Emotion Detection in Music

Undergraduate Semester Project

Fall Semester 2015-16

Author:
Anurag Sharma
Shivam Khandelwal

Supervisors:
Dr. Tanaya Guha
Dr. Subhajit Dutta

Abstract

Music has played an important role in development of human culture and society. This is primarily because of its ability to affect our mood and elicit emotions. Understanding its inherent capability to induce feelings and sentiments is helpful in music understanding and developing music retrieval and classifier systems. In developing automated systems to music emotion recognition one of the core issues is that there is no consensus on which factors in the music contribute to the expression of emotions the most making it difficult problem to find a robust MER. In this report we review the state of the art techniques used in building MER and address one of the core issues of selecting useful features for prediction emotional content in a music clip without losing useful information.

Contents

1 Introduction

There has been a exponential growth in the digital music libraries over the past decade. With this growth has come the expansion in music information retrieval (MIR) research and development of automated systems for cataloging and organizing music and related data. MIR has numerous application in building search engines for music and recommender systems for the music industry. It also helps in gaining insight in the relationships among the acoustic, electrical features of sound and auditory perception.

Some categories of music information encoding and retrieval like genre classification, artist or instrument identification have received more attention because the information needed to build the automated systems is easily quantified. Emotional content and perception of sentiments in a musical clip, on the other hand, is highly subjective and difficult to quantify.

Prediction of emotion expression of a music clip involves a review of the auditory perception, psychology of emotion and musical theory. In this report we present the overview of the commonly used quantification methodologies for human perception of emotion. We also explore the relationship between the acoustic features of the music and the quantified values of perception of emotion. We provide a mathematical background for the techniques commonly employed in building music emotion recognition (MER) system and highlight the difficulties encountered in the task of automated emotion prediction.

Remainder of the report proceeds as follows. We describe the framework for emotion recognition, tags and acoustic signal content based approaches in the remaining sections of this chapter. In Chapter 2 we describe the mathematical models used in the report. We describe the data set and experimental setup in Chapter 3 and 4 respectively Finally we discuss results and insights in Chapter 5 of this report and conclude the report with remarks and possible directions for future work.

Mood adjectives used in the MIREX Audio Mood Classification task. [5]

Clusters	Mood Adjectives
Cluster 1	RMS Energy
Cluster 2	passionate, rousing, confident, boisterous, rowdy
Cluster 3	rollicking, cheerful, fun, sweet, amiable, good natured
Cluster 4	literate, poignant, wistful, bittersweet, autumnal, brooding
Cluster 5	humorous, silly, campy, quirky, whimsical, witty, wry
Cluster 6	aggressive, fiery, tense/anxious, intense, volatile, visceral

Table 1.1: The above table is taken from [6].

1.1 Emotion in Music

The sound of music can arouse profound emotions in listeners. Some studies have regarded emotional content in the music as the most important characteristic for its aesthetic value. [1] Some authors also describe music as a ”language of emotions”. [2] This emotion evoked by music while listening to it is what we call emotion in music in our study.

1.2 Representation of Emotion

Typically two main models of representation of emotion in music has been used in the literature. Brief overview of the two models is as follows.

1.2.1 Categorical Description

In categorical description of music emotion, subjects are asked to select words from a predefined list which are closest to the feelings experienced by them on listening the soundtrack. Using these annotations the soundtracks are classified into different moods and emotion categories. One of the earliest study on emotion content of Music was done by Hevner [3] in which he used 66 adjectives for subjects to choose from for a music clip and grouped them in 8 categories. Recently Zenter et al. [4] also did a study to compile a list of music-relevant emotion terms study their frequency.

