small number of them may result in unacceptable performance since the former may confuse the classifiers, while the latter cannot evaluate the data in a truthful way. In our experiments, as shown in Table 3.1, IG and GR are among the best techniques in our datasets. However, in some situations, other techniques, including PLS will produce better results. In addition, several classifiers are used to classify music data using the mood labels and find the strongest classifiers among them for the next experiments. Table 3.2 depicts that, it is not wrong if we say MCC works well in classifying music data based on mood tags, while other types of classifiers (i.e. CVR, SMO, WAODE, NB, NBU) reached remarkable certainty. Thus, whether a particular classifier performs well is highly dataset-dependent and needs to be determined through experiments.

3.4 Summary

The extensive experiments in our study show that MMC is a hard task in practice. In our study, some current classification techniques are compared to understand which combinations of feature-selection or dimensionality-reduction techniques and classifiers work better and have the strength to perform well in mood classification. While we see that there could make some improvements through a particular combination of classifier and data preparation technique, the improvements are quite marginal, calling for more advanced techniques and methodologies, if any, to tackle the mood classification problem. As detecting mu-

sic piece′s mood is a subjective issue, one needs to contemplate different phenomena all together to achieve better classification accuracy. In the next chapter, we plan to use advanced techniques, such as ensemble techniques, to find a better classification accuracy on the mood of music data.

Chapter 4

Improving Mood Classification in Music

In this chapter, two empirical methods are proposed and implemented to improve the accuracy of MMC. These two ensemble approaches are focused on how several classifiers are combined to classify some unseen music tracks into existing categories. Moreover, several existing methods in this area are introduced to compare with our approaches. All implementations were done in JAVA using WEKA library Version 3.8.0.

4.1 Data

Using the introduced music benchmarks in Chapter 2, MSD and Last.fm, three subsets were created in this chapter: 5120M, 2000Head, and 2000Tail. All these subsets in this study are divided into two parts, training and testing, which are used in our experiments.

4.1.1 5120M Dataset

As previously discussed, MSD is one of the most popular benchmarks in music data mining. However, the lack of music tags in this benchmark is a significant issue. Moreover, Last.fm has almost the same music tracks, including some music tags for each piece of music. After considering the similar music tracks in MSD and Last.fm, all music tracks which contain those four mood tags were selected. There were two major problems for this dataset:

- Some music tracks have more than one emotional tag.

- Several tags have the same concept, but adopt different words. For example, for a piece of music, users may assign these tags: happy, happiness, cheerful, delighted, elated, glad, joyful, pleasant, or upbeat.

To tackle the first problem, redundant and repeated tags are removed, so the remaining tags have the same concept. For the second problem, all synonym tags are removed, leaving only the main tag. To illustrate, in the last example, *Happy* is preferred as the most appropriate tag for that music track. Another solution for the second problem is to define a new taxonomy for labels and ask users to assign tags based on it. This solution is related to a different area of MIR, and we do not focus on it in this study. Afterwards, an equal number of music tracks are picked up for each tag to create a balanced dataset. Altogether, there are 1280 music tracks for *Happy*, and the same numbers for each of the other three mood tags. During the experiments, the dataset needs to be divided into two parts for training and testing. Accordingly, we use a training subset with 4,000 music tracks and also a testing subset with 1,120 music tracks.

4.1.2 Subsets of 5120M Dataset

As mentioned before, there are not many datasets that can be used to evaluate mood classification in music. Hence, the number of music pieces is not something to be compared with the real-world data. Thus, two smaller datasets were created based on the 5120M dataset to get more results using them in different experiments. The first subset is called 2000Head, which contains the first 500 music tracks of each mood tag. Also, the second subset is called 2000Tail, which comprises the last 500 music tracks of each mood tag.

4.2 Proposed Ensemble Methods of Voting in this study

Both empirical ensemble approaches in this section are a variation of the voting mechanism, as discussed in Subsection 2.5.10.

4.2.1 First Approach: Plurality Voting

The usage of Plurality Voting backs to 1871, when this method used as a parental vote to increase the birth rate in France [15]. Plurality Voting is a voting technique, which announces a candidate with the highest numbers of votes as a winner. This simple voting method has significant benefits when it is used as a combination method in an ensemble technique.

In this approach, the candidates are mood tags. Moreover, an ensemble technique uses several classifiers. Suppose there are five classifiers that classify a dataset. The result of each classifier shows the probability of choosing each tag for individual pieces of music. The tag with the highest probability is selected as the more relevant tag. To simplify the process of Plurality Voting, assume it is not possible to have two tags with the same probability value. In other words, there is always one tag which gets the highest probability. Other situations will be discussed later in this chapter. At the end, one tag wins the election for each piece of music. This is called the predicted tag.

If two or more tags have an equal chance to be selected, the algorithm selects the one which has a greater probability value. If two or more tags have the same probability value, the algorithm selects both of them. In this situation, minimum and maximum ranges calculate for the accuracy, because it is possible that one of them is the correct tag and others are incorrect, or all of them are incorrect tags.

Same as Bagging, all decisions have equal weight, but the main difference is that in Plurality Voting there is no limitation for the number of classifiers to choose a tag. In other words, in Plurality Voting, a tag could be assigned to a song when it obtained the highest number of votes, either more than half of the classifiers vote for it or less than half of the classifiers choose it. Note that calculating the accuracy using Plurality Voting method needs a clasifier, which produces some probability measures. The algorithm of plurality voting as an ensemble technique is shown as follows:

[15] http://www.ipod-library.net/articles/Plural_voting.

