Personal and Ubiquitous Computing (2022) 26:1237–1249
https://doi.org/10.1007/s00779-020-01393-4

ORIGINAL PAPER

Emotional classification of music using neural networks with the MediaEval dataset

Yesid Ospitia Medina[1,2] · José Ramón Beltrán[3] · Sandra Baldassarri[3]

Received: 7 December 2019 / Accepted: 7 March 2020 / Published online: 15 April 2020

Abstract
The proven ability of music to transmit emotions provokes the increasing interest in the development of new algorithms for music emotion recognition (MER). In this work, we present an automatic system of emotional classification of music by implementing a neural network. This work is based on a previous implementation of a dimensional emotional prediction system in which a multilayer perceptron (MLP) was trained with the freely available MediaEval database. Although these previous results are good in terms of the metrics of the prediction values, they are not good enough to obtain a classification by quadrant based on the valence and arousal values predicted by the neural network, mainly due to the imbalance between classes in the dataset. To achieve better classification values, a pre-processing phase was implemented to stratify and balance the dataset. Three different classifiers have been compared: linear support vector machine (SVM), random forest, and MLP. The best results are obtained with the MLP. An averaged F-measure of 50% is obtained in a four-quadrant classification schema. Two binary classification approaches are also presented: one vs. rest (OvR) approach in four-quadrants and binary classifier in valence and arousal. The OvR approach has an average F-measure of 69%, and the second one obtained F-measure of 73% and 69% in valence and arousal respectively. Finally, a dynamic classification analysis with different time windows was performed using the temporal annotation data of the MediaEval database. The results obtained show that the classification F-measures in four quadrants are practically constant, regardless of the duration of the time window. Also, this work reflects some limitations related to the characteristics of the dataset, including size, class balance, quality of the annotations, and the sound features available.

Keywords Music emotion recognition (MER) · Emotion classification · Prediction · Music features · Multilayer perceptron

1 Introduction

In the last years, the music industry has been experiencing many important changes as a result of new user requirements and the wide range of possibilities offered by emerging devices and technologies [12]. These technologies allow users to access huge databases of musical pieces through different kind of applications. The facility of creation, accessing, and distributing music, as well as the effectiveness of search engines on musical repositories are current challenges of music industry, with different stakeholders, such as composers, producers, and emerging artists, waiting for innovative solutions [41]. The main features of digital music consumption platforms, such as Spotify, Youtube music, or Deezer, are closely related to the way they present their contents and allow access to them. In many cases, recommender system strategies are applied in order to help listeners explore large music repositories in order to suggest songs according to their requirements and preferences. However, knowing users' taste it is not enough to recommend a suitable song for a person in a particular moment. Moreover, it must be taken into account that music is considered an art that can produce emotional responses or induce listeners' emotions [8, 36]. This close connection

✉ Yesid Ospitia-Medina
yesid.ospitiam@info.unlp.edu.ar

José Ramón Beltrán
jrbelbla@unizar.es

Sandra Baldassarri
sandra@unizar.es

1 Universidad Nacional de La Plata, La Plata, Argentina

2 Universidad Icesi, Cali, Colombia

3 Universidad de Zaragoza, 50004 Zaragoza, Spain

between music and emotions is explained as the relationship between the diverse intrinsic musical features (such as tone, mode, and tempo) and the emotional perception of the listener [31]. Therefore, although music is typically classified by genre [22, 32] or considering cultural aspects [45], recent studies suggest that people actually choose musical pieces according to their mood [17, 35], in some cases to reaffirm an emotional experience of the moment and in others to counteract it [38]. For these reasons, and the astounding growth of musical platforms and applications, music emotion recognition (MER) has become an increasingly popular research topic in order to identify the strategies to provide good solutions that improve the listeners' user experience [1, 35]. The automatic recognition of emotions in music, as a field of research, requires a deep analysis of music content and features. The features used in MER can be classified as low level (associated with signal and sound processing) or high level (associated with musical elements such as tone, mode, and tempo). There are different software applications and libraries that can be used to extract the audio features of the songs. Depending on the advanced level of use of these libraries, they can be classified in the same way in low and high level, whereas the level is related to necessary advanced technical knowledge [26]. In order to detect and recognize emotional information from the contents, the values of the features must be analyzed and processed to determine and classify the song within specific emotional categories or classes [30]. Different combinations of feature values are normally related to output values that can be of approximate value or class, depending on the emotional classification model selected [10], that can be categorical, in which emotions are described with a discrete number of classes, or dimensional, in which emotions are described as numerical values of valence and arousal.

In this paper, an emotional classification system for the songs of the MediaEval dataset [39] has been designed and implemented by using neural networks. As a first approach, a dimensional prediction system was developed, but the results obtained led us to analyze the characteristics of the musical data in the MediaEval database. It was observed that the dataset was not balanced in the dimensional space, limiting a good accuracy in classification. So that, data selection and data balance strategies have been incorporated in our system in order to get a suitable dataset for the classification tasks. Moreover, as the MediaEval database provides a dynamic temporal evaluation of the emotional annotation and the features of each song, we have also analyzed the emotional categorical classification regarding the analysis of temporal window.

The paper is organized as follows: Section 2 presents some relevant work on the recognition of emotions considering prediction and classification approaches. Section 3 offers a description of the dataset of musical pieces as well as its process of labeling the emotions. Section 4 presents the system developed for emotion prediction in music and its limitations. Section 5 shows the design and implementation of the classification system. In Section 6, the results obtained and a discussion of the main problems and challenges detected are exposed. Finally, Section 7 highlights the conclusions obtained from this work and the proposed future work.

2 Related work

As was previously mentioned, MER is a multidisciplinary field of research that has been widely expanded in the last years. Although there are many works in this field [19, 42, 43], it is not easy to compare their performance because of their methodological differences in data representation, annotation, selection of features, and emotional models, that leads to different evaluation metrics, making almost impossible to compare the accuracy of the algorithms applied [1]. Additionally, the methods and experiments proposed in different works are very difficult to replicate or benchmark since most of them use private datasets or different public datasets, with not enough songs and different low and high-level features [27].

Despite the problems for comparing the different works, in the following, a selection of MER systems are analyzed, depending on whether it is a prediction or a classification approach. A brief summary of the selected emotional prediction systems is presented in Table 1 while emotional classifications systems are presented in Table 2. These tables show the most important characteristics of the dataset: number of songs, length of each song, number of sound features, state of data balancing, and kind of annotation (static or dynamic), but also, classification or prediction techniques and their respective evaluation metrics. In the case of classification system (Table 2), the classes used in the work are also specified: quadrant, cluster and, in some cases, average of all quadrants.