1.2.2 Dimensional Representation

Some researches suggest that emotion can be quantified as a value on a continuous scale in some predefined dimensions. Most notable of these dimensional models is the two dimensional Valence - Arousal Model established by Russel. [7] In this model any music clip is represented in a Arousal - Valence plane where Arousal broadly refers to energy or excitement content of the clip and Valence caters to the positive or negative nature of the emotion expressed in the music. Accordingly a high arousal and high valence valued song corresponds to a music which elicits ecstatic emotion from the listener. On the other hand a low arousal and high valence song corresponds to a calm or soothing melody. Figure 1.1 represents the generally agreed upon emotions corresponding to A-V values.

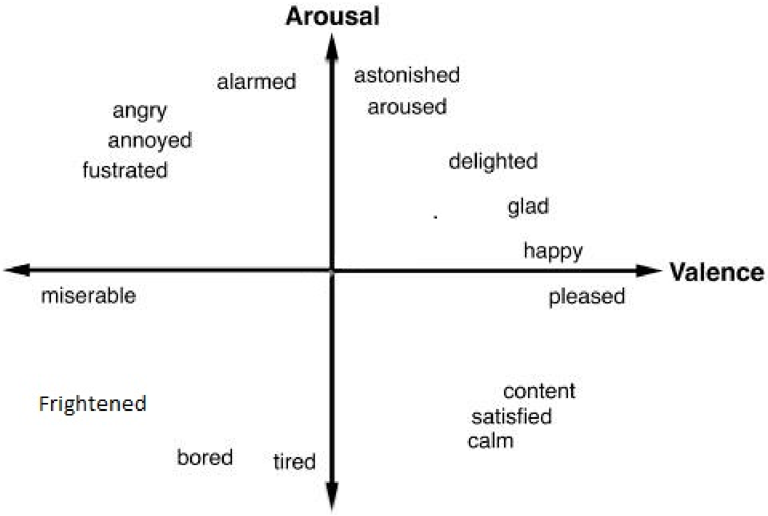


Figure 1.1: The above figure taken from [6] describes the emotions which correspond to different Arousal-Valence values.

1.3 Content Based Audio Analysis

A longstanding goal of music information retrieval research community has been to find a relationship between the signals and features derived from processing the audio file of the music clip and perceived emotions by the listeners. This can help in building automated music tagging systems and can be used on huge online music libraries to tag songs with the emotional content they contain. Emotions perceived on listening to a song are affected by the tone, loudness, pitch, timber, harmony etc. of the soundtrack. Much of the research towards building content based systems has been directed towards building informative acoustic features. Table 1.2 gives an overview of the acoustic features used commonly for different emotions. Apart from the features mentioned in the table, some functionals and statistics like range, min, max etc. of the features are also used as separate features in studies.

We use following low level features in our study. The detailed descriptions of the features can be found in [8].

- **pcm RMSenergy**: Root-mean-square signal frame energy
- **mfcc**: Mel-Frequency cepstral coefficients 1-12
- **pcm_zcr**: Zero-crossing rate of time signal (frame-based)
- **voiceProb**: The voicing probability computed from the ACF
- **F0**: The fundamental frequency computed from the Cepstrum

Common acoustic feature types for emotion classification

Type	Features
Dynamics	RMS Energy
Timbre	MFCCs, spectral shape, spectral contrast
Harmony	Roughness, harmonic change, key clarity, majorness
Register	Chromagram, chroma centroid and deviation
Rhythm	Rhythm strength, regularity, tempo, beat, histograms
Articulation	Even density, attack slope, attack time

Table 1.2: The above table is taken from [6].

These features encompass a wide range of domains which include dynamics, timbre, harmony, register, rhythm, and articulation. [9]. Some other studies adopt a more general approach to feature extraction and employ dimensionality reduction techniques. [10] In the next chapter we describe few techniques employed in our study.

2 Methodologies Used

2.1 Principal Component Analysis

Principal component analysis (PCA) is a statistical technique that analyzes a data table with possibly inter - correlated observations. The main idea of principal component analysis is to reduce the dimensionality of a data set which consists of a large number of correlated variables.

