CS6735: Programming Project Report

Identify the Emotion from Music with Machine Learning

Data Scraper
Krishno Dey, Taylor Short, Aditya Thakur, Rafid Farhan

October 2023

1 Introduction

Music one of the main human traditions that has stayed relevant throughout the existence of humanity. It helps connect people with each other and helps connect people with themselves through identity, emotions, and relaxation. Regardless of any individual's reason to listen to music, the surplus of easily accessible music ensures that everyone can find a piece they enjoy. Given the variety of music available, people often rely on music recommendation algorithms to tailor their playlists. While there are many music recommendation metrics including genre, artist, and lyrical content; a music's emotion alone can help guide people to explore new music that matches their familiar style.

Given the large scale of music in existence, machine learning is needed to efficiently classify it. Since emotion in music is largely subjective, it is also important to identify which features best align with music emotion. Our project implements six separate traditional machine learning algorithms including Logistic Regression (LR) [1], K-Nearest Neighbor (KNN) [2], Decision Tree (DT) [3], Random Forest(RF) [4], Gaussian Naive Bayes(NB) [5], and Support Vector Machine (SVM) [6]. Using these models we attempt to predict the emotion in Turkish music and determine the most accurate machine learning algorithm for this problem. We also investigate the neural network models and compare the results to the traditional algorithms. Due to the large input feature set, our implementation reduces the input features from 50 to 40 to improve accuracy and determine the best features. We also apply 10-fold cross-validation on each of the models. The algorithms are trained with a set of 320 Turkish songs and their accompanying emotion categories and evaluated on their prediction of the remaining 80 data samples. Throughout the rest of the article, we use the terms algorithms and models interchangeably.

2 Experimental Setup

In this section, we present the dataset we used in this study and provide an overview of the feature selection methods attempted. We also list the traditional machine learning and neural network-based models that we used in our research and our reasoning for choosing them.

2.1 Data

The Turkish music dataset, developed by Er et al.[2], contains 400 total music samples. The samples are divided into four types with 100 songs each: happy, sad, relax, and angry. Each data sample contains 50 numerical independent features with the goal of predicting the categorical output feature. It contains vocal and non-vocal features across a variety of genres of Turkish music. The distribution of the dataset is shown in the table 1. Although the dataset is relatively small for a machine learning algorithm, it makes a significant contribution to the domain due to the scarcity of the of the publicly available categorized Turkish music data. Obtaining emotionally categorized music data can be difficult because it often has to be hand-labelled. Emotions identified in music are subjective, and can vary between cultures and generations. Additionally, music often changes pace during the course of the song and can contain a variety of emotions within a single piece. Even with carefully filtered high-quality audio, categorizing

Class	# of Samples
Angry	100
Happy	100
Relax	100
Sad	100

Table 1: Data Distributions

music into emotions presents a challenge. We rely on our sample data being consistent and accurate in order to properly train a model to accurately predict emotion in music.

2.2 Feature Selection

With over 50 independent variables per music sample from our dataset, we employed feature selection to reduce irrelevant features and attempt to improve the overall accuracy. We tested three feature selection methods: ANOVA F1, Mutual Information (MI), and Chi-Square. For each feature selection method we selected the 40 most important features. The ANOVA F1 eliminated some of the Mel-frequency cepstral coefficients (MFCC), while the MI and Chi-square kept those features and eliminated 10 features consisting of harmonic change and chromogram means. Section 3 presents the performance of the models including with and without feature selection. Regardless of the eliminated features, the three feature selection methods did not have any significant impact on the machine learning models or their training time as the dataset is quite small. The machine learning models produced better results when run with all 50 features.

2.3 Machine Learning(ML) Algorithms

The primary goal of our experiment is to understand the relationship between music features and emotions and to predict the emotion in a given Turkish music piece. Some considerations for selecting a machine learning model for the emotion identification problem included the small dataset size and limited number of features. For this reason, we initially only considered traditional machine learning techniques. The small dataset would be unable to perform well on a neural network without modification or data augmentation.