Although there are some works that are focused only on prediction like [11, 14, 29] or [7], and that are some others centered only on solving classification problems like [13, 28, 44]; there are also several works that make predictions and later extend their systems to achieve classifications, such as [37] or [2]. Generally, these works that include prediction and classification use datasets annotated in dimensional models, in which a valence and arousal (V/A) coordinate system is established [34]. Additionally, some of these works implement a data pre-processing phase, but

Table 1 Emotional prediction systems. reference metrics: root mean square error (RMSE), averaged random distance (ARD), and determination coefficient (R^2)

Article	Songs	length	Features	Balanced	Annotation	C. technique	RMSE	ARD	R^2
Fernandes [11]	194	25 s	454	No	Static	SVM	[V : 0.24, A : 0.22]	–	–
Schmidt [37]	240	15 s	–	–	Dynamic	SVR	–	0.238	–
Panda [29]	189	25 s	556	No	Static	SLR, KNN, SVR	–	–	[V:40.6%, A:67.4%]
Bai [2]	744	45 s	548	–	Static	SVR, RFT, PCA	–	–	[V:29.3%, A:62.5%]
Grekow [14]	324	6 s	654	Yes	Static	SMOreg	–	–	[V:58%, A:79%]
Hennequin [7]	18,644	30 s	–	–	Static	ConvNet	–	–	[V:17.9%, A:23.5%]

only aimed to identify classes (quadrants and clusters), although this may not be necessary if the database is also categorically annotated with a discrete set of emotions [10].

The type of annotation of the emotions is another important criterion of comparison, as well as a key success factor in the field of MER. Although there are some works that consider the composer intention [15], in most of the cases, the annotations are carried out according to the emotional perception of the listener. Among these works, there are mainly two approaches for annotating the songs: static and dynamic. In the static annotation process, the user sets one value of valence and one for arousal for indicating the user's emotional perception, more related to the mood response to the music. In a dynamic annotation process, the user gives a dynamical temporal evaluation about his/her emotional perception, in a continuous way. Most of the emotional prediction and classification works are based on static annotation [2, 7, 11, 14]. On the other hand, Schmidt et al. used dynamic annotation, with a time window that varies from 2 to 15 s [37]. However, there are some works that apply temporal dynamic annotation, but this continuous annotation is later averaged, and therefore, at the end, there is only one value of valence and arousal. This is the case of the work of Panda et al. [29], in which dynamic annotation is implemented, but these annotations are averaged and used as a single global value within the prediction model, so that the work is classified as "Static annotation".

Besides the dynamic or static approach for annotation, for obtaining good results in a classification system and, in particular, in music emotion classification, it is fundamental to have a large number of data, and moreover, with almost the same amount of data for each emotion. Data balancing is a fundamental issue in classifying systems, because it determines the values obtained in the success rates [23].

Generally, in the majority of the prediction systems, the amount of data is a very important criterion because it determines its generalization capacity. However, if the availability of data per categorical annotation is not almost equally distributed, the minority classes has a negative influence on the performance of a classifying system, because automatic learning techniques have a tendency to specialize in the prediction of majority classes.

The information about the data balancing in the studied works is also shown in Tables 1 and 2. The value in the balanced column indicates the original status of the data balancing before any pre-processing phase. It can be observed that for some works no details on data balancing are incorporated, especially in prediction systems (see Table 1) in which it has no major relevance. Regarding classification systems (see Table 2), the data of the works [13, 28] and [44] are balanced and do not require any special treatment. Instead, [27] presents unbalance data, and although there is no detail of the strategy used to balance the data, apparently some songs are removed from the original dataset. And, finally, there are other works without information about the balance of the data [2, 37].

Regarding the results of the revised MER works, it can be observed that either for prediction (Table 1) or classification (Table 2) success rates vary from low to average. There are only a few works that have higher values, but it is difficult to get any conclusion, due to the lack of uniformity in the datasets and the unbalanced nature of the majority of the public available datasets.

In this work, the MediaEval dataset has been employed and a brief description of it is given in the next section. This database has been chosen because it has many songs, and they are annotated dynamically and with dimensional parameters.

3 Music dataset

The dataset used in this work is the free available MediaEval one [39]. It contains 1802 songs in MPEG layer 3 (MP3) format, with a sampling frequency of 44100 Hz, each one with values of 260 low-level features. Each song is analyzed

Table 2 Emotional classification systems. Reference metrics: accuracy, F-measure

Article	Songs	Length	Features	Balanced	Annotation	C. technique	Classes	Accuracy	F-measure
Schmidt [37]	240	15 s	–	–	Dynamic	SVR	4 quadrants (averaged)	50.18%	–.
Panda [28]	903	30 s	253	Yes	Static	SVM	5 cluster	–	[0.37, 0.37, 0.61, 0.40, 0.53]
Grekow [13]	324	6 s	471	Yes	Static	SMO	4 quadrants	[0.81, 0.90, 0.87, 0.77]	[–,0.72, 0.65, 0.54]
Zhang [44]	400	35 s	–	Yes	Static	RFC	4 quadrants (averaged)	83.29%	–
Bai [2]	744	45 s	548	–	Static	SVR, RFT, PCA	4 quadrants (averaged)	59.2%	–
Panda [27]	900	30 s	898	No	Static	SVM	4 quadrants	–	[0.77, 0.85, 0.71, 0.68]

for 45 s and the value of each feature is extracted every 0.5 s (500 ms). The process of extracting the sound features was carried out through the openSMILE[1] software.

The MediaEval dataset stores the values of the emotional perception of each song with a dimensional emotional model, annotating a value of valence and arousal (V/A). The annotation process is done with a dynamic approach, which means that every 500 ms the user indicates his/her emotional perception by setting a V/A coordinate. Although users have to annotate their emotions, in many cases, it seems that they indicate their mood, as will be later commented. It is important to highlight that each song is annotated by multiple users, so the dynamic annotation is available of each user, as well as the averaged annotation of all users for each time window length of a particular song. Also, for each song, there is an averaged V/A value. The process of calculating this averaged V/A implies, first, to average the annotations of all the annotators for each time window and then to calculate the average of all the resulting V/A values for all time window length. In this way, each final annotation could be related to the average value of each sound feature, which is calculated from all time windows.

The amount and distribution of the musical pieces on the V/A emotional plane is shown in Fig. 1. As can be observed, there is a clear unbalanced data distribution of the songs between quadrants. This situation generally represents a problem in classification algorithms and this fact will be reviewed in detail in Sections 4 and 5.