At this time, it is necessary to preserve as much as possible of the variation of original data set. Mathematically, PCA depends upon the eigen-decomposition of positive semi-definite matrices and upon the singular value decomposition (svd) of rectangular matrices.

The number of principal components is less than or equal to the number of original variables.

Suppose the data is given in p variables, $X^T = (X_1,\ X_2,\ \ldots,\ X_p)$. Let Σ be the co-variance matrix of X with eigen values $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. Let $Y_1, Y_2, \ldots, Y_p$ be he principal components. All the principal components are linear combinations of variables, $Y_i = A_i^T X$.

We want to maximize $\text{Var}(Y_i) = A_i^T \Sigma A_i$ subject to $A_i^T A_i$ and $\text{cov}(Y_j, Y_i) = 0\ \forall j \leq i$.

Maximization yields $Y_i = e_i^T X$ with $\text{Var}(Y_i) = \lambda_i$, where e_i is the eigen vector corresponding to λ_i. For further details interested reader is referred to [11] and [12].

2.2 Elastic Net Regression

The elastic net is a regularized regression method that linearly combines the L^1 and L^2 penalties of the *lasso* and *ridge* methods. The Lasso method uses the penalty function based on $\|\beta\|_1 = \sum_{i=1}^{n} \beta_i$, while Ridge regression uses $\|\beta\|^2$ for penalty term.

The use of the above penalty function has many limitation. In particular, for ”large p, small n” case the Lasso selects at most n variables before it saturates. Also, Lasso tends to select one variable (and ignore others) out of the group of highly correlated variables ($corr \geq 0.75$). [13]

To overcome these limitations, Elastic Net uses linear combination of both of these penalty terms.The estimates from Elastic Net method are defined by:

$\hat{\beta} = ArgMin_{\beta}(\|y - X\beta\|^2 + \lambda_1\|\beta\| + \lambda_2\|\beta\|^2)$

The quadratic penalty term makes the loss function strictly convex, and it therefore has a unique minimum.

2.3 Support Vector Regression

Support Vector Machines(SVM) can also be applied to regression problems by the introduction of alternate cost or loss functions.

The model produced by support vector classification depends only on a subset of the training data, because the cost function for building the model does not care about training points that lie beyond the margin. Analogously, the model produced by SVR depends only on a subset of the training data, because the cost function for building the model ignores any training data close to the model prediction.

Consider the problem of approximating the set of data with a linear function f,

$$f(x) = w \cdot x + b \tag{2.1}$$

The optimal regression function is given by the minimum of the functional,

$$minimize, \phi(\omega, \varepsilon) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\varepsilon_i + \varepsilon_i^*) \tag{2.2}$$

Subject to conditions,

1. $y_1 - \omega \cdot x_i - b \leq \epsilon + \varepsilon_i$
2. $\omega \cdot x_i + b - y_i \leq \epsilon + \varepsilon_i^*$
3. $\varepsilon_i\varepsilon_i^* \geq 0$

The constant C determines the trade of between the flatness of f and the amount up to which deviations larger than ϵ are tolerated.

Optimization problem (2.2) can be solved easily by Lagrangian dual formulation.

$$L = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\varepsilon_i+\varepsilon_i^*) - \sum_{i=1}^{n}\alpha_i(y_i-\omega\cdot x_i-b+\epsilon+\varepsilon_i) + \sum_{i=1}^{n}\alpha_i^*(\omega\cdot x_i+b-y_i+\epsilon+\varepsilon_i^*) + \sum_{i=1}^{n}(\eta\varepsilon+\eta^*\varepsilon^*) \tag{2.3}$$

where, $\alpha, \eta, \alpha^*, \eta^* \geq 0$

Now, b can be computed by Karush-Kuhn-Tucker(KKT) conditions which state,

$\alpha_i(\omega \cdot x_i + b - y_i + \epsilon + \varepsilon_i) = 0$, $\alpha_i^*(y_i - \omega \cdot x_i - b + \epsilon + \varepsilon_i^*) = 0$,

$(C - \alpha_i)\varepsilon_i = 0$, $(C - \alpha_i^*)\varepsilon_i^* = 0$

Hence b can be computed as :

$b = y_i - \omega \cdot x_i - \epsilon$ for $\alpha_i \in (0, C)$,

$b = y_i - \omega \cdot x_i + \epsilon$ for $\alpha_i^* \in (0, C)$

For further details and an insightful description reader is encouraged to refer to [14].

