MIR TASK: EMOTION DETECTION IN MUSIC: A COMPARISON OF SVM, DEEP NET, AND ...

Chung Kang Chen
NYU
ckc432@nyu.edu

Daniel Kachler
NYU
dk4075@nyu.edu

Dhruv Vajoeyi
NYU
dv2086@nyu.edu

ABSTRACT

Tasks in music information retrieval (MIR) can help extract features that provide data useful for mood based music classification. Features can be used for emotion detection which can help link music consumption to its user's emotions, and to classify music per the composition's mood or emotion using Russell's circumplex model of affect for criterion (Russell, J. A. 1980). A useful way to extract these features is through machine learning (ML) models. Commonly applied ML models are based on gaussian processes. A framework is proposed and compared by implementing support vector machine (SVM), K-nearest neighbour and Neural Networks for accuracy utilizing a feature set.

1. INTRODUCTION

Music information retrieval (MIR) can provide valuable information that helps quantify audio data. Music emotion Recognition (MER) is an area of interest to MIR research because of the relationship between music and emotions. Previous research has used different MIR features in order to classify and label music per the composition's perceived emotions as related to the eight most common musical dimensions according to musicology literature (melody, harmony, rhythm, dynamics, timbre, expressivity, texture, and form). The musical dimensions as related to their MIR features were used in previous works as the input datasets of Gaussian based Machine learning (ML) in order to classify the datasets in different moods according to the Russell's circumplex model of affect, a representation of the cognitive structure that people utilize in conceptualizing affect (Russell, 1980). The features utilized in previous work range from low-level features such as MFCC, or spectral; to high-level features such as genre or danceability.

Previous works and methodology will be discussed, and compared. The proposed framework will implement ML models based on Gaussian processes and will compare the results of support vector machine (SVM), K- nearest neighbor (KNN) and Neural Networks for accuracy utilizing a dataset of the top 100 features extracted from a dataset of music called the 4Q audio emotion dataset (Russell's model) (2018). This dataset consists of the top 100 features regarded as relevant for MER, basic, mid-level and high-level features were extracted. In order to obtain results, Russell's circumplex model was used for criteria of different emotions. Russell's circumplex model is a dimensional model of emotion taxonomy.It is argued that dimensional models give less ambiguous results.

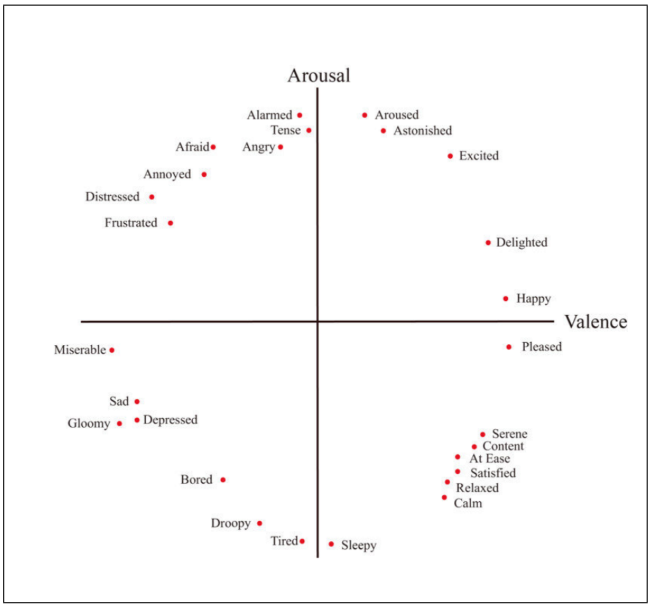

Figure 1. Russell's circumplex model of affect

Russell's circumplex model divides different emotions based on four quadrants of a cartesian plane which has an x-axis for valence and a y-axis for arousal (Fig.1). "The two proposed dimensions are valence (pleasant-unpleasant) and activity or arousal (aroused-not aroused), or AV" (Panda, R., et al 2018). Widely regarded and accepted as a standard in the cognitive sciences for emotion classification. The circumplex was proposed to give psychologists and psychiatrists a roadmap, and to classify emotions as conveyed by non-cognitive professionals in their daily lives. The objective of this work is to compare the results of the different ML models, and critically analyze the models and the results through the proposed framework.Previous works have performed the same task utilizing other machine learning models as opposed to the SVM proposed framework.