1. apply *n* classifiers on one piece of the dataset
2. count the number of suggested tags
3. **if** there is a tag which obtained the highest number of votes
4. choose it as the predicted tag
5. **else if** there are two or more tags which obtained the same number of votes
6. choose the tag with the highest probability value
7. **else if** the probability value of tags are equal
8. choose all tags with the same probability value

In order to obtain the final accuracy, we should count the number of music pieces, which are assigned tag correctly.

4.2.2 Second Approach: Borda Count

Borda Count is known as a social-choice theory. It is not clear who is the "founding father" of this theory, but it is noted that Borda Count was introduced in 18th-century by Jean-Charles de Bord, who was a French mathematician [89]. This voting method computes the probability of choosing a tag over some others. It is possible to have more than one tag with the same chance to be selected as the predicted tag. Therefore, to show the accuracy of this method, a minimum and maximum period is used. We used this technique in the MMC concept to improve the accuracy of mood classification in music data. The process of finding the winner(s) based on Borda Count is described below.

A group of classifiers is applied to each song in the dataset. Each classifier calculates the chance of each tag to be selected as the predicted tag. Table 4.1 demonstrates a sample result of applying *n* classifiers to the dataset, where candidates are different tags. After that, all tags are sorted in descending order for each piece of music. For example, the order of choosing tags by $Classifier$ 1 for one song of the dataset is $Happy$, *Relax*, *Sad*, and *Angry*, respectively. In the next step, the number of classifiers is counted, which produced the same order of tags for that song. Then, a score is calculated for each candidate in this way: if there are *n* candidates in the election, each candidate will get *n* points for the highest selection chance, *n*-1 points for the second highest selection chance, and so on.

Table 4.1: Sample results of applying n classifiers to one song of the dataset.

Classifiers / Place	Classifier 1	Classifier 2	.....	Classifier n
First	Happy (0.8)	Happy (0.9)		Relax (0.63)
Second	Relax (0.15)	Sad (0.1)		Happy (0.24)
Third	Sad (0.05)	Relax (0.0)		Angry (0.07)
Fourth	Angry (0.0)	Angry (0.0)		Sad (0.06)

Table 4.2: 350 voters participate in an election with 4 candidates.

Summation of Votes / Place	100	50	70	130
First	H	H	S	R
Second	R	S	A	H
Third	S	A	R	A
Fourth	A	R	H	S

To clarify this method, take a look at the example below:

Table 4.2 shows an example with 350 voters and 4 tags, H, S, A, R. The second column of the table indicates that 100 voters chose H as the first tag, and R, S, A were chosen as the second, third, and fourth tags, respectively. Since the number of candidates is four, 4 points are given to the first candidate, 3 points in the second one, and 2 and 1 to the third and fourth ones, respectively. In other words, a candidate gets 1 point when it has the lowest chance, 2 points when it is placed next-to-the-lowest chance, and so on. For example, the calculation of Table 4.2 is:

Candidate H: 4(100 + 50) + 3(130) + 2(0) + 1(70) = 1060

Candidate S: 4(70) + 3(50) + 2(100) + 1(130) = 760

Candidate A: 4(0) + 3(70) + 2(50 + 130) + 1(100) = 670

Candidate R: 4(130) + 3(100) + 2(70) + 1(50) = 1010

To calculate the final result for Candidate H, we consider how many times the music piece is tagged as H for Candidate 1 to Candidate 4. In the first row, Candidate 1, we can see the H under column 100 and 50. As discussed, Candidate 1 has 4 points multiplied to the summation of votes. In the second row, Candidate 2, we just have one H under column 130, and 3 points are given to this candidate. There is no H in the third row. And in the last row, candidate 4, we have H under column 70, and this candidate has 1 point. We can see Candidate H obtains the highest result compared to others. Therefore, the candidate H is the winner of the election since its Borda Count value is the highest among all candidates.

4.3 Experiments on Ensemble Techniques

To evaluate our proposed algorithms, two groups of five classifiers are chosen for applying to three datasets: 5120Mood, 2000Head, and 2000Tail. The results show the effectiveness of these algorithms, when compared to some existing combination methods in an ensemble technique. Participating classifiers in these experiments are BN, MLP, CVR, J48, and RandF for the first group, NB, SMO, DTable, Logistic, and RandT for the second group. The complete name of the classifiers was provided in Chapter 2 (Table 2.1). We chose these classifiers for two reasons. First, they were used by many researchers in MIR. Second, they represent each of the different main categories of WEKA classifiers, so we have a chance to see how different categories of WEKA classifiers behave in MMC. Among existing ensemble techniques, some of them are chosen to compare the results. WEKA contains three main ensemble techniques; namely, Boosting, Bagging, and Stacking. We introduced them in Chapter 2. In addition, three other combination methods - Minimum Probability, Maximum Probability, and Majority Voting - were selected to compare their accuracy with the proposed approaches in this study. In this study, experiments are conducted from two perspectives:

- Using different datasets which were presented.
- Applying various types of feature selection techniques which are independent of feature selection and optimal feature selection.

It is very important which classifiers are grouped together in an ensemble technique, since they may improve classification. The idea behind this kind of grouping is to use classifiers from different categories in each group, instead of choosing classifiers randomly, to get more accurate results.

4.3.1 Results for the 5120Mood Dataset

Table 4.4 shows the results of voting techniques on the 5120Mood dataset without applying any feature selection techniques. It is possible that using feature selection techniques may improve the outcome of classification, because more acoustic features in a dataset may confuse the classifiers. It is depicted that Plurality Voting and Borda Count methods reach acceptable results compared to other methods of combination, such as Min, Max, and Majority voting. However, these results are not acceptable since they are lower than in the accuracy of applying just simple classifiers in some cases. To illustrate, SMO, Logistic, and RandF achieve better results, rather than all combination methods. The logic behind the combination methods is that the accuracy of combined classifiers should be higher than the ones when applying single classifiers. Furthermore, there are some accuracies under single Stacking, group Stacking, Bagging and Boosting columns which confirm this conclusion.