2.4 Random Forests

Random forests are an ensemble learning method for classification, regression that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random forests correct for decision trees' habit of overfitting to their training set.

Ensemble learning are the methods that generate many classifiers and aggregate their results.Two well known methods are *boosting* and *bagging*. In *boosting*, successive trees give extra weight to points incorrectly predicted by earlier predictors. In the end, a weighted vote is taken for prediction. In *bagging*, successive trees do not depend on earlier trees each is independently constructed using a bootstrap sample of the data set. In the end, a simple majority vote is taken for prediction.

Random forests add an additional layer of randomness to bagging. In addition to constructing each tree using a different bootstrap sample of the data, random forests change how the classification or regression trees are constructed. In standard trees, each node is split using the best split among all variables. In a random forest, each node is split using the best among a subset of predictors randomly chosen at that node.

The algorithm is as follows:

Let the number of trees be n_t

1. Take bootstrap samples from the original data n_t times.
2. For each of the bootstrap samples, train an unpruned classification or regression tree such that at each node randomly sample predictors and choose the best split from among those variables.
3. For predictions in new-data aggregate the predictions of the n_t trees (i.e. majority votes for classification, average for regression).

The bootstrapping procedure increases model performance because it decreases the variance of the model, without increasing the bias.

The predictions of a single tree are highly sensitive to noise in its training set but the average of many trees is not, as long as the trees are not correlated. But simply training many trees on a single training set would give strongly correlated trees. Hence, bootstrap sampling is a way of de-correlating the trees by training them using different samples.[15]

2.5 Partial Least Squares Regression

PLS regression is a recent technique that generalizes and combines features from principal component analysis and multiple regression. It is particularly useful when we need to predict a set of dependent variables from a (very) large set of independent variables (i.e., predictors).

Let X be the data matrix and Y be the response vector. PLS regression searches for a set of components (called latent vectors) that performs a simultaneous decomposition of X and Y with the constraint that these components explain as much as possible of the co-variance between X and Y. This step generalizes PCA. It is followed by a regression step where the decomposition of X is used to predict Y.

Simultaneous decomposition of predictors and dependent variables: PLS regression decomposes both X and Y as a product of a common set of orthogonal factors and a set of specific loadings. So, the independent variables are decomposed as $X = TP^T$ with $TT^T = I$ with I being the identity matrix. By analogy with pca T is called the score matrix, and P the loading matrix (in PLS regression the loadings are not orthogonal). Likewise, Y is estimated as $\hat{Y} = TBC^T$ where B is a diagonal matrix with the regression weights as diagonal elements. The columns of T are the latent vectors.

Underlying model:

$$X = TP^T + E \tag{2.4}$$

$$Y = UQ^T + F \tag{2.5}$$

Where, X and Y are data matrix and response respectively, T and Q are projections of X (the X Score) and the projections of Y (the Y score), P and Q are orthogonal *loading* matrices and Matrices E and F are error terms assumed to be i.i.d normal variables. The decompositions of X and Y are made so as to maximize the co-variance between T and U. For further details refer to [16].

In the next chapter we describe the experimental setup and the data used in our study.