4 Dimensional emotion prediction

The main objective of a dimensional emotional prediction system is to provide the value of valence and arousal, depending on the low-level features of the songs. Once that value is obtained, it is desirable to properly identify the corresponding emotion in the V/A plane. This implies, therefore, the need to assess the success rate of the location of the predicted value in the V/A plane. This section presents the results previously obtained by the authors developing an emotional prediction system and the limitations that were revealed [24].

In that work, a dimensional model was designed to predict emotions for MER by implementing an emotion prediction system based on a multilayer perceptron (MLP) [24]. The main functional phases that compose the emotional prediction system are presented in Fig. 2.

The data analysis module is responsible for loading and pre-processing the data of the dataset and applying principal component analysis (PCA) to reduce the dimension of the features space of the MediaEval dataset. The classifier

[1] http://opensmile.sourceforge.net/.

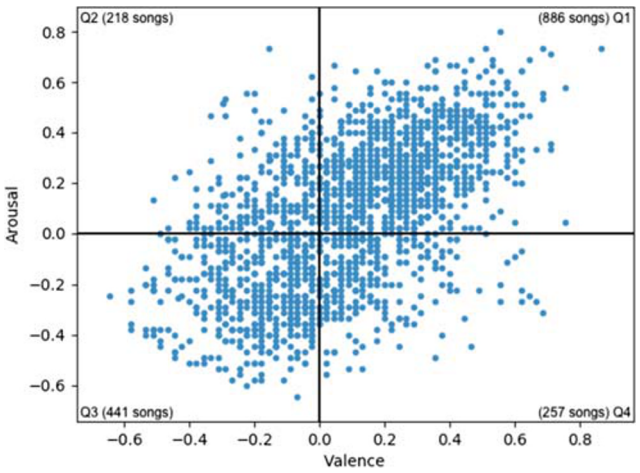

Fig. 1 MediaEval songs distributed in the V/A dimensional space

system design module is responsible for defining the model, the training process, and the metrics to analyze the success rates. In this case, the definition of the model of an MLP [33] includes the number of layers, the activation function used by neurons, and the learning rate for the training process. Depending on the results obtained, the predictive model is configured with new adjustments and the training process is repeated until improvement is achieved.

The prediction system has been implemented using an MLP. The structure of the neural network is presented in Fig. 3. In this figure, it is possible to observe the input neurons, which represent the sound features of the songs; the MLP requires 260 input neurons for the 260 sound features available in the MediaEval dataset. However, it is important to highlight that the implementation of PCA could reduce the number of features; consequently, a lower number of input neurons would be needed. Finally, the MLP has only one neuron output to show the valence or arousal value. For this reason, at least two neural networks are necessary to implement the emotion prediction system.

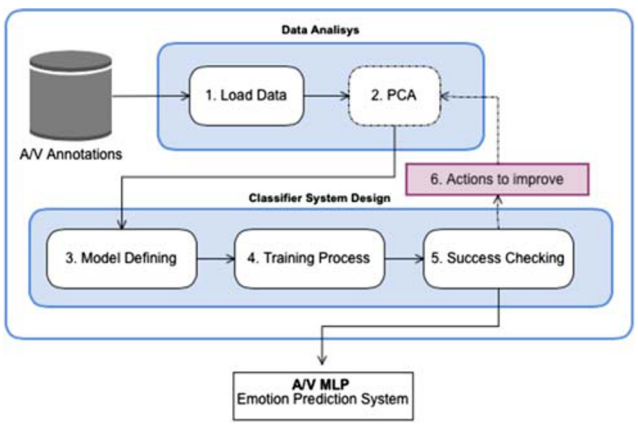

Fig. 2 Main phases of the emotion prediction process [24]

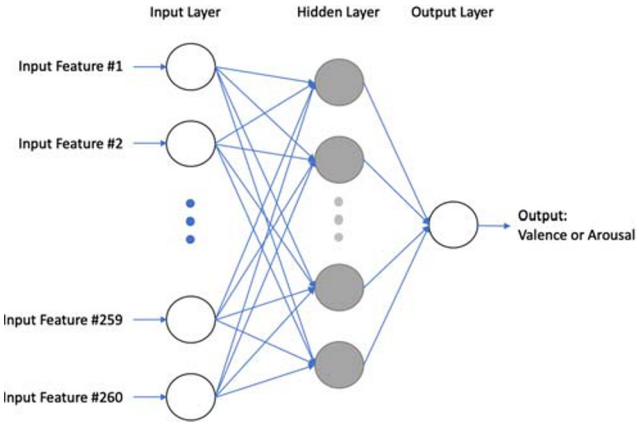

Fig. 3 Multilayer perceptron

The models presented were trained in a conventional way using 80% of the dataset for the training process and 20% for the testing process. The training cycles were determined through the stabilization of the loss metric; in general, the training process stopped when the metric improved very slightly. This constant review over the loss metric was very important to avoid overfitting. With respect to PCA, it was setting with 95% for variance retention. The optimal parameter settings of these models are presented in Table 3, specifying learning rates (LR), hidden layers (HL), and neurons in the hidden layer.

The usual error metric to evaluate success rate with respect to the approximate value predictions is the root mean square error (RMSE). This metric uses a numerical scale between 0 and 1, closer values to 1 indicate higher level error, meanwhile closer values to 0 indicate lower level error. In our case, after applying independent models for valence and arousal (V/A), it was possible to obtain RMSE values of 0.23 and 0.24 respectively (see Table 3).

Additionally, the dataset was divided into four subsets of data, constituting four quadrants to later train independent prediction models for both dimensions (V and A). It was possible to reach RMSE values between 0.11 and 0.16. The parameter settings of these models are presented in Table 4.

These numerical results seemed promising, so that, the predicted V/A values were considered for doing a test of emotional classification by quadrant. As an extension of the prediction system, a very basic conditional rule engine was implemented. Depending on the coordinate obtained

Table 3 Model parameters settings for the best results. LR: learning rate. HL: number of hidden layers. Neurons: number of neurons in the hidden layers. RMSE: root mean squared error

Model	LR	HL	Neurons	RMSE
Valence	0.070	1	64	0.23
Arousal	0.060	1	64	0.24

Table 4 Model parameters settings for best results by quadrants. LR: learning rate. HL: number of hidden layers. Neurons: number of neurons in the hidden layers. RMSE: root mean squared error

Model	Quadrant	LR	HL	Neurons	RMSE
Valence	Q1	0.070	1	128	0.14
	Q2	0.030	1	64	0.11
	Q3	0.070	1	64	0.14
	Q4	0.050	1	128	0.14
Arousal	Q1	0.070	1	64	0.16
	Q2	0.060	1	128	0.14
	Q3	0.060	1	64	0.14
	Q4	0.040	1	128	0.11

by the models presented above in Table 4, the rule system determined the quadrant. The results are shown in Fig. 4, in which it can be observed that the predicted values (orange dots) are not suitable gathered around the quadrants. In particular, the classification accuracy rate obtained by quadrant is quadrant 1 (Q1) = 75%, quadrant 2 (Q2) = 5%, quadrant 3 (Q3) = 35%, and quadrant 4 (Q4) = 21%.