This paper is organized as follows. Section 2 reviews the related works, section 3 discusses the utilized dataset,

Attribution: Chung Kang Chen, Daniel Kachler, Dhruv Vajoeyi. "MIR Task: Emotion Detection in Music: a comparison of SVM, Deep Net, and ...", 21st International Society for Music Information Retrieval Conference, Montréal, Canada, 2020.

section 4 goes over the employed methods, section 5 discusses the experiment's results, section 6 is the conclusion and section 7 for future work.

2. RELATED WORKS

2.1 Neural Network

In previous Studies, Continuous Conditional Neural Fields offers improvement over the previous approaches. In more recent studies, a combination of CNN and recurrent neural networks (RNN) has proven to achieve better results with even fewer parameters.

Following these findings, the proposed framework proposes a new way to achieve similar or better results when compared to recent models by training a CNN-RNN structured model directory on raw audio data. The proposed neural network is obtained by stacking a one dimensional convolution layer (1D-CNN) with a time distributed fully connected (TD-FC) layer and a bidirectional Gated Recurrent Unit (BiGRU) and another max-out fully connected (MFC) layer with two output units.

The model was trained with Back-Propagation Through Time (BPTT) and Adam optimizer (Kingma & Ba, 2014). To prevent the system from over-fitting, the mean squared error (MSE) was used in the training phase and the root mean square error (RMSE) was used in the evaluation phase.

Surprisingly, the results were considerably better than CRNN and CRNN-NB architectures with fewer parameters (3.3k) compared to CRNN, which has 30k parameters and CRNN-NB, which has 17k parameters.

From previous research, researchers concluded that raw audios might not be the best choice when it comes to training neural networks (Malik et al., 2017). It, however, worked surprisingly well in the valence dimension. While this does not prove that raw audio is a better choice for training neural networks, this may lead to the conclusion that hand-crafted features don't describe the valence aspect of the audio signal in an effective way.

2.2 Support-Vector Machine

Previous studies have demonstrated the semantic gap between audio features and MER. "As an example, MFCCs belong to tone colour but do not give explicit information about the source or material of the sound. Nonetheless, they can implicitly help to distinguish these. This is an example of the mentioned semantic gap, where high level concepts are not being captured explicitly with the existent low level features" (Panda, R., et al 2018). An example of how low level features cannot capture high level concepts. The study carried out by Panda, R., et al (2018) presents another problem in MER, the datasets being utilized often give ambiguous results of perceived emotion.

These findings led the researchers to propose a framework. The authors built their own dataset that uses simple taxonomy through semi-automatic construction of the aforementioned dataset (reducing resources), obtaining a dataset containing hundreds of songs, constructed for public use. In order to validate the emotion annotations a blind test was carried out in which the subjects were given randomly distributed and selected 30 seconds clips of music and annotated the clips according to Russell's circumplex model of affect. It is important to mention that clips perceived as unclear by the subjects were also marked (labeled as unclear).

1702 features were then extracted utilizing MIR Toolbox, Marsyas and PsySound. Duplicate or similar features were removed; features irrelevant to the study were also removed from the results. After constructing the dataset Support Vector Machines were used for classification, which correlated the dataset to the emotions in the circumplex. According to the authors an SVM network proved to be robust and better at classifying in four quadrants than other models from the authors reviewed literature and their own work.

Figure 2 from Panda, R., et al's 2018 work displays the standard and novel audio features extracted and organized by musical concept. Most are tonal features. Figure 3 of the aforementioned authors work shows the best 30 features to distinguish each quadrant, grouped by categories. Baseline features obtained an f-score of 67.5 percent with SVM and 70 standard features. The same SVM implementation had an f-score of 71.7 percent utilizing a total of 800 features. The same implementation on both novel and standard features gave an f-score of 76.4 percent utilizing a low number of features; 100 (29 novel and 71 baseline features) as opposed to the 800 used in the previously mentioned result.

From their results and previous research the authors concluded that by implementing features related to higher level musical concepts. Utilizing two of the less represented concepts in MER feature extraction, features related to musical expressive performance techniques such as vibrato, tremolo, and glissando and features related to musical texture combined with baseline features provide more accurate results. The optimal results were obtained by utilizing 29 novel and 71 baseline features, and results of other larger feature sets gave lower f-scores.