3 Dataset Description

The data consists of 1000 songs which have been selected from Free Music Archive (FMA), an online library of high-quality music which if freely accessible. The dataset was developed to 1,000 songs. However, a set of duplicates were later discovered and removed, which reduced the size of the dataset to 744 songs.The extracted 45 seconds excerpts are all re-encoded to have the same sampling frequency, i.e, 44100Hz. The 45 seconds excerpts are extracted from uniformly distributed random starting point in a given song.

The continuous annotations were collected at a sampling rate which varied by browsers and computer capabilities. Therefore, annotations were re-sampled and the averaged annotations with 2Hz sampling rate were generated. The standard deviation of the annotations are also provided. The continuous annotations $\in [-1, 1]$ and due to instability of the annotations at the start of the clips excludes the first 15 seconds. To combine the annotations collected for the whole song, on nine points scale, the average and the standard deviation of the ratings were reported ranging from [1, 9].

3.1 Data Format

All the songs received 10 annotations from which the average and standard deviation were provided. There are Five annotation files in *csv* files. Four files contain average and standard deviation of arousal and valence continuous annotation for each song (*valence cont average.csv,valence cont std.csv, arousal cont average.csv, arousal cont std.csv*) *whole song annotations.csv* contains the average and standard deviation of the static annotations.

3.2 Features

Features extracted by *openSMILE3* are provided. Each *csv* file has a dimension of 744 (Songs) × 6670 (Features). The files contain a header describing the columns (features). After the @data field the data is contained as a comma separated table with rows representing feature vectors and columns features/attributes as described in the header. There are data files for 60 time instances: 15s to 45s every 0.5s.

Full description of the data set can be found at [17].

4 Experimental Setup

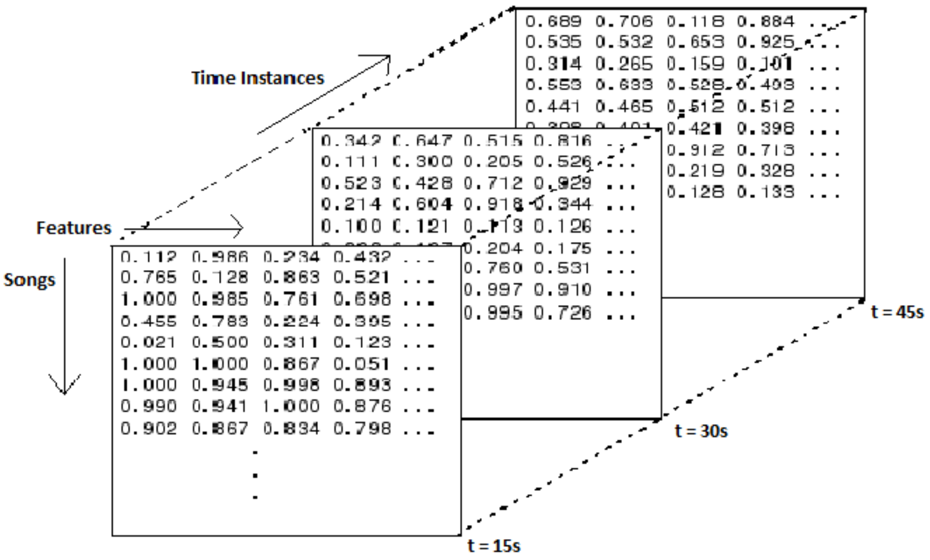
To summarize the last section, the data set can be precisely described as follows:

1. Training data set consists of 744 songs
2. Each song is 30s long clip which is segmented into 0.5s intervals (hence, 60 time instances)
3. For each time instance, we have following :
   - A 2-D data matrix with rows and columns corresponding to songs IDs and different audio features respectively
   - Annotations of Arousal and Valence

Mathematically, Let the 60 time instances be denoted by $t_1, t_2 \dots t_n$. Let number of features or predictors in the model be P and number of songs be n. For each t_i, we train a model M_i i.e. $Y_i = f(X_i)$ where,

- Y_i = Arousal or Valence annotation corresponding to t_i for each song. Y_i is a $n \times 1$ vector.
- X_i = Data matrix with predictors or features for each song (columns = predictors and rows = song IDs). X_i is a $n \times P$ matrix.