Having a small dataset, we could also reasonably expect multiple traditional machine learning model to be efficient. Ultimately, we decided to test a range of different models including Logistic Regression (LR), K-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest(RF), Gaussian Naive Bayes(NB), and Support Vector Machine (SVM). We chose this diverse approach as an exploration in machine learning and to verify our initial predictions on machine learning algorithm selection. Our experiment then compares the performance of these models and discusses the reasons for their performance. While comparing the results we consider the result of LR as the baseline for our study.

2.4 Neural Network based Models

We started out by experimenting with a straightforward 1-layer Artificial Neural Network (ANN) [7] model in our study of neural network-based models given our small dataset. We chose a

straightforward design for this model with one hidden layer. We trained the ANN with a learning rate of 0.001 and a batch size of 32 for 50 epochs. We also tried using a multi-layered ANN to explore a more intricate design. This model has more hidden layers, which allows it to detect more complex patterns in the data. Setting the learning rate for this ANN to 0.001 and training it for 50 epochs allowed us to fine-tune the hyperparameters to maximize their performance.

Furthermore, we looked at whether recurrent neural networks (RNNs) [8] were a good fit for our goal. RNNs are excellent for sequential data and have the capacity to identify temporal relationships in our dataset. To achieve the ideal balance between convergence and overfitting, we trained an RNN model with a particular number of epochs and changed the learning rate. In this instance, we also trained for 50 epochs at a learning rate of 0.001.

3 Experimental Results

In this section, we present the results of the study and compare each of the classical machine learning and neural network-based models. We used Accuracy, Micro Precision, Micro Recall, and Micro F1 score as performance measures to evaluate and compare the models.

3.1 Result of Classical Machine Learning(ML) Algorithms

We applied Logistic Regression (LR), K-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest(RF), Gaussian Naive Bayes(NB), and Support Vector Machine (SVM) classical machine learning algorithms to the Turkish music dataset. A comparison of our three feature selection algorithms in table 4 suggests that there is no significant effect on machine learning performance for this dataset. However, per Table 2 and Table 4 we can see that Support Vector Machine (SVM) has a slightly better performance using feature selection. Even though the increase in accuracy is insignificant, the precision, recall and F1 score of the sad class is significantly increased due to the feature selection. However, since we have a balanced dataset, not many significant changes can be observed, and the results of different evaluation metrics stays fairly consistent in the Table 4. Ultimately, since none of the feature selection algorithms improved performance significantly and did not really reduce the training time, feature selection was omitted from our final machine learning algorithms, and performance metrics were obtained using all 50 features. We also applied 10-fold cross-validation.

Table 2 shows the performance of all the ML algorithms detecting emotion from music. The Logistic Regression (baseline) and Random Forest models produce the best performance, and the precision, recall, and F1 scores of both are similar. Both Linear Regression and Random Forest achieved 82.50% accuracy and excellent precision, recall and f1-scores for each 4 classes. These results are unsurprising as Logistic Regression works well on small and simple datasets. On the other hand, the performance of KNN and DT is quite poor and well below the baseline result. The result of NB and SVM was also below the baseline result of the LR model. It was also found that cross-validation does not have any significant influence over the performance except for SVM. The machine learning models have a higher accuracy when identifying the angry and happy classes. This may be due to the feature distribution of the dataset.

Algorithm	Label	A1	A2	P	R	F1
LR	Angry	82.50	79.50	83.00	86.00	84.00
	Happy			89.00	89.00	89.00
	Relax			94.00	79.00	86.00
	Sad			70.00	76.00	73.00
KNN	Angry	60.00	62.50	59.00	73.00	65.00
	Happy			65.00	83.00	73.00
	Relax			50.00	47.00	49.00
	Sad			67.00	3800	48.00
DT	Angry	68.75	67.50	76.00	86.00	81.00
	Happy			76.00	72.00	74.00
	Relax			63.00	63.00	63.00
	Sad			58.00	52.00	55.00
RF	Angry	82.50	79.25	87.00	91.00	89.00
	Happy			84.00	89.00	86.00
	Relax			84.00	84.00	84.00
	Sad			74.00	67.00	70.00
NB	Angry	75.00	76.00	83.00	91.00	87.00
	Happy			68.00	94.00	79.00
	Relax			72.00	68.00	70.00
	Sad			77.00	48.00	59.00
SVM	Angry	73.75	79.50	82.00	82.00	82.00
	Happy			80.00	89.00	84.00
	Relax			75.00	79.00	77.00
	Sad			56.00	48.00	51.00