Several abbreviations are used in the tables in this chapter, and the complete words and the corresponding abbreviations are shown in Table 4.3.

From the results of classification on this dataset, we can see in Table 4.5 that classifying the data after applying PCA to it obtained the highest accuracies in Plurality Voting and Borda Counting by 95.66% and 93.05%, respectively. These accuracies are greater than other current methods of combination which are implemented in this study, as shown in

Table 4.3: Complete words for used abbreviations in the tables of this chapter.

No.	Classifier Name	Abbreviation
1	Group 1	G1
2	Group 2	G2
3	Classifier	Cla
4	Accuracy	Acc
5	Single Stacking	SS
6	Group Stacking	GS
7	Bagging	Ba
8	Boosting	Bo
9	Minimum	Min
10	Maximum	Max
11	Majority	Maj
12	Plurality	Plu
13	Borda Count	Bor
14	Logistic	Log
15	RandF	RF
16	DTable	DT
17	RandT	RT

Table 4.4: Results of voting techniques on 5120Mood dataset without any feature selection.

	Cla	Acc	SS	GS	Ba	Bo	Min	Max	Maj	Plu	Bor
G 1	BN	42.66	36.76	42.71	43.16	42.66	16.1	43.59	44.47	49.93	49.41
	MLP	45.18	39.41		50.49	47.52					
	CVR	45.23	38.18		50.25	46.52					
	J48	40.21	30.61		45.55	45.8					
	RF	50.06	42.38		51.02	50.04					
G 2	NB	43.14	37.85	44.96	43.34	43.14	15.02	40.39	45.92	49.6	49.33
	SMO	52.19	44.79		51.95	52.19					
	DT	42.11	34.02		42.77	42.11					
	Log	51.64	46.09		51.93	51.64					
	RT	37.64	25		43.95	39.3					

Table 4.5: Results of voting techniques on 5120Mood dataset with PCA feature selection.

	Cla	Acc	SS	GS	Ba	Bo	Min	Max	Maj	Plu	Bor
G 1	BN	45.51	40.14	43.52	45.84	45.49	2.85	91.44	92.7	95.66	93.05
	MLP	44.88	35.63		50.96	46.31					
	CVR	42.62	36		49.36	45.14					
	J48	36.91	27.77		43.67	42.64					
	RF	49.69	41.78		50.76	49.86					
G 2	NB	39.67	36.92	45.98	39.69	39.67	8.3	96.45	59.3	64.64	67.46
	SMO	52.23	47.99		52.03	52.23					
	DT	42.36	33.07		45.02	42.36					
	Log	52.15	45.16		51.86	52.15					
	RT	33.57	25		40.49	36.97					

Table 4.6: Results of voting techniques on 5120Mood dataset with GR feature selection.

	Cla	Acc	SS	GS	Ba	Bo	Min	Max	Maj	Plu	Bor
G 2	NB	43.14	37.85	45	43.34	43.14	8.76	87.55	59.63	65.46	67.54
	SMO	52.21	45.68		51.95	52.21					
	DT	42.11	33.75		42.62	42.11					
	Log	51.64	46.09		51.93	51.64					
	RT	37.19	25		42.7	40.41					

Table 4.5. We conclude that Plurality Voting is more compatible with a PCA feature selection on this dataset. The results of this experiment, and other experiments that we will see in this section, shows that PCA could improve the accuracies of the proposed methods in this study.

Table 4.6 shows how the second group of classifiers produced the best results via the GR feature selection technique on the 5120Mood dataset. From our results, it is clear that the accuracy of Plurality Voting and Borda Counting vary considerably across the table. The only exception is that the outputs of these methods are lower than in the Max combination method. It is not necessarily a weakness of the proposed method, only because the combination of this group of classifiers generates these results. Generally, Plurality Voting and Borda Counting have better results compared to the other approaches of combining classifiers in ensemble techniques.

Table 4.7: Results of voting techniques on 2000Head dataset without any feature selection.

	Cla	Acc	SS	GS	Min	Max	Maj	Plu	Bor
G 1	BN	39.7	33.15	37.85	16	42.1	39.5	44.98	46.58
	MLP	42.85	34.05						
	CVR	43.65	37.2						
	J48	38.45	29.55						
	RF	45.45	37.65						
G 2	NB	40.85	33.3	39.65	15.15	38.9	41.2	46.9	47.88
	SMO	48.8	37.8						
	DT	39.95	33.05						
	Log	47.35	40.5						
	RT	36.65	25						

4.3.2 Results for the 2000Head Dataset

Table 4.7 depicts how the proposed approaches perform on the 2000Head dataset without any feature selection techniques. From the results, it can be seen that Plurality Voting and Borda Count achieve higher accuracies compared to all other combination methods in both groups of classifiers. However, some stand-alone classifiers reach comparable accuracies compared to Plurality Voting and Borda Count. For example, RandF in simple, Bagging, and Boosting states work better than Plurality Voting; also, MLP and CVR obtain the best results in Bagging situations, even more than Borda Counting in the first group of classifiers. However, SMO and Logistic from the second group achieve superior results compared to the proposed methods, while Plurality Voting and Borda Count are still preferred rather than the other combination methods.

Based on the given data in Table 4.8, the highest accuracy of Plurality Voting is 97.93%, when PCA applied to the data. As we discussed earlier, PCA had a positive effect on acoustic features in MMC. As shown in Table 4.8, Plurality Voting has improved. Moreover, Borda Counting obtains the best result after applying PCA too, by 95.98%.