The data can be visualized using the following figure.



4.1 Feature selection and dimensionality reduction

The original data set consisted of large number of features ($n << p$), we used following techniques for feature selection.

- Removed all the variables with variance less than a given threshold (0.05).
- Removed all the sets of highly correlated variables in the data but kept 1 member of each set. The two variables are assumed to be highly correlated if correlation value between them is greater than 0.75 ($\rho \geq 0.75$).
- Removed the bottom 50% of variables ranked according to correlation with Y (A/V values).

For dimensionality reduction, we use Principal Component Analysis.

4.2 Model description and assumptions

We used two types of dataset for training models. In case 1, we used whole dataset (without feature selection) and in case 2 we use reduced dataset. In both the cases we are using PCA for dimensionality reduction.

We develop 60 models corresponding to each time instances for prediction of A/V values. Here we have assumed the independence of A/V values for each interval i.e. the A/V value at a given time instant t_i depends only on the features corresponding that instant i.e. X_i. It does not depend on X_{i-1} or X_{i+1}.

Once, models for each t_i are trained, we use them for predicting A/V values in test dataset which consists of 1000 songs.

4.3 Incorporating temporal information

As we know that human emotions do not change rapidly, hence the A/V value at a given t_i depends on its value at t_{i-1} and t_{i+1}. To take into account the temporal information, we have used Gaussian filtering in both directions. This ensures that the A/V values fluctuate smoothly with time. The smoothing length parameter is chosen so as to give the best results.

4.4 Evaluation criteria for models

Different models are evaluated by using correlation (ρ) as the criteria. Here ρ is the correlation between predicted A/V values and ground truth. This criteria has been used in literature for evaluating models for music emotion recognition.

Thus we predicted A/V values for each of the 1000 songs and calculated correlation between predicted values and ground truth for each song. The ρ reported in next chapter is the mean of correlation values for each song.

5 Results and Conclusions

5.1 Using whole dataset - without feature selection

Table 5.1: A/V Correlation Values

	Correlation Values	
Methods (with PCA)	Arousal	Valence
Multiple Linear Regression	0.082 ± 0.220	0.015 ± 0.071
Lasso Regression	0.096 ± 0.045	0.023 ± 0.011
Elastic Nets	0.103 ± 0.092	0.047 ± 0.059
Random Forest	**0.155 ± 0.113**	**0.077 ± 0.045**
SVR		
polynomial	0.098 ± 0.112	0.022 ± 0.054
rbf (kernel)	0.106 ± 0.215	0.034 ± 0.021
Baseline	**0.050 ± 0.430**	**-0.020 ± 0.590**

Using PCA of the whole data set for Arousal and Valence

The above table summarizes the results for the case 1 when the whole dataset (all the features) was used for model development. PCA was used for dimensionality reduction in this case. Since the number of features were comparable to number of observations even after using PCA, we can observe a high standard error for each correlation value.

As we can see the Random Forest method works best for both arousal and Valence prediction and there is a significant increase in correlation when compared with baseline.

5.2 Using Reduced dataset - after feature selection

Table 5.2: A/V Correlation Values

	Correlation Values	
Methods (with PCA)	Arousal	Valence
Multiple Linear Regression	0.103 ± 0.17	0.019 ± 0.04
Lasso Regression	0.196 ± 0.095	0.038 ± 0.015
Elastic Nets	0.215 ± 0.108	0.059 ± 0.021
Random Forest	0.209 ± 0.081	**0.094 ± 0.033**
SVR		
polynomial	0.098 ± 0.812	0.022 ± 0.054
radial	**0.223 ± 0.076**	0.074 ± 0.029
Baseline	**0.050 ± 0.430**	**-0.020 ± 0.590**

