IAT481 Assignment 3: Speech Emotion Recognition Dataset Analysis

Wei Xing Deng #301442155

Word count: 837

School of Interactive Arts and Technology, Simon Fraser University

IAT 481W D100: Exploring Artificial Intelligence: Its Use, Concepts, and Impact

Professor: Dr. O. Nilay Yalcin

TA: Maryiam Zahoor

Feb 8, 2024,

Introduction

In this report, I analyzed the comparing performance of machine learning models trained on two different datasets: the RAVDESS-only and a combined dataset between RAVDESS and EmoDB. My findings reveal a notable increase in model accuracy with the combined dataset for several classifiers, such as RandomForestClassifier, SVC, and KNeighborsClassifier.

Comparing Models:

After comparing the results from training models on two different datasets, the combined dataset offers better performance across the board compared to the RAVDESS dataset. For instance, the accuracy score for the RandomForestClassifier jumped from 51.92% with just RAVDESS to 63.58% with the combined dataset. This trend is seen with the KNeighborsClassifier and SVC with a linear kernel too, where their scores improved from 51.57% and 49.83% to 57.62% and 55.96%, respectively, when using the combined dataset. This improvement likely stems from the combined dataset's richer and more diverse data, which provides a better-rounded view for the models to learn from. It seems like having more varied data helps the models to not just focus on the specific characteristics of the RAVDESS dataset but to understand broader patterns. This way, they become better at making predictions on data they haven't seen before, which explains the boost in their performance.

Figure.1

Out[ ]:

	Classifier	Accuracy Score
4	RandomForestClassifier	51.92%
0	KNeighborsClassifier	51.57%
1	SVC	49.83%
2	SVC RBF kernel	47.74%
3	DecisionTreeClassifier	37.63%
6	GaussianNB	26.48%
7	QuadraticDiscriminantAnalysis	24.04%
5	AdaBoostClassifier	23.34%

Note. From *RAVDESS-Only dataset of Colab* [Screenshot], by D. Weixing, 2024,

Figure.2

	Classifier	Accuracy Score
4	RandomForestClassifier	63.58%
0	KNeighborsClassifier	57.62%
1	SVC	55.96%
2	SVC RBF kernel	46.36%
5	AdaBoostClassifier	43.38%
3	DecisionTreeClassifier	41.72%
7	QuadraticDiscriminantAnalysis	36.75%
6	GaussianNB	35.10%

Note. From *Combined dataset of Colab* [Screenshot], by D. Weixing, 2024,

SVC:

After I ran the SVC model on the combined dataset, I saw a similarity with the accuracy results from the RAVDESS-only dataset given by our lab. Both datasets gave 100.00% accuracy on the training set, but the test set accuracy was different. The RAVDESS-only dataset holds 56.45% test set accuracy, which is better than the 50.33% from the combined dataset. This finding taught me that adding more diverse data does not necessarily result in improved accuracy on unseen data. I would expect that having more data available would improve the accuracy of the model, as more data should lead to better learning; but the model would need to be simple enough to generalise well with added diverse data, resulting in less than accurate test accuracy.

Figure.3

```
SVC Model's accuracy on training set is 100.00%
SVC Model's accuracy on test set is 56.45%
```

```
SVC Model's accuracy on training set is 100.00%
SVC Model's accuracy on test set is 50.33%
```

Note. From *both* Combined and *RAVDESS-Only dataset of Colab* [Screenshot], by D. Weixing, 2024,

Knn (k Nearest Neighbors):

Incidentally, by remembering my experience with the kNN classifier on those parameters from the lab - which were used to create the graph above - it's easy to see what difference choosing or flattening a dataset can make to the type of model that results. For example, when these predefined settings (`n_neighbors` was set to 5; `distance` was used to weight the algorithm; and a `brute` algorithm was used) were applied to create the classifier, the test set accuracy drastically increased from 49.48 per cent (RAVDESS-only dataset) to 58.28 per cent (combined dataset), emphasising how beneficial a more varied dataset was, as the different yet similar combinations of words probably gave the model some real fodder to learn off of, making it more likely to generalise well on the unseen data.

Figure.4

```
Default kNN Model's accuracy on training set is 66.90%
Default kNN Model's accuracy on test set is 43.55%

kNN Model's accuracy on training set is 100.00%
kNN Model's accuracy on test set is 49.48%
```

```
Default kNN Model's accuracy on training set is 71.23%
Default kNN Model's accuracy on test set is 53.64%

kNN Model's accuracy on training set is 100.00%
kNN Model's accuracy on test set is 58.28%
```

Note. From *both* Combined and *RAVDESS-Only dataset of Colab* [Screenshot], by D. Weixing, 2024,

Random Forest:

After applying the Random Forest classifier to both a specialized dataset (RAVDESS-only) and a more diverse one (combined dataset), I noticed interesting patterns in performance metrics. The initial setup of the Random Forest showed that models trained on a broader set of data points—evidenced by the jump in test set accuracy from 51.57% for RAVDESS to 60.26% for the combined dataset—tend to have a superior ability to generalize. This improvement underscores the value of incorporating a wide array of training examples. Furthermore, when I refined the model's settings by adjusting parameters such as the number of estimators and the criterion used for splitting, the test accuracy improved even more, reaching 63.25% with the combined dataset, compared to 52.96% with just RAVDESS. These findings highlight how a strategic approach to both the selection of training data and the tuning of model parameters can synergistically enhance the performance of machine learning models, particularly in tasks requiring nuanced differentiation across varied inputs.

Figure.5

```
Default Random Forest Model's accuracy on training set is 100.00%
Default Random Forest Model's accuracy on test set is 51.57%

Random Forest Model's accuracy on training set is 100.00%
Random Forest Model's accuracy on test set is 52.96%
```

```
Default Random Forest Model's accuracy on training set is 100.00%
Default Random Forest Model's accuracy on test set is 60.26%

Random Forest Model's accuracy on training set is 100.00%
Random Forest Model's accuracy on test set is 63.25%
```

Note. From *both* Combined and *RAVDESS-Only dataset of Colab* [Screenshot], by D. Weixing, 2024,

Conclusion:

To sum up my exploration, the takeaway is clear: a richer, more varied dataset significantly boosts the accuracy of classifiers such as RandomForestClassifier, SVC, and KNeighborsClassifier. This highlights the critical role of dataset diversity in optimizing the performance of machine learning models.

Reference

Python 2.7 Tutorial. (2024). Pitt.edu. https://sites.pitt.edu/~naraehan/python2/tutorial9.html

OpenAI. (2023). ChatGPT (Mar 14 version) [Large language model]. https://chat.openai.com/chat