Among the second group of classifiers, Plurality Voting has the best result, with 67.3% in two situations, first after applying IG (Table 4.9) and second after applying SYM (Table 4.10). For Borda Counting in the second group of classifiers, 69.7% is the highest accuracy

Table 4.8: Best results of Plurality Voting and Borda Counting on 2000Head dataset via PCA.

	Cla	Acc	SS	GS	Ba	Bo	Min	Max	Maj	Plu	Bor
G 1	BN	43.95	34.6	38.15	46.3	43.95	2.03	96.35	96.2	97.93	95.98
	MLP	41.75	33.8		47.95	44.85					
	CVR	42.2	33.35		46.3	44.45					
	J48	35.3	26.7		39.05	41.1					
	RF	45.55	36.85		48.15	46.6					
G 2	NB	39.8	37.05	39.65	40	39.8	6.33	97.88	60.2	65.85	69.63
	SMO	47.95	39.15		47.75	47.95					
	DT	38.7	31.75		41.75	38.7					
	Log	47.4	38.8		46.95	47.4					
	RT	32.7	25		39.6	32.8					

where the SYM (Table 4.10) feature selection is performed before classification, which is better than the Min and Majority methods, but lower than in Max approach. Table 4.9 and Table 4.10 indicate these outcomes.

As we see, in Table 4.9 and Table 4.10 the Max probability method obtained a high accuracy, around 90.5%. In the case of applying IG and SYM to the dataset, we suggest using the Max probability method, since it has a greater chance to predict the tag correctly compared to other methods.

Table 4.9: Best results of Plurality Voting on 2000Head dataset via IG.

Classifier	Accuracy	SS	GS	Ba	Bo	Min	Max	Maj	Plu	Bor
NB	40.85	33.3		40.75	40.85					
SMO	48.95	38.2		48.7	48.95					
DTable	39.65	33.8	36.75	42.15	39.65	7.4	90.53	60.85	67.3	69.65
Logistic	47.35	40.5		47.8	47.35					
RandT	34.45	25		41.75	36.1					

4.3.3 Results for the 2000Tail Dataset

The proposed approaches, Plurality Voting and Borda Count, also have been applied to the 2000Tail dataset. As previously noted, the first group of results is related to when no

Table 4.10: Best results of Plurality Voting and Borda Counting on 2000Head dataset via SYM.

Classifier	Accuracy	SS	GS	Ba	Bo	Min	Max	Maj	Plu	Bor
NB	40.85	33.3		40.75	40.85					
SMO	48.75	37.8		48.7	48.75					
DTable	39.85	34.3	39.45	42.15	39.85	7.38	90.53	60.85	67.3	69.7
Logistic	47.35	40.5		47.8	47.35					
RandT	35.8	25		42.45	36.85					

Table 4.11: Best results of Plurality Voting and Borda Counting on 2000Tail dataset without any feature selection.

	Cla	Acc	SS	GS	Ba	Bo	Min	Max	Maj	Plu	Bor
G 1	BN	43.7	34.55	42.1	43.5	43.7	16	43.78	42.95	48.38	48.9
	MLP	45.9	33.55		49.95	45.9					
	CVR	45.2	36.7		49.8	45.95					
	J48	38.1	28.8		44.45	43.65					
	RF	48.5	42.8		50.75	49.6					
G 2	NB	43.6	35.9	40.9	43.6	43.6	14.8	39.75	44.65	48.65	49
	SMO	52.15	42.65		51.1	52.15					
	DT	41.25	33.6		43.55	41.25					
	Log	50.95	44.95		50.65	50.95					
	RT	36.2	25		43.15	37.25					

feature selection technique is applied to the dataset. From Table 4.11, we see that generally the accuracy of the proposed methods is considerably better than the other combination methods in both groups of classifiers. Plurality Voting obtains an accuracy of more than 48%, and the Borda Count achieves an accuracy of around 49%. They are at least 4% higher than other implemented ensemble techniques.

The second experiment focused on obtaining the best outcomes after applying different feature selection methods. By looking at Table 4.12, we see that the Plurality Voting achieved the highest accuracy, around 96.85%, for the first group of classifiers. Although it is higher than Max method, around 1%, this improvement is still considerable when the accuracy of combination methods is greater than 90%. As shown in Table 4.12, Borda Counting, with 94.68%, has a lower result when compared to Max when PCA is applied

Table 4.12: Best results of Plurality Voting on 2000Tail dataset by applying PCA.

	Cla	Acc	SS	GS	Ba	Bo	Min	Max	Maj	Plu	Bor
G 1	BN	45.15	37.65	44.2	47.3	45.15	2.15	95.78	94.6	96.85	94.68
	MLP	43.1	37.35		50.4	46.3					
	CVR	44.75	36		48.7	44.8					
	J48	37.45	27		44	43.05					
	RF	47.65	42		50.4	49.25					
G 2	NB	42.65	37.05	45.35	42.6	42.65	7.03	96.95	62.2	68.1	69.5
	SMO	50.75	41.45		50.65	50.75					
	DT	43.95	33.95		43.7	43.95					
	Log	51.45	44.6		50.95	51.45					
	RT	33.6	25		40.95	33.8					

Table 4.13: Best results of the Borda Count on the 2000Tail dataset by applying IG.