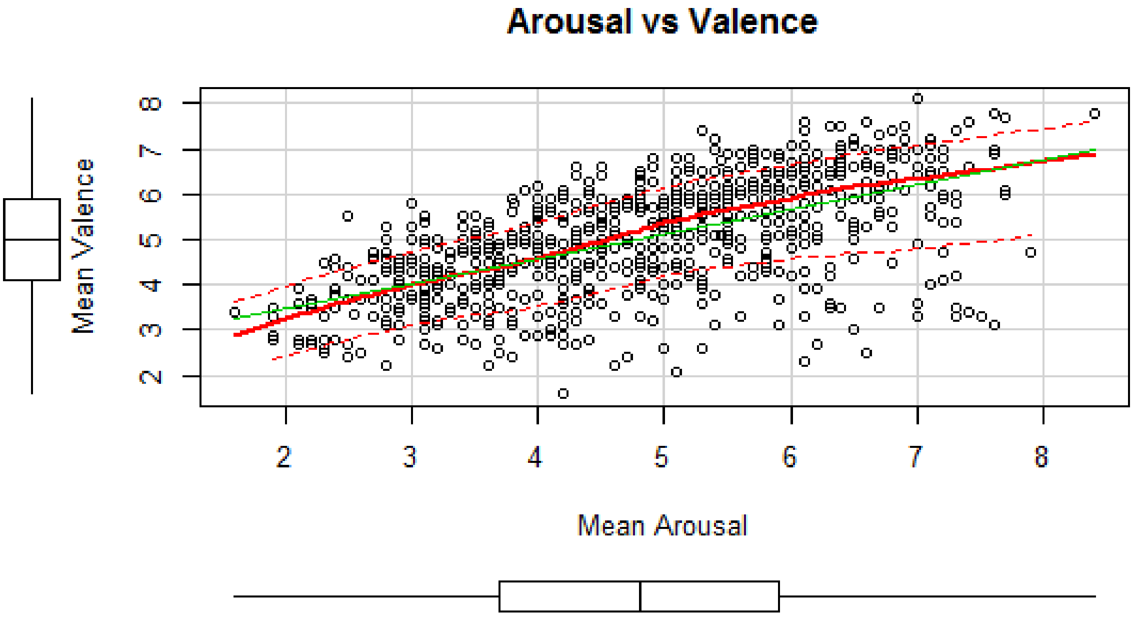
USing PCA of the reduced dataset for the Arousal and Valence

The above table summarizes the results for the case 2 when the feature selection techniques were used for removing the least useful features to reduce the data set size. PCA was used for dimensionality reduction in this case also. Since the number of features were much less than number of observations in this case, we can observe the the standard error is much less as compared to previous case.

As we can see the Random Forest method works best for Valence prediction and the SVR with radial kernel gives best results for Arousal prediction. Also, there is a significant increase in correlation when compared with results using whole dataset and baseline. This can be explained using the concept of overfitting. In case 1 overfitting occurs since the model is excessively complex, such as having too many parameters relative to the number of observations which is not the case with case 2. Also it is well known that the model with overfit has poop predictive performance as it can as amplify minor fluctuations in the data.

5.3 Relationship between Arousal and Valence

We conjectured that additional information about valence can be gained by supplementing arousal data. To observe whether the relationship between these variables exist, we plotted the mean arousal Vs. mean valence for all the songs.



As we can see that there is a strong relationship between arousal and valence values. Arousal is directly proportional to valence. The correlation between these two variables also confirm that these are indeed related. This correlation value comes out to be **0.557**. Hence, Arousal can be used to predict Valence and vice-versa.

5.4 Predicting Valence using Arousal

The relationship between A/v values can be exploited to improve prediction of valence. We used arousal as one of the features for predicting valence.

Table 5.3: Valence Correlation Values

Method	Correlation
SVR	0.096 ± 0.038
Random Forest	**0.126 ± 0.031**

Valence prediction using Arousal

As evident from the correlation values the arousal as a feature improves performance of valence prediction model.