As a conclusion, although the RMSE value obtained by the emotional prediction system seems to be correct, it was not possible to extend the functionality of this system to achieve success in emotional classifications in four or more quadrants through logical rules. The extension of a prediction system to a classification system by obtaining a prediction of the arousal and valence values is not enough to make a successful classification. Therefore, this work is extended in order to design and implement an emotional classification system, that will be explained in detail in the following section.

5 Categorical emotion prediction

In order to improve the emotional classification results a new system is designed, which is trained with the same MediaEval dataset. The main methodological tasks of the classification system are shown in Fig. 5. Four different processes are performed: labeling, data selection and data balance (that conforms data pre-processing phase) and, finally, the classification process.

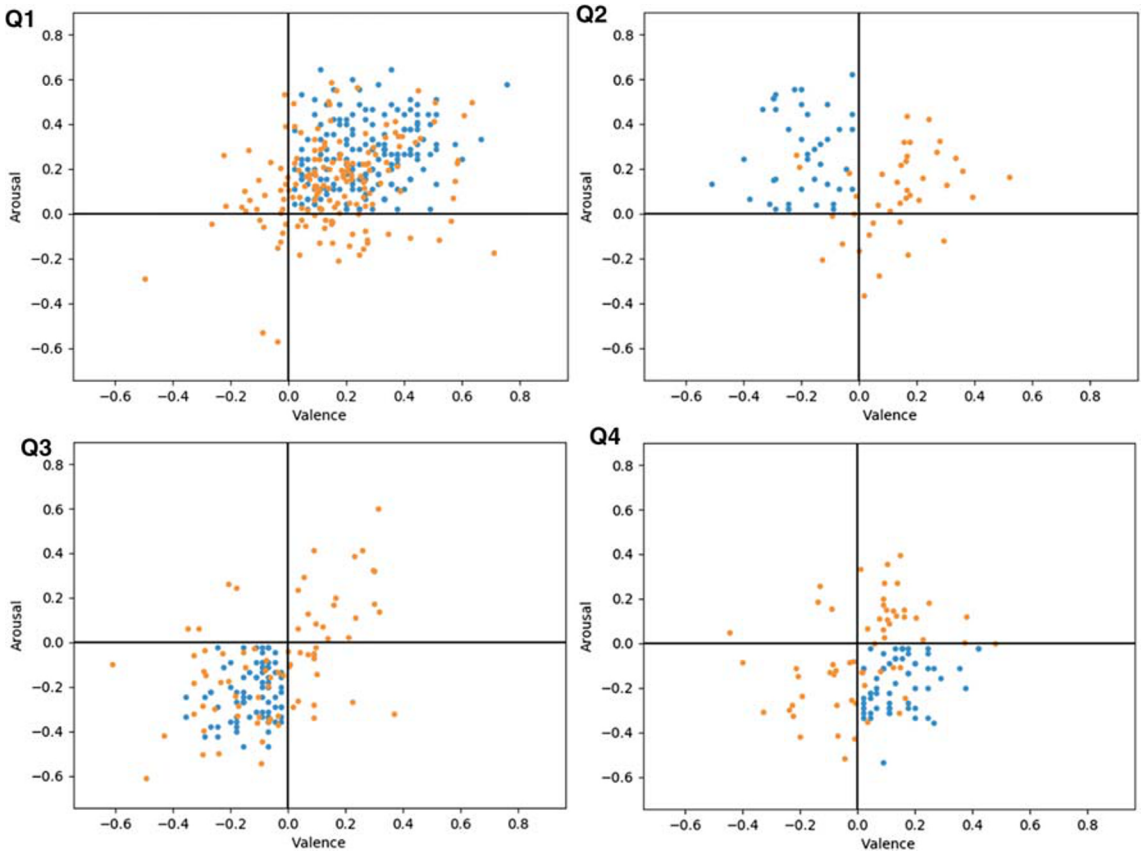

Fig. 4 Classification based on prediction. Q1 top-left, Q2 top-right, Q3 down-left, and Q4 down-right. The quadrant locations of the predicted values is far form being adequate for an emotion classification system (especially Q2 and Q4). Real values in blue dots, predicted values in orange dots

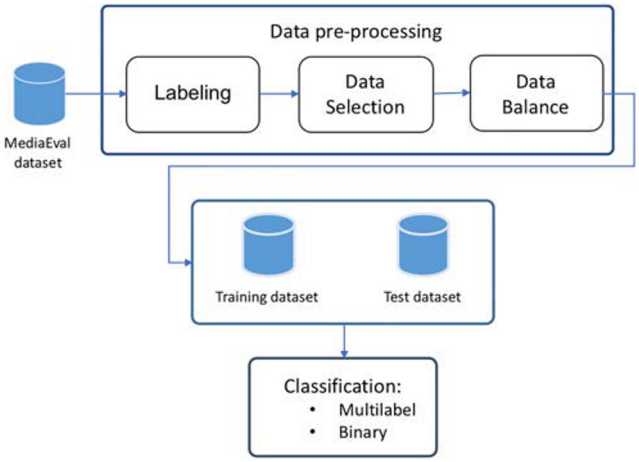

Fig. 5 Block diagram of the classifier tasks

This system has been developed in the Python language, and the main libraries used for each phase are referenced in Table 5.

5.1 Data pre-processing

In this initial phase, some transformations on the original MediaEval dataset are performed. The main objective is to obtain a properly divided training dataset and test dataset. Due to the known unbalanced nature of the dataset, some balancing strategies have been implemented and evaluated.