	Cla	Acc	SS	GS	Ba	Bo	Min	Max	Maj	Plu	Bor
G 1	BN	43.7	34.55	41.05	43.5	43.7	7.53	77.58	93.4	95.2	93.63
	MLP	45	36.2		50.7	44.85					
	CVR	43.45	37.75		49.2	47.2					
	J48	36.7	28.65		43.65	44.25					
	RF	48.5	39.85		50.5	48.55					
G 2	NB	43.6	35.9	44.7	43.6	43.6	9.2	84.78	63.2	70.2	70.43
	SMO	52.05	42.95		51.15	52.05					
	DT	40.85	35.35		43.55	41.05					
	Log	50.95	44.95		50.65	50.95					
	RT	37.15	25		44.1	37.85					

to the dataset, but it is better than all other methods when IG was chosen as the feature selection, which is depicted in Table 4.13. The following two tables, Table 4.12 and Table 4.13, are the best results for the first group of classifiers.

To compare the results of combining the second group of classifiers, Table 4.14 presents the highest accuracy of Plurality Voting, which is 70.25%, when the SYM feature selection is applied to the dataset. This is less than the Max value, and the reason is that classifiers in the second group cannot work well together. Therefore, changing one or more classifier(s) may considerably increase the accuracy of Plurality Voting. As for the Borda Count combination method, it is noticeable that this approach reached the best results, 70.43%, by

Table 4.14: Best results of both proposed methods on 2000Tail dataset by applying SYM.

	Cla	Acc	SS	GS	Ba	Bo	Min	Max	Maj	Plu	Bor
G 1	BN	43.7	34.55	41.9	43.5	43.7	7.2	78.33	94.55	96.13	93.63
	MLP	45.3	37.85		50.05	44.4					
	CVR	43.05	36.75		49.55	45.8					
	J48	36.75	29.5		43.3	44.45					
	RF	50.75	42.15		49.55	48.95					
G 2	NB	43.6	35.9	42.15	43.6	43.6	9.2	84.78	63.25	70.25	70.43
	SMO	52.1	42.7		51.15	52.1					
	DT	40.85	35.3		43.55	41.05					
	Log	50.95	44.95		50.65	50.95					
	RT	37.7	25		43.85	35.7					

Table 4.15: Best results of Borda Count on 2000Tail dataset by applying GR.

	Cla	Acc	SS	GS	Ba	Bo	Min	Max	Maj	Plu	Bor
G 1	BN	43.7	34.55	39.45	43.5	43.7	7.3	77.98	94.45	96.33	93.93
	MLP	45.7	37.2		51	45.7					
	CVR	43.3	37.5		49.95	45.05					
	J48	36.85	29.55		43.5	44.25					
	RF	48.8	37.4		50.15	49.55					
G 2	NB	43.6	35.9	40.55	43.6	43.6	9.23	84.78	63.25	70.2	70.43
	SMO	52	43.05		51.15	52					
	DT	40.85	34.5		43.55	41.05					
	Log	50.95	44.95		50.65	50.95					
	RT	35.65	25		44.95	36.45					

applying SYM, GR, as well as IG.

Based on the given data in Table 4.14, Table 4.15, and Table 4.13, there are no substantial changes in the results of the classifiers. In other words, the accuracy of individual classifiers - Stacking, Bagging, Boosting, Min, Max, Majority Voting, Plurality Voting, and Borda Count - have a slight change or remain unchanged. It means that the performances of these three feature selection techniques (SYM, GR, and IG), which use the same searching method, are very similar on this dataset.

4.4 Discussions

Mood classification is one of the hardest problems in MIR since it is related to many factors. It is the reason why not many researchers work in this area of music classification. Several previous studies were discussed in Chapter 2 and Chapter 3, but the number of studies in MMC is lower than the number of studies in genre and style classification in MIR. In this chapter, two combination methods were introduced for ensemble techniques. The idea of using several classifiers in classification has become more popular in the recent years in MIR. As discussed in this chapter, using a combination of classifiers has many benefits, such as the growth in accuracy of classification. After considering the outcomes of all experiments in Section 4.3, it can be seen that the two new proposed methods for combining several classifiers in ensemble technique can significantly improve the results. The proposed methods reached the best results, when compared to other combination methods. Also, they achieve better results over the ensemble techniques in WEKA, such as bagging, boosting, and stacking. The key issue is which groups of classifiers are appropriate to combine. In general, the proposed methods are acceptable when compared to other methods and algorithms.

Chapter 5

Conclusion

Automatic classification of music tracks is one of the biggest challenges in recent years, especially for music recommender systems and websites. There are many applications in this area to manage music data and classify them, but some principal issues still remain, particularly in MMC. One reason is that MMC is affected by many items from acoustic features to psychological effects to emotional situations, so they make this task more complex. More complications increase the requirement of having an automatic tag annotation system for MMC. In this study, automatic tag annotation mechanisms of MMC are investigated from a computational viewpoint.

5.1 Contribution

The main goal of this study is to improve the accuracy of MMC. To achieve this end, some machine learning algorithms and data mining tools were adopted to increase the performance of classification. In this thesis, experiments were executed on several benchmarks in MIR to achieve the most reliable results. At the first step, some preliminary experiments were done to find what types of classifiers, feature selection techniques, dimensionality reduction methods, and artificial intelligence algorithms work better in MMC, as showen in Chapter 3. Thus, some of them were selected for advanced experiments in Chapter 4. The selected methods contain:

- discretization techniques as a preprocessing task.
- CFS, IG, GR, Chi, as well as SYM techniques for selecting features.

- PCA and PLS as dimensionality reduction methods.
- NB, HNB, NBU, BN, DT, DTable, CVR, Logistic, MLP, SMO, RandF, RandT, JRip, MCC, AODE, WAODE, RS, Ridor, LogitBoost, LADTree, DTNB, Bagging, Boosting, Dagging, Stacking, RF, and END as classification algorithms.
- Cross-validation as evaluation method.