5.5 Conclusion and Future direction of work

We draw following insights in the problem discussed above through our study. Firstly we observe that the high number of features can be reduced to a more manageable set without significant loss of information. This greatly helps in using different methods in the reduced feature set. It is also evident from the results that using the complete feature set introduces overfitting in the training model and hence poorer results for the test dataset. We successfully manage to reduce our dataset to fewer features without using the domain knowledge.

Secondly it is insightful to see the use of predicted arousal values as a feature in prediction of valence. This is an important finding and can be used to model join distribution of Arousal - Valence values.

Finally we also studied the behavior and stability of different regression models subject to our dataset. In coherence with other studies SVM with radial basis kernel and Random Forests models outperform other models in the terms of correlation of the predicted values.

In subsequent studies we wish to apply our methods in conjugation with domain knowledge based feature reduction. We also plan to broaden our set of models used on the data set to include some other famous regression methods like Partial Least Square regression, Quantile Regression etc.

Another possible direction of work is to device new methods to incorporate temporal information in the dataset and extend them beyond Gaussian filtering which we currently use.

Acknowledgement

Firstly we would like to thank Prof. Tanaya Guha for giving us an opportunity to work in her lab under her able guidance. She directed our research toward problems that were both interesting and solvable. We are grateful for her patience and persistence with us as her students.

We would also like to thank our co-advisor, Dr. Subhajit Dutta for his insightful comments on the various methods and possible directions of approach that we could use in our study.

Finally We would like to thank Department of Mathematics and Statistics, IIT Kanpur for giving us the resources, infrastructure and freedom in curriculum to choose our own direction of work according to our personal tastes and interests.

Bibliography

[1] Patrik N Juslin. From everyday emotions to aesthetic emotions: towards a unified theory of musical emotions. *Physics of Life Reviews*, 10(3):235–266, 2013.

[2] Deryck Cooke. The language of music. 1959.

[3] Kate Hevner. Experimental studies of the elements of expression in music. *The American Journal of Psychology*, pages 246–268, 1936.

[4] Marcel Zentner, Didier Grandjean, and Klaus R Scherer. Emotions evoked by the sound of music: characterization, classification, and measurement. *Emotion*, 8(4):494, 2008.

[5] XHJS Downie, Cyril Laurier, and MBAF Ehmann. The 2007 mirex audio mood classification task: Lessons learned. In *ISMIR 2008: Proceedings of the 9th International Conference of Music Information Retrieval*, page 462. Lulu. com, 2008.

[6] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. Music emotion recognition: A state of the art review. In *Proc. ISMIR*, pages 255–266. Citeseer, 2010.

[7] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[8] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia*, pages 1459–1462. ACM, 2010.

[9] Tuomas Eerola, Olivier Lartillot, and Petri Toiviainen. Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. In *ISMIR*, pages 621–626, 2009.

[10] Yi-Hsuan Yang, Yu-Ching Lin, Heng-Tze Cheng, I-Bin Liao, Yeh-Chin Ho, and Homer H Chen. Toward multi-modal music emotion classification. In *Advances in Multimedia Information Processing-PCM 2008*, pages 70–79. Springer, 2008.

[11] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010.

[12] Chang-kee Lee. Some methods of dimension reduction.

[13] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

[14] Steve R Gunn et al. Support vector machines for classification and regression. *ISIS technical report*, 14, 1998.

[15] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[16] Hervé Abdi. Partial least square regression (pls regression). *Encyclopedia for research methods for the social sciences*, pages 792–795, 2003.

[17] Mohammad Soleymani, Micheal N Caro, Erik M Schmidt, Cheng-Ya Sha, and Yi-Hsuan Yang. 1000 songs for emotional analysis of music. In *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia*, pages 1–6. ACM, 2013.