- Labeling. Each song is labeled with the appropriate quadrant or binary class to which it belongs.
- Data selection. The proportion of data selected for training, validating, and testing is defined. By default, 80% for training and validation and 20% for testing. The validation dataset is established through the early stop parameter, and typically is 20% of the training dataset. The division of data between training and testing is done with a stratified approach, allowing datasets to be separated while maintaining class proportions [16].
- Data balance. Different strategies have been implemented to evaluate the class distribution in the training dataset. It is important to highlight that unbalanced datasets have a negative impact on the classification metrics of minority classes. [5, 23]. For data balancing, three strategies are tested and evaluated: over-sampling, under-sampling, and a combination of over-sampling and under-sampling [21]. Each of the above strategies has different algorithms that are also evaluated according to the results obtained with the classification system in each of their experiments, as will be presented in Section 6.

Table 5 Implemented libraries

Phase	Libraries
Data pre-processing	model_selection.StratifiedShuffleSplit
	imblearn.over_sampling
	imblearn.under_sampling
	imblearn.combine
Classification	klearn.neural_network
	sklearn.ensemble.RandomForestClassifier
	sklearn.svm

5.2 Classification

Three classifiers are implemented and evaluated: linear SVM [6], random forest [25], and a MLP [20]. It is important to note that for all the experiments the training dataset is completely separated from the test dataset in a stratified and random way. This a very important point to guarantee the generalizing capacity of the classifier. In a first step, the three proposed classifiers are evaluated with the dataset annotated with 4 quadrants. Then, the MLP classifier with a binary classification strategy is considered. Finally, the influence of temporal averaging with different analysis windows on dynamic emotional classification is analyzed. The values of the main settings for each of the implemented classifier are presented in Table 6.

For the linear SVM and random forest case there are not many parameterization possibilities compared to MLP. However, one of the most relevant parameters for this work, which is also available in both classifiers, is the possibility of assigning weights to the classes. As mentioned above, the MediaEval dataset is unbalanced, and for this reason, the balancing techniques available for each classifier are of great importance in this study and must be activated.

Table 6 Settings values for each classifier

Classification technique	Parameter	Value
SVM	class_weight	Balanced
Random forest	n_estimators	100
	max_depth	2
	random_state	0
	class_weight	Balanced
MLP	hidden_layer_sizes	(160)
	max_iter	500
	verbose	True
	activation	relu
	early_stopping	True
	validation_fraction	0.2
	tolerance	0.0001
	n_iter_no_change	20

Table 7 Multi-class classifier with balancing strategies for the MLP classifier

Balancing S.	Algorithm	**F**-measure	Avg **F**
None	None	[0.74, 0.19, 0.60, 0.07]	0.40
OverSampler	None	[0.74, 0.19, 0.60, 0.07]	0.40
	Resample	[0.73, 0.35, 0.62, 0.25]	0.49
	SMOTE	[0.73 ,0.34, 0.69, 0.21]	0.49
	ADASYN	[0.71, 0.33, 0.60, 0.30]	0.48
	BorderlineSMOTE	[0.73, 0.29, 0.59, 0.29]	0.47
UnderSampler	RandomUnderSampler	[0.62, 0.25, 0.58, 0.32]	0.44
	ClusterCentroids	[0.51, 0.29, 0.53, 0.25]	0.39
Combination	SMOTEEN	[0.64, 0.36, 0.57, 0.32]	0.47
	SMOTETomek	[0.72, 0.31, 0.62, 0.34]	**0.50**

Best result is in bold

With respect to MLP, the most relevant parameters to avoid overfitting and to improve the capacity of generalization are *early_stopping*, *tolerance*, and *n_iter_no_change*.

The settings values of each classifier are presented in Table 6. Some of the considerations presented in [18] were applied for the definition of the optimal neural network architecture, specifically with respect to the determination of the number of hidden layers and their respective number of neurons.

5.2.1 One vs. rest scheme for improving the classification process

The success rates of two possible approaches in the design of classification systems are also analyzed: binary classifiers and multi-class classifiers. Although the use of binary classifiers may require more computational resources (various classifiers are executed in parallel), it generally allows higher success rates in the classification process [4, 9]. In addition, the use of binary classifiers is also considered for reducing the impact of classification complexity due to unbalanced data [40]. With this precedent, in addition to designing a multi-class classifier, the results obtained with binary classifiers by quadrant and binary classifiers for valence and arousal are also analyzed.

5.2.2 Emotional classification over time

In Section 3, it was mentioned that the MediaEval dataset was dynamically annotated. Even though each song has a continuous 45-s annotation, it is important to discuss the proper window length in which the emotional annotation is the most accurate. For this reason, the classifying system is trained with time windows between 0.5 and 10 s, with 0.5 s of increment. A new set of data is generated for each of the time windows of different duration. The new data are obtained by averaging in time the corresponding values of both the acoustic features and the emotional annotations. Finally, for each case, the F-measure is obtained and its relation to the window length is determined. For this dynamic classification process, the stratification algorithm needs to be customized to guarantee a correctly separation between training and test data. First, a stratification is performed from the averaged database, and then all the time windows of each song are completely separated and added to the corresponding dataset (training or test).

Table 8 Comparison of multi-class classifier implementing SVM, Random Forest, and MLP

Classifier	Balancing strategy	**F**-measure	Avg **F**
SVM	None	[0.75, 0.12, 0.61, 0.17]	0.41
	class_weight	[0.69, 0.30, 0.62, 0.24]	0.46
Randon forest	None	[0.77, 0.04, 0.60, 0.21]	0.40
	class_weight	[0.78, 0.19, 0.64, 0.17]	0.44
MLP	None	[0.74, 0.19, 0.60, 0.07]	0.40
	SMOTETomek	[0.72, 0.31, 0.62, 0.34]	**0.50**

Best result is in bold

Table 9 Binary classifiers by quadrant

Quadrant	Balancing strategy	Algorithm	**F**-measure	Avg **F**
I	None	None	[0.72, 0.76]	**0.74**
II	None	None	[0.93, 0.04]	0.49
	OverSampler	Resample	[0.91, 0.32]	0.62
		SMOTE	[0.91, 0.37]	0.64
		ADASYN	[0.89, 0.30]	0.60
		BorderlineSMOTE	[0.91, 0.39]	**0.65**
	UnderSampler	RandomUnderSampler	[0.72, 0.24]	0.48
		ClusterCentroids	[0.54, 0.24]	0.39
	Combination	SMOTEEN	[0.83, 0.23]	0.53
		SMOTETomek	[0.89, 0.33]	0.61
III	None	None	[0.89, 0.60]	0.75
	OverSampler	Resample	[0.88, 0.64]	**0.76**
		SMOTE	[0.87, 0.63]	0.75
		ADASYN	[0.86, 0.61]	0.74
		BorderlineSMOTE	[0.88, 0.64]	**0.76**
	UnderSampler	RandomUnderSampler	[0.83, 0.63]	0.73
		ClusterCentroids	[0.75, 0.56]	0.66
	Combination	SMOTEEN	[0.83, 0.65]	0.74
		SMOTETomek	[0.86, 0.61]	0.74
IV	None	None	[0.92, 0.04]	0.48
	OverSampler	Resample	[0.89, 0.22]	0.56
		SMOTE	[0.89, 0.28]	**0.59**
		ADASYN	[0.89, 0.18]	0.54
		BorderlineSMOTE	[0.89, 0.28]	**0.59**
	UnderSampler	RandomUnderSampler	[0.75, 0.30]	0.53
		ClusterCentroids	[0.48, 0.27]	0.38
	Combination	SMOTEEN	[0.81, 0.28]	0.55
		SMOTETomek	[0.88, 0.23]	0.56