3. DATA SET

The data set contains 900 audio clips with annotations based on the four emotional classes from Russel's emotion circumplex. Existing audio features include baseline feature and novel features proposed in the "Novel Audio Features for Music Emotion Recognition" (Panda el al., 2018). It was built in a semi-automatic method, using AllMusic annotations along with simpler human validation to reduce the resources required to build a fully manual data set. There are three steps for construction of the data set:

Step 1: Query each of the 289 distinct emotion tags for the top songs in the AllMusic API. 89 percent had an associated audio sample and 98 percent had genre tags, with 28646 distinct artist tags present.

Step 2: Map all the tags into Russell's 4 Valence Arousal quadrants using Warriner's adjectives list (Warriner et al., 2013) that contains 13915 English words with ratings arousal, valence, and dominance. Which resulted

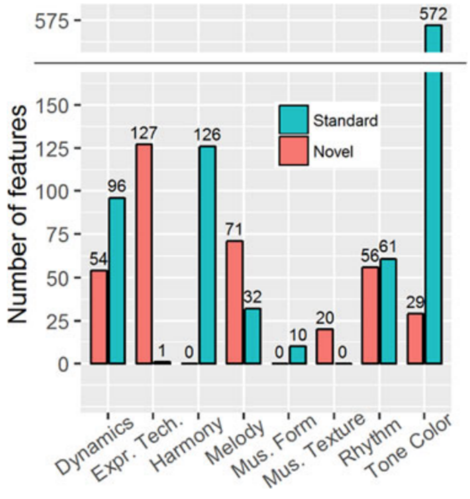

Fig. 2. Feature distribution across musical concepts.

Figure 2. Panda, R., Malheiro, R., Paiva, R. P. (2018)

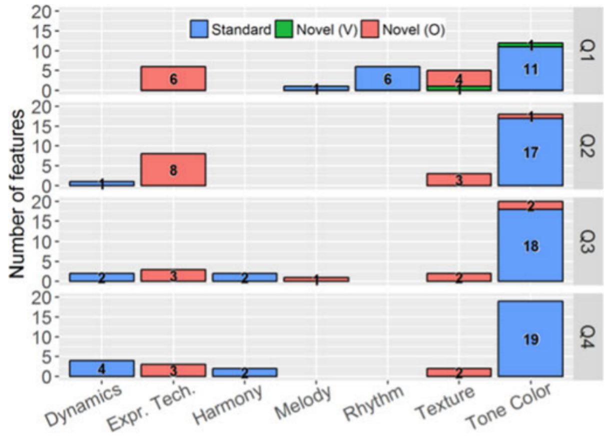

Fig. 3. Best 30 features to discriminate each quadrant, organized by musical concept. Novel (O) are extracted from the original audio signal, while Novel (V) are extracted from the voice-separated signal.

Figure 3. Panda, R., Malheiro, R., Paiva, R. P. (2018)

in 200 common words of intersection between Warriener's list and the AllMusic annotation.

Step 3: The set of metadata, audio clips, and emotion tags with arousal, valence, and dominance values was then processed and filtered. Songs that did not have a dominant quadrant and duplicated were removed from the data set and the data set was re-balanced to have a more even distribution amongst all quadrants. Finally the data set was sub-sampled to consist of 2200 songs that are evenly distributed in each quadrant, which were then manually validated.

Using the MIR Toolbox, Marsyas, and PsySound, 898 standard features were extracted from the audio data to serve as the baseline.

3.1 Novel Features

As we human beings rely on cues and higher-leveled musical concepts to determine the emotions in the music, novel features were computed for the data set using traditional melodic, dynamic, and rhythmic features. The novelty features being proposed are musical texture features and expressivity features.

3.1.1 Musical Texture Features

Musical texture features are features that are related with the music layers of the song. Using a sequence of multiple frequency estimation to measure the number of simultaneous layers in each frame of audio signal, they have extracted following features:

Musical Layers (ML) statistics: Number of layers in a frame is the number of f0 estimates in that frame. 6 usual statistics regarding the distribution of the estimates were then computed: MLmean, MLstd, etc.

Musical Layer Distribution (MLD): Musical layers in each frame were divided into 4 classes: no layers, a single layer, two simultaneous layers, and three or more layers. MLD is the percentage of the frames in each of the classes.

Ratio of Musical Layer Transitions (RMLT): the information of the transition from a specific layer to another layer.

3.1.2 Expressivity Features

As common expressions such as vibrato, tremolo, and articulations are a common part of music which links to emotions. Here are the novelty functions of expressive techniques:

Articulation Features: Classifying all the transitions between notes in the song clip, other features as listed below are extracted.