Table 2: Performance of Different Machine Learning Algorithms on the Music Dataset. A1: Accuracy, A2: 10 fold cross-validation average accuracy P: Micro-Precision, R: Micro-Recall, F1: Micro-F1

3.2 Result of Neural Network Models

We applied three neural network-based models and their performance is presented in table 3. The table shows that neural networks do not produce much of a better performance than classical machine learning models. Although a very straightforward layer ANN produced a very good performance compared to the ML algorithms. From the table 3 we can see that single-layer ANN achieves 80.00% accuracy on the test and 80.62% on training data, which means there is no overfitting. When we increase the number of hidden layers, for example in ANN and RNN the models overfit on test data. That happens because the dataset is fairly small and if we try to add more hidden layers in NN, the model seems to overlearn the features from the train data.

In our study, the NN models can not beat the performance of ML algorithms like LR and RF. Although a single-layer ANN produces fairly good results. Increasing the complexity of NN models (ANN and RNN) results in overfitting on train data.

From our study, we can say that using large models like NN on a smaller dataset like the one we are using does not produce better results. Particularly, the lower accuracy is not enough to justify the computational complexity that comes with the NN models. Overall our study suggests that using NN models for a smaller dataset might not be a good approach.

4 Conclusions

In this study, we are trying to identify emotion in Turkish music. For this study we used a dataset developed by Er et al. [2], that 50 features that can be used to detect 4 classes of

Algorithm	Label	T_Acc	Acc	P	R	F1
1 layer ANN	Angry	80.62	80.00	78.00	82.00	80.00
	Happy			89.00	94.00	92.00
	Relax			80.00	84.00	0.82
	Sad			72.00	62.00	67.00
Multi-ANN	Angry	92.19	81.25	94.00	77.00	85.00
	Happy			89.00	89.00	89.00
	Relax			84.00	84.00	84.00
	Sad			64.00	76.00	70.00
Multi-ANN	Angry	100.00	80.00	83.00	86.00	84.00
	Happy			89.00	89.00	89.00
	Relax			93.00	68.00	79.00
	Sad			64.00	76.00	70.00

Table 3: Performance of Neural Network models on the Music Dataset. T_Acc: Training Accuracy, Acc: Test accuracy, P: Micro-Precision, R: Micro-Recall, F1: Micro-F1

emotion namely Relaxed, Happy, Sad, and Angry. We first employed classical machine learning models to observe the classification result. As the result suggests most of the LR and RF models produce the best result with an accuracy of 82.50 and a excellent micro f1 score. In addition to that, we also applied 3 different feature selection methods namely ANOVA F1 square, mutual, info square, and chi-square. As the result suggests the feature selection method does improve the performance of these models. Later we also applied neural network-based models without any feature selection method and we observed that these models did not produce a commendable performance compared to the ML algorithms. Among the three neural network models, single-layer ANN produced the best result.

5 Limitations and Future Works

Despite achieving good classification performance with ML algorithms, there are still limitations that require further investigation. We considered the performance of the Logistic Regression model as the baseline of our study. RF was the only model to match the performance of the baseline. The performance of the other models KNN, DT, NB, and SVM fell short of our baseline accuracy. Future work should investigate the reason behind the under-performance of the alternative machine learning models on this dataset. The biases towards angry and happy classes also remains unexplained. Further exploration of the data should also be undergone to provide a viable reason for these biases and attempt to mitigate this in future models.

Neural networks should also be explored further for predicting emotion in music. Although our single-layer ANN performs close to the baseline, more complex neural networks like multi-layered ANN and RNN seem to overfit on training data. Although we suspect the under-performance to be due to the small dataset size, future work should investigate this shortfall. Researchers should work towards building a more robust dataset with more samples, or otherwise augment the data to further investigate deep learning in this area.