Best result are in bold

6 Results and discussion

This section presents the results obtained from the experiments described previously. These results are, in turn, divided into two sections: classification system with average values per song (Section 6.1) and classification system analyzed by window length (Section 6.2).

6.1 Classifying system with averaged values per song

This section presents the classification results obtained by working with the MediaEval dataset of 1802 songs and 260 low-level features. The values of emotional annotation and sound features were averaged over the duration of the song obtaining a set of single values per song.

Table 10 Binary classifiers for valence and arousal

Classifier	*F*-measure
(V-,V+)	(0.69, 0.77)
(A-,A+)	(0.66, 0.72)

This work has considered the implementation of three different classifiers. As indicated in Section 5.2, SVM and Random Forest have the possibility to adjust class weights to treat the problem of class unbalancing. In the case of MLP, this parameter is not available, so the imbalanced-learn[2] library has been used to deal with data unbalancing. This library has different strategies, which have been evaluated in order to identify the best alternative for the MLP classifier. The results obtained are shown in Table 7. These results show that, without applying data balancing, the classes less represented in the dataset are not properly classified. Applying some data balancing strategy, the classification performance improves substantially, although it remains low.

[2] https://imbalanced-learn.readthedocs.io/en/stable/.

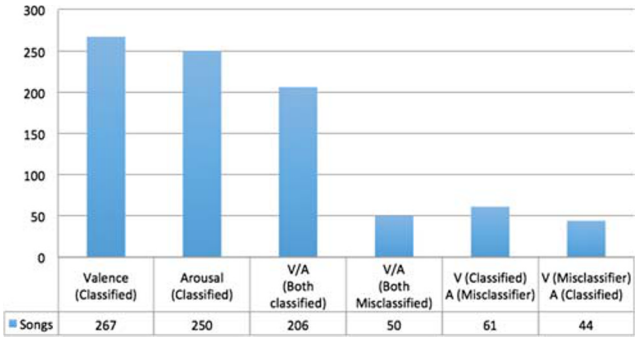

Fig. 6 Success rate in binary classifier V/A

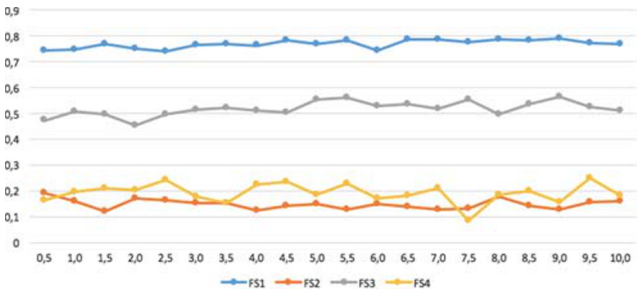

Fig. 8 Performance for different window lengths with stratification by song. Horizontal axis is the averaging window length in seconds and the vertical axis is the F-measure for each class

Once the best balancing strategy has been identified for the MLP classifier, the three different proposed classifiers have been compared. The F-measures for each quadrant are calculated, and the classification performance of the classifiers is summarized in Table 8. The experiment in which the best-averaged F-measures are obtained for all the classes is the one that implements MLP with SMOTETomek [3]. However, it can be observed that, for classes Q1 and Q3, other classifiers could classify with better results (Random Forest with the class_weight function), but with the less represented classes in the dataset very low F-measure values are obtained.

To improve the previous results, a set of binary classifiers has been implemented. Table 9 presents four binary classifiers applying MLP as the classification algorithm, as well as all data balancing strategies. These results are obtained through the one vs. rest approach, in which independent classifiers are designed to identify each class (quadrant). The best F-measures for each quadrant are indicated in italics. Because the majority of the data belongs to class I, it is not convenient to apply balancing strategies, and there are no experiments with balancing strategies for that class. In this case, the best common balancing strategy for all the classes is Oversampling with BorderlineSMOTE, instead of the combination SMOTETomek balancing strategy of the previous four-quadrant classifier.

Finally, Table 10 presents binary classifiers for valence and arousal and their respective F-measures obtained for the same test dataset. These results are quite good, due to the fact that the data of the dataset are distributed more evenly between half-planes of the V/A plane.

In Fig. 6, from the same test dataset of 301 songs, it can be observed that 267 songs (74%) are correctly classified in valence, 250 songs (69%) are correctly classified in arousal, 206 songs (57%) are correctly classified in both dimensions, 50 songs (14%) are not classified in either of them, 61 songs (17%) are correctly classified in valence but they are not correctly classified in arousal, and 44 songs (12%) are not correctly classified in valence but they are correctly classified in arousal.

6.2 Classifying system analyzed by the window length

The final objective of this research is to analyze the success rates in the classification process through the dynamic emotional annotation process available in the MediaEval dataset. As presented in Section 5.2.2, an averaged value of low-level features and emotional annotations have been obtained for different window lengths. In Figs. 7, 8, and 9, the horizontal axis represents the time span of the averaging window in seconds. So, these figures represent

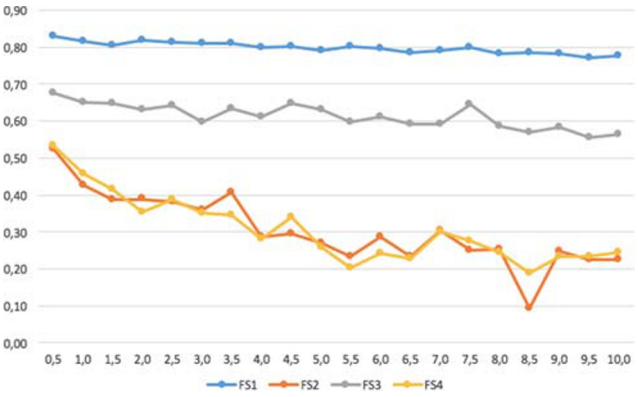

Fig. 7 Performance for different window lengths without stratification by song. Horizontal axis is the averaging window length in seconds and the vertical axis is the F-measure for each class