Staccato Ratio (SR), Legato Ratio (LR), and Other Transition Ratio (OTR): The feature that indicates the ratio of each transition type to the total number of transitions.

Staccato Notes Duration Ratio (SNDR), Legato Note Duration Ratio (LNDR), and Other Transition Notes Duration Ratio (OTNDR): The ratio of the duration of notes with specific articulation to the duration of all the notes.

Glissando Presence (GP): Detection of whether a song piece has glissando in it.

Glissando Extent (GE) statistics: 6 usual statistics are calculated for notes containing glissando.

Glissando Duration (GD) and Glissando Slope (GS) statistics: 6 usual statistics are calculated for glissando duration and glissando slope.

Glissando Coverage (GC): Global coverage of glissandos.

Glissando Direction (GDIR): Global direction of glissandos.

Glissando to Non-Glissando Ratio (GNGR): The ratio of notes containing glissando to the total number of notes.

Vibrato Presence (VP): Detection of whether a song piece has vibrato in it.

Vibrato Rate (VR) statistics: 6 usual statistics were calculated based on the vibrato rate of each note.

Vibrato Extent (VE) and Vibrato Duration (VD) statistics: 6 usual statistics were calculated based on the vibrato extent and vibrato duration of each note.

Vibrato Coverage (VC): Global coverage of vibratos.

High-Frequency Vibrato Coverage (HFVC): Measures vibrato coverage restricted to notes over note C4 (261.6Hz).

Vibrato to Non-Vibrato Ratio (VNVR): The ratio of notes containing vibrato to the total number of notes.

Vibrato Notes Base Frequency (VNBF): Similar to the vibrato rate statistics, 6 usual statistics were calculated for the base frequency of all notes with vibrato.

4. METHOD

We evaluated the performance of different models on the novel features dataset to determine how effectively different Machine Learning models can make use of handcrafted features to extract emotion information.

4.1 Dataset

From the novel features dataset, we used annotations.csv which listed an ID for each of 900 tracks with a label∈{Q1, Q2, Q3, Q4}, and top100_features.csv where each row includes a track id and real numbered values for the selected top 100 features.

The data was then processed by merging both tables into one feature set, and one-hot encoding the labels.

4.2 Process

For each model, we perform a 10-fold cross-validation split on the dataset. The average f1-score and confusion matrix is computed across the 10 trials and is used to evaluate and compare the models.

4.3 Models

4.3.1 Gaussian Kernel SVM

The SVM is implemented in Python using the sklearn package with an instantiation of the sklearn.svm.SVC object using a Radial Basis Function kernel.

4.3.2 K-Nearest Neighbors

The KNN model is implemented in Python using the sklearn package with an instantiation of the sklearn.neighbors.KNeighborsClassifier with a parameter of k=50.

MFCC1(mean)
Average Power Spectrum(median)
Music Layers(mean)
Spectral 2nd Moment(median)
Spectral Skewness(std)
Spectral Skewness(max)
Rolloff(mean)
Music Layers(std)
Rolloff(MeanA/StdM)
Tremolo Notes(mean)
SFM_15(mean)
Average Power Spectrum(mean)
Loudness(skewness)
MFCC1(max)
State Transitions(per sec)
Spectral Entropy(std)
Spectral Centroid(std)
SFM_12(mean)
Vibrato Extent(std)
Tremolo Coverage(C4+)

Table 1. Top 20 Features

4.3.3 Neural Network

We use Tensorflow and Keras to implement and tune the parameters of a Neural Net.

For each Neural Net we are training over 10 epochs using Root Mean Squared Propagation and a Categorical Cross-Entropy loss function.

For all internal activation layers, we use the ReLU function, and a softmax activation for the output.

- Simple Model: A dense input layer of 50 neurons and an output of 4 neurons.
- Multi-Layer Model: A dense input layer of 50 neurons, A hidden layer of 20 neurons and an output of 4 neurons.
- Dropout Model: Attempt to rectify overfitting of training data using a dropout layer with dropout rate=0.5

5. RESULTS

5.1 SVM

Average F1 Score: 0.749

	Q1	Q2	Q3	Q4
Precision	0.716	0.894	0.733	0.670
Recall	0.818	0.822	0.671	0.684
F1	0.763	0.856	0.701	0.677