In addition, two proposed approaches were introduced in Chapter 4 and are focused on combining several classifiers in an ensemble technique to improve the accuracy of the MMC.

5.1.1 Perusing Different Approaches

At the beginning of this thesis before performing any experiments, there were some difficulties that needed to be solved. The most important one was related to data. The lack of reliable music tracks with assigned emotional tags is a major problem in mood classification in MIR. Most of the quality datasets do not have any mood tag. Sometimes users attached mood tags to the pieces of music in the datasets, but they may not be acceptable. It is because users assigning tags needs overall knowledge about music, while sometimes they do not have enough knowledge related to music specifications of a song. We needed to find a reliable data for our experiments, so some famous benchmarks were selected to create some new and more precise datasets. Another basic problem was related to finding the best features and feature sets. In order to tackle this problem, some previous works in different areas of music classification were studied, such as genre classification, mood classification, and musicology papers. Based on the gathered information on those studies, some groups of feature sets were selected, and a series of experiments were conducted in Chapter 3 to test the suitability of them.

In addition, some of these sets contained noisy and redundant data, so they are not useful for finding the most relevant feature sets. Therefore, feature selection techniques and dimensionality reduction methods were used to refine the features by different methods, such

as ranking or searching. After that, classifiers were applied to a dataset that contains the highest ranked acoustic features. It depends on the nature of experiments. In some cases, the accuracy of each classifier should be compared to others, while in others, a group of several classifiers is applied to a dataset using a combination method and the accuracy of the combination method should be compared to the accuracy of individual classifiers. Lastly, evaluation techniques were used to test our classification strategies.

By consideration of previous work in Chapter 2 and the consequences of the primary experiments in Chapter 3, we see that the current classification methods could not achieve acceptable results in mood classification on music data. After finding the weaknesses, three approaches were proposed in Chapter 4 to provide some solutions for the MMC problem in MIR, and they result in significantly improved accuracies.

5.1.2 Improving Accuracy of Mood Classification

Two approaches were proposed to improve the accuracy of the MMC, which are focused on improving the accuracy of ensemble methods. As discussed in the previous chapters, the ensemble technique applies several classifiers to a dataset. The accuracy is based on the performance of the classifiers in this technique. The results show that the existing ensemble technique does not achieve reasonable accuracy by combining current methods in the MMC.

We proposed two new ensemble techniques for combining the classifiers. Both ideas come from statistics and are variants of voting methods. The first idea, Plurality Voting, combines several classifiers by using an optimistic perspective. All classifiers were applied to the datasets to predict a tag for each piece of music. Finally, a tag that obtained more votes or the highest probability is selected as the mood tag for the music piece. In the second idea, Borda Count, the final tag was selected by computing a formula based on the probability value vector, which is a weighting algorithm.

5.2 Future Work

Music covers a wide spectrum of acoustic features, where each of them is a world with its own difficulties and complexities. More concentration on audio features has definite positive effects in music classification.

Another important part of MIR which needs to be taken into account is gathering real-sized audio repositories which include reliable acoustic features. One of the most serious problems in MIR is the lack of reliable data. A more technical attempt can be finding more accurate ways to combine several classifiers. As an example, there is no research to show that how many and what type of classifiers should be applied to reach the highest accuracy on a dataset.

To improve the efficiency of this work, we plan to use more classification algorithms and investigate other combination methods in ensemble techniques. Further focus on the acoustic features and classification measures may help us to achieve more accurate results on mood classification in MIR. Another idea is to use more musical knowledge to classify genre and instrument classification. Some feature selection techniques, dimensionality reduction methods, and classifiers were used in this study, while other techniques may promote the performance of the MMC.

Bibliography

[1] Amélie Anglade, Emmanouil Benetos, Matthias Mauch, and Simon Dixon. Improving music genre classification using automatically induced harmony rules. *Journal of New Music Research*, 39(4):349–361, 2010.

[2] Hasitha Bimsara Ariyaratne, Dengsheng Zhang, and Guojun Lu. A class centric feature and classifier ensemble selection approach for music genre classification. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 666–674. Springer, 2012.

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

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

[5] Jeronimo Barbosa, Cory McKay, and Ichiro Fujinaga. Evaluating automated classification techniques for folk music genres from the brazilian northeast. *Computer Music: Beyond the frontiers of signal processing and computational models*, 2015.

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

[7] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of International Society of Music Information Retrieval*, volume 2, page 10, 2011.

[8] Jennifer Bollerman. New grove dictionary of music and musicians online. *Reference & User Services Quarterly*, 40(4):377–377, 2001.

[9] Zehra Cataltepe, Yusuf Yaslan, and Abdullah Sonmez. Music genre classification using midi and audio features. *European Association for Signal Processing Journal on Advances in Signal Processing*, 2007(1):1–8, 2007.

[10] Weiwei Cheng and Eyke Hüllermeier. Combining instance-based learning and logistic regression for multilabel classification. *Proceedings of Machine Learning*, 76(2-3):211–225, 2009.

[11] Sander Dieleman, Philémon Brakel, and Benjamin Schrauwen. Audio-based music classification with a pretrained convolutional network. In *Proceedings of 12th International Society of Music Information Retrieval Conference*, pages 669–674. University of Miami, 2011.

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

[13] James Dougherty, Ron Kohavi, Mehran Sahami, et al. Supervised and unsupervised discretization of continuous features. In *Machine Learning: Proceedings of the Twelfth International Conference*, volume 12, pages 194–202, 1995.

[14] J Stephen Downie. Music information retrieval. *Annual review of information science and technology*, 37(1):295–340, 2003.

[15] Antti Eronen and Anssi Klapuri. Musical instrument recognition using cepstral coefficients and temporal features. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages II753–II756. IEEE, 2000.