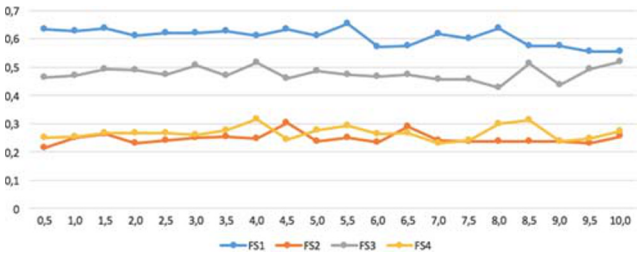

Fig. 9 Performance for different window lengths with stratification by song and SMOTETomek. Horizontal axis is the averaging window length in seconds and the vertical axis is the F-measure for each class

the classification behavior, in terms of *F*-measure, when more or less amount of temporal data are taken into account in a dynamic emotional classification.

Although in Fig. 7 high *F*-measure values can be observed for minority classes with short window lengths, it is important to highlight that for this experiment the stratification process did not guarantee complete separation of the songs. The dataset is chosen from randomly separated time windows without taking into account the precaution of completely separating the songs in the training and test datasets. This generates very good results, but they are due to an overfitting that questions the ability of generalization of the classifier.

In Figs. 8 and 9, a stratification process is carried out before averaging. This ensures proper separation of all time windows of each song in the training and test dataset. Figure 8 shows the different values obtained of *F*-measure according to the variation of the window length. Additionally, Fig. 9 shows the same information but considering the data balancing through the SMOTETomek technique. These results show that in general, the variation of the window length does not significantly improve the success rates in the classification process. On the other hand, it can be seen that data balancing slightly improves the *F*-measures values of minority classes as presented in Table 7 for a multi-class classifier with averaged values per song.

7 Conclusions and future work

This work has focused on the development of a music emotional classifier using the MediaEval dataset. It is based on a previously designed dimensional prediction system using neural networks, in particular, a multilayer perceptron (MLP). The prediction results shows that in spite of being possible to reach low RMSE values (0.11) in the prediction of valence and arousal, they are not valid enough to properly identify an emotion on the V/A plane. The accuracy rates obtained in a four-quadrant classification gives an averaged value of 34%. These results lead to the development of a categorical classifier based on the same dataset.

Three different classifiers have been designed and evaluated: SVM, Random Forest, and MLP. The pre-processing stage of the dataset has been analyzed and it is a key element in the development of this work. A stratification strategy has been implemented by randomly and proportionally dividing the dataset into the training data and the test data. Due to a clearly unbalanced dataset, different balancing strategies has been evaluated, especially in the case of the MLP classifier. It is shown that the success rates between the three analyzed classifiers are not very different, giving a slightly higher *F*-measure to the MLP classifier. But, even using balancing strategies, the less represented classes, that correspond with quadrants Q2 and Q4, get a poor result in terms of F-measure (31% and 34%, respectively, in the best scenario).

In order to improve the results obtained, a set of binary classifiers have been implemented and evaluated using the MLP classifier, which has provided the best results. The classification proposed is a one vs. rest approach giving the best-averaged *F*-measure for the four quadrants of 69%. Therefore, a set of four binary MLP classifiers, trained with the MediaEval dataset, and balanced with the BorderlineSMOTE algorithm could give us a good enough emotional classification of four quadrants. A simpler MLP binary classifier in terms of valence and arousal gives similar results to properly identify the positive and negative regions of the V/A plane (valence is correctly classified at 73%, and arousal at 69%). Finally, it has been shown that using an adequate stratification and balancing scheme, there are very few differences in the classification results using the dynamic annotation provided by the MediaEval dataset. The *F*-measures are almost constant regardless of the analysis time window used to carry out the classification. However, a deeper analysis considering variations and differences in time window length will be done. It is possible to consider that the labeling process was affected by the listeners' mood or something else, which could negatively impact the quality of the emotional annotations in the music pieces.

It is important to note that the design of appropriate emotionally annotated musical databases is one of the most important challenges facing the development of emotional classifiers based on music. Despite the effort made for years, the current databases available to the scientific community either have small amount of data or have labeling problems or are not adequately balanced between the different emotions. From our point of view, the development of this type of database is essential for the progress of research in the music emotion retrieval field. Beyond the possible advances in the development of a dataset, in the near future, our focus will be set in including the results obtained with our classification system in the development of a personalized emotional music recommender system.

Funding information This work has been partly financed by the Spanish Science, Innovation and University Ministry (MCIU), the National Research Agency (AEI), and the EU (FEDER) through the contract RTI2018-096986-B-C31 and the contract TIN2015-72241-EXP, and by the Aragonese Government and the European Union through the FEDER 2014-2020 "construyendo Europa desde Aragón" action (Group T25_17D)