Table 2. SVM: Results by Quadrant

	Q1	Q2	Q3	Q4
Q1	184	15	5	21
Q2	29	185	6	5
Q3	18	6	151	50
Q4	26	1	44	154

Table 3. SVM: Confusion Matrix

5.2 K-Nearest Neighbors

Average F1 Score: 0.676

	Q1	Q2	Q3	Q4
Precision	0.611	0.911	0.594	0.656
Recall	0.831	0.729	0.618	0.524
F1	0.704	0.810	0.606	0.583

Table 4. kNN: Results by Quadrant

	Q1	Q2	Q3	Q4
Q1	187	10	16	12
Q2	51	164	5	5
Q3	36	5	139	45
Q4	32	1	74	118

Table 5. kNN: Confusion Matrix

5.3 Neural Net

The best result was obtained using the dropout model, an input layer of 50 and a dropout layer of 0.5.
Average F1 Score: 0.719

	Q1	Q2	Q3	Q4
Precision	0.710	0.854	0.717	0.618
Recall	0.796	0.809	0.551	0.72
F1	0.751	0.831	0.623	0.665

Table 6. Neural Net: Results by Quadrant

	Q1	Q2	Q3	Q4
Q1	179	19	6	21
Q2	31	182	5	7
Q3	19	10	124	72
Q4	23	2	38	162

Table 7. Neural Net: Confusion Matrix

5.4 Summary

The results show that the SVM outperforms the Neural Net when handling the novel feature dataset. The F1-Score across all quadrants is better in the case of the SVM, but significantly more so in the case of Q3 (Low Arousal-Low Valence).

In all models, accuracy suffers greatly in Q3 and Q4. The confusion matrices indicate that there Q3 and Q4 are frequently mistaken for eachother.

Q2 is the most accurately predicted class across all models.

6. CONCLUSION

From the above work we conclude that the novel features provide value in estimating Music Emotion, even achieving positive results with a simple k-Nearest Neighbors approach.

The Neural Network was ultimately outperformed by the SVM and this likely comes down to the Neural Network requiring a lot more data to tune it's parameters. Even a shallow neural network has to tune many neurons. During training we noticed the accuracy metrics during training were very high, indicating that the model was overfitting to the training data. This is an issue when there are many features but not much data. Until methods for extrapolating the novel features are released, we are reliant on this database to train our models, and thus the SVM remains most effective. Fitting and predicting on such a small database was also much more efficient through an SVM, with the Neural Net requiring a few minutes to train, but the SVM fitting instantly.

We also reviewed tracks from the different quadrants and noticed that Q3 and Q4 were emotionally ambiguous, so it is no surprise that none of the models were particularly accurate in that domain. It may be of more use to use real number values and attempt regression problems on those values instead of classification. This would address ambiguity since songs in Q3 or Q4 could still be close to the center of the valence scale. This again relies on annotators to provide valence-arousal values, and would need extensive surveying and scientific method applied to come close to an agreed upon value.

7. FUTURE WORK

We would like to explore data augmentation methods which preserve the emotion of the audio. This would tackle the problem of unavailability of emotion annotations leading to small datasets. Augmentations such as pitch shifting and time-stretching both potentially alter the perceived emotion of a track. Potential options for augmentation include adding noise or relative volume adjustment.

We also want to try training a convolutional neural network on the raw audio signal from the novel features dataset. Extracting multiple features makes sense for improving the performance of an SVM. However, a convolutional neural network is able to extrapolate hidden features as parameters for its convolutional and hidden layers. While these features may not be naturally explainable, they may be useful for the task of Music Emotion Recognition, and thus handcrafting novel features may have no benefit for complex machine learning models.

8. REFERENCES

[1] Ba J. Kingma, D. P. Adam: A method for stochastic optimization. volume 1412, 2014.

[2] Adavanne S. Drossos K. Virtanen T. Ticha D. Jarina R. Malik, M. Stacked convolutional and recurrent neural networks for music emotion recognition. *arXiv*, 1706(02292), 2017.

[3] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[4] Jarina R. Chmulik M. Kuba M. Orjesek, R. Dnn based music emotion recognition from raw audio signal. *IEEE)*, pages 1–4, 2019.

[5] Malheiro R. Paiva R. P. Panda, R. Novel audio features for music emotion recognition. *IEEE Transactions on Affective Computing*, 11(4):614–626, 2018.

[6] Malheiro R. M. Paiva R. P. Panda, R. Audio features for music emotion recognition: a survey. *IEEE Transactions on Affective Computing*, 2020.

[7] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(1163), 1980.

[8] Kuperman V. Brysbaert M. Warriner, A. B. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207, 1980.