[16] Yazhong Feng, Yueting Zhuang, and Yunhe Pan. Music information retrieval by detecting mood via computational media aesthetics. In *Proceedings of IEEE/WIC International Conference on Web Intelligence*, pages 235–241. IEEE, 2003.

[17] Rebecca Fiebrink and Ichiro Fujinaga. Feature selection pitfalls and music classification. In *Proceedings of the International Society on Music Information Retrieval*, pages 8–12, 2006.

[18] Rebecca Fiebrink, Cory McKay, and Ichiro Fujinaga. Combining d2k and jgap for efficient feature weighting for classification tasks in music information retrieval. In *Proceedings of International Society of Music Information Retrieval*, pages 510–513, 2005.

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

[20] Mark A Hall and Lloyd A Smith. Practical feature subset selection for machine learning. Springer, 1998.

[21] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[22] Xiao Hu and J Stephen Downie. Exploring mood metadata: Relationships with genre, artist and usage metadata. In *Proceedings of International Society of Music Information Retrieval*, pages 67–72, 2007.

[23] Xiao Hu and J Stephen Downie. Improving mood classification in music digital libraries by combining lyrics and audio. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, pages 159–168, 2010.

[24] Xiao Hu and J Stephen Downie. When lyrics outperform audio for music mood classification: A feature analysis. In *Proceedings of 11th International Society of Music Information Retrieval*, pages 619–624, 2010.

[25] Xiao Hu and Jin Ha Lee. A cross-cultural study of music mood perception between american and chinese listeners. In *Proceedings of International Society of Music Information Retrieval*, pages 535–540, 2012.

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

[27] Gabriela Husain, William Forde Thompson, and E Glenn Schellenberg. Effects of musical tempo and mode on arousal, mood, and spatial abilities. *Music Perception: An Interdisciplinary Journal*, 20(2):151–171, 2002.

[28] Sudhir B Jagtap et al. Census data mining and data analysis using weka. *arXiv preprint arXiv:1310.4647*, 2013.

[29] Adit Jamdar, Jessica Abraham, Karishma Khanna, and Rahul Dubey. Emotion analysis of songs based on lyrical and audio features. *International Journal of Artificial Intelligence and Applications(IJAIA)*, 6(3), 2015.

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

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

[32] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques. *An International Journal of Computing and Informatics (Informatica)*, 31(1):249–268, 2007.

[33] Cyril Laurier, Jens Grivolla, and Perfecto Herrera. Multimodal music mood classification using audio and lyrics. In *Proceedings of Seventh International Conference on Machine Learning and Applications.*, pages 688–693. IEEE, 2008.

[34] Cyril Laurier, Olivier Lartillot, Tuomas Eerola, and Petri Toiviainen. Exploring relationships between audio features and emotion in music. *Proceedings of the 7th Triennial Conference of European Society for the Cognitive Sciences of Music*, 2009.

[35] Cyril Laurier, Mohamed Sordo, Joan Serra, and Perfecto Herrera. Music mood representations from social tags. In *Proceedings of International Society of Music Information Retrieval*, pages 381–386, 2009.

[36] Teresa Lesiuk. The effect of music listening on work performance. *Psychology of music*, 33(2):173–191, 2005.

[37] Tao Li and Mitsunori Ogihara. Detecting emotion in music. In *Proceedings of International Society of Music Information Retrieval*, volume 3, pages 239–240, 2003.

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

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

[40] Shin-Cheol Lim, Jong-Seol Lee, Sei-Jin Jang, Soek-Pil Lee, and Moo Young Kim. Music-genre classification system based on spectro-temporal features and feature selection. *IEEE Transactions on Consumer Electronics*, 58(4):1262–1268, 2012.

[41] Beth Logan et al. Mel frequency cepstral coefficients for music modeling. In *Proceedings of International Society of Music Information Retrieval*, 2000.

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

[43] Marcus-Christopher Lud and Gerhard Widmer. Relative unsupervised discretization for association rule mining. In *Proceedings of European Conference on Principles of Data Mining and Knowledge Discovery*, pages 148–158. Springer, 2000.

[44] Michael I Mandel and Dan Ellis. Song-level features and support vector machines for music classification. In *Proceedings of International Society of Music Information Retrieval*, volume 2005, pages 594–599, 2005.

[45] Michael I Mandel and Daniel PW Ellis. Multiple-instance learning for music information retrieval. In *Proceedings of International Society of Music Information Retrieval*, pages 577–582, 2008.

[46] Benoit Mathieu, Slim Essid, Thomas Fillon, Jacques Prado, and Gaël Richard. Yaafe, an easy to use and efficient audio feature extraction software. In *Proceedings of International Society of Music Information Retrieval*, pages 441–446, 2010.

[47] Brian McFee, Thierry Bertin-Mahieux, Daniel PW Ellis, and Gert RG Lanckriet. The million song dataset challenge. In *Proceedings of the 21st International Conference on World Wide Web*, pages 909–916. ACM, 2012.

[48] Cory McKay, Rebecca Fiebrink, Daniel McEnnis, Beinan Li, and Ichiro Fujinaga. Ace: A framework for optimizing music classification. In *Proceedings of International Society of Music Information Retrieval*, pages 42–49, 2005.

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

[50] Cory McKay and Ichiro Fujinaga. 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.

[51] Cory McKay, Ichiro Fujinaga, and Philippe Depalle. jaudio: A feature extraction library. In *Proceedings of the International Conference on Music Information Retrieval*, pages 600–603, 2005.

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

[53] David Moffat, David Ronan, and Joshua D Reiss. An evaluation of audio feature extraction toolboxes. *Proceedings of the 18th International Conference on Digital Audio Effects.*, 2015.