References

1. Aljanaki A, Yang YH, Soleymani M (2017) Developing a benchmark for emotional analysis of music. PLoS ONE 12(3):1–22

2. Bai J, Peng, Shi J, Tang D, Wu Y, Li J, Luo K (2016) Dimensional music emotion recognition by valence-arousal regression. In: 2016 IEEE 15th international conference on cognitive informatics & cognitive computing (ICCI*CC). IEEE, Palo Alto, pp 42–49
3. Batista GE, Bazzan AL, Monard MC (2003) Balancing training data for automated annotation of keywords: a case study. In: WOB, pp 10–18
4. Berstad TJD, Riegler MA, Espeland H, De Lange T, Smedsrud PH, Pogorelov K, Stensland HK, Halvorsen P (2019) Tradeoffs using binary and multiclass neural network classification for medical multidisease detection. In: Proceedings - 2018 IEEE international symposium on multimedia. ISM 2018, pp 1–8
5. Bi J, Zhang C (2018) An empirical comparison on state-of-the-art multi-class imbalance learning algorithms and a new diversified ensemble learning scheme. Knowl-Based Syst 158(June):81–93
6. Chang CC, Lin CJ (2011) LIBSVM. ACM Trans Intell Syst Tech 2(3):1–27
7. Delbouys R, Hennequin R, Piccoli F, Royo-Letelier J, Moussallam M (2018) Music mood detection based on audio and lyrics with deep neural net. In: Proceedings of the 19th international society for music information retrieval Conference, Paris, pp 370–375
8. Deng JJ, Leung CH, Milani A, Chen L (2015) Emotional states associated with music: classification, prediction of changes, and consideration in recommendation. ACM Trans Interactive Intell Sys(TiiS) 5(1):4
9. Diaz-Vico D, Figueiras-Vidal AR, Dorronsoro JR (2018) Deep MLPs for imbalanced classification. In: Proceedings of the International Joint Conference on Neural Networks 2018-July
10. Eerola T, Vuoskoski JK (2011) A comparison of the discrete and dimensional models of emotion in music. Psychol Music 39(1):18–49
11. Fernandes J (2010) Automatic playlist generation via music mood analysis. Msc. thesis, University of Coimbra
12. Garcia-Gathright J, St. Thomas B, Hosey C, Nazari Z, Diaz F (2018) Understanding and evaluating user satisfaction with music discovery. In: The 41st international ACM SIGIR conference on research & development in information retrieval - SIGIR '18. ACM Press, New York, pp 55–64
13. Grekow J (2015) Audio features dedicated to the detection of four basic emotions pp 583–591. Springer International Publishing, Cham
14. Grekow J (2017) Audio features dedicated to the detection of arousal and valence in music recordings. In: 2017 IEEE International conference on innovations in intelligent systems and applications (INISTA). IEEE, Gdynia, pp 40–44
15. Huang S, Zhou L, Liu Z, Ni S, He J (2018) Empirical research on a fuzzy model of music emotion classification based on pleasure-arousal model. In: 2018 37th Chinese Control Conference (CCC) 2018-july, IEEE, pp 3239–3244
16. Jing L, Tian K, Huang JZ (2015) Stratified feature sampling method for ensemble clustering of high dimensional data. Pattern Recogn 48(11):3688–3702
17. Kamasak ME (2018) Emotion based music recommendation system using wearable physiological sensors. IEEE Trans Consum Electron, pp 196–203
18. Karsoliya S (2012) Approximating number of hidden layer neurons in multiple hidden layer BPNN architecture. Int J Eng Trends Tech 3(6):714–717
19. Kim YE, Schmidt EM, Migneco R, Morton BG, Richardson P, Scott J, Speck JA, Turnbull D (2010) Music emotion recognition: a state of the art review. In: Information retrieval, ismir, pp 255–266
20. Kingma DP, Ba J, Adam A (2014) Method for stochastic optimization. In: 3rd International Conference on Learning Representations. ICLR 2015 - Conference Track Proceedings, pp 1–15
21. Kotsiantis S, Kanellopoulos D, Pintelas P (2006) Handling imbalanced datasets : a review. Science 30(1):25–36
22. Li T, Ogihara M, Li Q (2003) A comparative study on content-based music genre classification. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, ACM, pp 282–289
23. Longadge R, Dongre S (2013) Class imbalance problem in data mining review. European Journal of Internal Medicine 24(1):e256
24. Medina Ospitia Y, Beltrán JR, Sanz C, Baldassarri S (2019) dimensional emotion prediction through low-level musical features. In: ACM audio mostly (AM'19), p. 4. Nottingha
25. Ng A, Soo K (2018) Random forests. In: Data science – was ist das eigentlich?!. Springer, Berlin, pp 117–127
26. Ospitia Medina Y, Baldassarri S, Beltrán JR (2019) High-level libraries for emotion recognition in music: a review. In: Agredo V, Ruiz P (eds) Human-computer interaction. HCI-COLLAB 2018. Springer, Popayán, pp 158–168
27. Panda R, Malheiro R, Paiva RP (2018) Musical texture and expressivity features for music emotion recognition, In: Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018, pp. 383–391. http://ismir2018.ircam.fr/doc/pdfs/250_Paper.pdf
28. Panda R, Paiva RP (2012) Music emotion classification: dataset acquisition and comparative analysis. In: 5th Int Conference on digital audio effects (DAFx-12), york, september 17-21, 2012, York
29. Panda R, Rocha B, Paiva RP (2013) Dimensional music emotion recognition: combining standard and melodic audio features. In: Proc 10th International Symposium on Computer Music Multidisciplinary Research, pp 1–11
30. Picard RW (1997) Affective computing. MIT press, Cambridge
31. Powell J (2012) Así es la música, titivilus edn, Titivilus Barcelona
32. Rajanna AR, Aryafar K, Shokoufandeh A, Ptucha R (2015) Deep neural networks: a case study for music genre classification. In: 2015 IEEE 14th international conference on machine learning and applications (ICMLA), IEEE, pp 655–660
33. Rumelhart DE, Hinton GE, Williams RJ (1988) Learning representations by back-propagating errors. cognitive modeling. Nature 1986b:533–536
34. Russell JA (1980) A circumplex model of affect. J Pers Soc Psychol 39(6):1161–1178
35. Schedl M, Zamani H, Chen CW, Deldjoo Y, Elahi M (2018) Current challenges and visions in music recommender systems research. Inter J Multimedia Information Retrieval 7(2):95–116
36. Scherer KR (2004) 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
37. Schmidt EM, Turnbull D, Kim YE (2010) Feature selection for content-based, time-varying musical emotion regression. In: Proceedings of the international conference on Multimedia information retrieval - MIR '10. ACM Press, New York, p 267
38. Sloboda JA (1986) The musical mind, Oxford psy edn. Oxford University Press, New York
39. Soleymani M, Aljanaki A, Yang YH (2016) DEAM: MediaEval database for emotional analysis in Music. pp 3–5
40. Vluymans S, Fernández A, Saeys Y, Cornelis C, Herrera F (2018) Dynamic affinity-based classification of multi-class imbalanced data with one-versus-one decomposition: a fuzzy rough set approach. Knowl Inf Syst 56(1):55–84
41. Yang D, Lee WS (2009) Music emotion identification from lyrics. In: 2009 11th IEEE international symposium on multimedia, IEEE, pp 624–629
42. Yang X, Dong Y, Li J (2018) Review of data features-based music emotion recognition methods. Multimedia Systems 24(4):365–389

43. Yang YH, Chen HH (2012) Machine recognition of music emotion : a review. ACM Trans Intell Syst Tech 3(3):30
44. Zhang F, Meng H, Li M (2016) Emotion extraction and recognition from music. In: 2016 12th international conference on natural computation, fuzzy systems and knowledge discovery, ICNC-FSKD 2016, pp 1728–1733
45. Zhou N (2019) Database design of regional music characteristic culture resources based on improved neural network in data mining. Pers Ubiquit Comput, pp. 1–12