Appendix A: Table of classification Result with Feature Selection

	L	Anova F1					Mutual Info					Chi Square				
		A1	**A2**	**P**	**R**	**F1**	**A1**	**A2**	**P**	**R**	**F1**	**A1**	**A2**	**P**	**R**	**F1**
LR	AG	81.3	79.5	83.0	86.0	84.0	81.3	79.5	83.0	86.0	84.0	81.25	79.50	83.0	86.0	84.0
	HP			84.0	86.0	89.0			84.0	86.0	89.0			84.0	86.0	89.0
	RL			89.0	84.0	86.0			89.0	84.0	86.0			89.0	84.0	86.0
	SD			70.0	67.0	68.0			70.0	67.0	68.0			70.0	67.0	68.0
KNN	AG	57.5	68.0	53.0	73.0	62.0	57.5	68.0	53.0	73.0	62.0	57.5	68.0	53.0	73.0	62.0
	HP			65.0	83.0	73.0			65.0	83.0	73.0			65.0	83.0	73.0
	RL			50.0	47.0	49.0			50.0	47.0	49.0			50.0	47.0	49.0
	SD			67.0	29.0	40.0			67.0	29.0	40.0			67.0	29.0	40.0
DT	AG	68.75	67.5	83.0	86.0	84.0	76.25	68.0	83.0	91.0	87.0	72.5	69.75	79.0	86.0	83.0
	HP			75.0	83.0	79.0			75.0	83.0	79.0			75.0	83.0	79.0
	RL			65.0	68.0	67.0			72.0	86.0	70.0			68.0	68.0	68.0
	SD			65.0	52.0	58.0			73.0	68.0	67.0			65.0	52.0	58.0
RF	AG	78.75	79.5	80.0	91.0	85.0	76.25	81.75	83.0	91.0	87.0	81.25	80.0	91.0	91.0	91.0
	HP			80.0	89.0	84.0			81.0	94.0	87.0			80.0	89.0	84.0
	RL			79.0	79.0	79.0			72.0	68.0	70.0			79.0	79.0	79.0
	SD			75.0	57.0	65.0			65.0	52.0	68.0			74.0	67.0	70.0
NB	AG	75.0	75.0	82.0	82.0	82.0	75.0	75.0	82.0	82.0	82.0	75.0	75.0	82.0	82.0	82.0
	HP			68.0	94.0	79.0			68.0	94.0	79.0			68.0	94.0	79.0
	RL			74.0	74.0	74.0			74.0	74.0	74.0			74.0	74.0	74.0
	SD			79.0	52.0	63.0			79.0	52.0	63.0			79.0	52.0	63.0
SVM	AG	78.75	80.5	86.0	82.0	84.0	78.75	80.25	86.0	82.0	84.0	78.75	80.25	86.0	82.0	84.0
	HP			84.0	89.0	86.0			84.0	89.0	86.0			84.0	89.0	86.0
	RL			88.0	74.0	80.0			88.0	74.0	80.0			88.0	74.0	80.0
	SD			62.0	71.0	67.0			62.0	71.0	67.0			62.0	71.0	67.0

Table 4: Comprehensive Breakdown of the Classification Results using feature selection on the Music Dataset. L: label, AG: Angry, HP: Happy, RL: Relax, SD: Sad, A1: Accuracy, A2: 10 fold cross-validation average accuracy, P: Micro-Precision, R: Micro-Recall, F1: Micro-F1

References

[1] Arun K. Kuchibhotla, Lawrence D. Brown, Andreas Buja, and Junhui Cai. All of linear regression, 2019.

[2] Mehmet Bilal Er and Ibrahim Berkan Aydilek. Music emotion recognition by using chroma spectrogram and deep visual features. *Int. J. Comput. Intell. Syst.*, 12:1622–1634, 2019.

[3] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986.

[4] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[5] Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

[6] Nello Cristianini and Elisa Ricci. *Support Vector Machines*, pages 928–932. Springer US, Boston, MA, 2008.

[7] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[8] Juergen Schmidhuber. On learning to think: Algorithmic information theory for novel combinations of reinforcement learning controllers and recurrent neural world models, 2015.