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

[55] Mohsen Naji, Mohammd Firoozabadi, and Parviz Azadfallah. Classification of music-induced emotions based on information fusion of forehead biosignals and electrocardiogram. *Cognitive Computation*, 6(2):241–252, 2014.

[56] Aziz Nasridinov and Young-Ho Park. A study on music genre recognition and classification techniques. *International Journal of Multimedia and Ubiquitous Engineering*, 9(4):31–42, 2014.

[57] Steven R Ness, Anthony Theocharis, George Tzanetakis, and Luis Gustavo Martins. 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.

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

[59] Nicola Orio and Roberto Piva. Combining timbric and rhythmic features for semantic music tagging. In *Proceedings of International Society of Music Information Retrieval.*, pages 77–82, 2013.

[60] Braja Gopal Patra, Dipankar Das, and Sivaji Bandyopadhyay. Automatic music mood classification of hindi songs. In *Proceedings of Sixth International Joint Conference on Natural Language Processing*, page 24, 2013.

[61] Tim Pohle, Elias Pampalk, and Gerhard Widmer. Evaluation of frequently used audio features for classification of music into perceptual categories. In *Proceedings of the Fourth International Workshop on Content-Based Multimedia Indexing*, volume 162, 2005.

[62] Christopher Rea, Pamelyn MacDonald, and Gwen Carnes. Listening to classical, pop, and metal music: An investigation of mood. *Emporia state research studies*, 46(1):1–3, 2010.

[63] Chris Sanden and John Z Zhang. 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.

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

[65] Klaus R Scherer. Which emotions can be induced by music? what are the underlying mechanisms? and how can we measure them? *Journal of New Music Research*, 33(3):239–251, 2004.

[66] Klaus R Scherer and Marcel R Zentner. Emotional effects of music: Production rules. *Proceedings of Music and emotion: Theory and research*, pages 361–392, 2001.

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

[68] William A Sethares. *Tuning, timbre, spectrum, scale*. Springer Science & Business Media, 2005.

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

[70] Carlos N Silla Jr, Celso AA Kaestner, and Alessandro L Koerich. Automatic music genre classification using ensemble of classifiers. In *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, pages 1687–1692. IEEE, 2007.

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

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

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

[74] Yading Song, Simon Dixon, and Marcus Pearce. Evaluation of musical features for emotion classification. In *Proceedings of International Society of Music Information Retrieval*, pages 523–528. Citeseer, 2012.

[75] Efstathios Stamatatos and Gerhard Widmer. Automatic identification of music performers with learning ensembles. *Artificial Intelligence*, 165(1):37–56, 2005.

[76] Jérôme Sueur, Thierry Aubin, and Caroline Simonis. Equipment review: seewave, a free modular tool for sound analysis and synthesis. *Proceedings of International Journal of Animal Sound and its Recording.*, 18(2):213–226, 2008.

[77] Konstantinos Trohidis, Grigorios Tsoumakas, George Kalliris, and Ioannis P Vlahavas. Multi-label classification of music into emotions. In *Proceedings of International Society of Music Information Retrieval*, volume 8, pages 325–330, 2008.

[78] Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet. Towards musical query-by-semantic-description using the cal500 data set. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 439–446. ACM, 2007.

[79] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302, 2002.

[80] Nicolas Wack, C Laurier, O Meyers, R Marxer, Dmitry Bogdanov, Joan Serra, E Gomez, and Perfecto Herrera. Music classification using high-level models. *Music Information Retrieval Evaluation Exchange (MIREX10)*, 2010.

[81] Jun Wang, Xiaoou Chen, Yajie Hu, and Tao Feng. Predicting high-level music semantics using social tags via on-tology-based reasoning. 2010.

[82] Shuo-Yang Wang, Ju-Chiang Wang, Yi-Hsuan Yang, and Hsin-Min Wang. Towards time-varying music auto-tagging based on cal500 expansion. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2014.

[83] I.H. Witten, E. Frank, and M.A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2011.

[84] Changsheng Xu, Namunu C Maddage, Xi Shao, Fang Cao, and Qi Tian. Musical genre classification using support vector machines. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal.*, volume 5, pages 429–432. IEEE, 2003.

[85] Changsheng Xu, Namunu Chinthaka Maddage, and Xi Shao. Automatic music classification and summarization. *IEEE transactions on speech and audio processing*, 13(3):441–450, 2005.

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

[87] Yusuf Yaslan and Zehra Cataltepe. Audio music genre classification using different classifiers and feature selection methods. In *Proceedings of 18th International Conference on Pattern Recognition (ICPR'06)*, volume 2, pages 573–576. IEEE, 2006.

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

[89] Manzoor Ahmad Zahid and Harrie De Swart. The borda majority count. *Information Sciences*, 295:429–440, 2015.

[90] Xiatian Zhang, Quan Yuan, Shiwan Zhao, Wei Fan, Wentao Zheng, and Zhong Wang. Multi-label classification without the multi-label cost. In *Proceedings of the SIAM International Conference on Data Mining*, volume 10, pages 778–789. SIAM, 2010.

[91] Yibin Zhang and Jie Zhou. A study on content-based music classification. In *Proceedings of Seventh International Symposium on Signal Processing and Its Applications.*, volume 2, pages 113–116. IEEE, 2003.

[92] Yibin Zhang and Jie Zhou. Audio segmentation based on multi-scale audio classification. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal.*, volume 4. IEEE, 2004.

[93] Yibin Zhang, Jie Zhou, and Zhaoqi Bian. Content-based classification and analysis on chinese traditional opera. *Computer Engineering*, 12:183–186, 2006.