A Customized Machine Learning Pipeline to Build State-of-the-Art Audio Classifiers

Sruthi Kurada

***Abstract* – Audio classifiers have many real-world applications, from informing medical diagnoses to revealing automobile malfunctions. In this study, I have explored strategies to build an accurate classifier to categorize environmental sounds from the UrbanSound8K dataset. Published classifiers on this ten-class dataset only have 50-79% accuracy. Through engineering a machine learning pipeline, I have built a state-of-the-art classifier with a 99% test-set accuracy on this dataset. In order to examine the general applicability of this pipeline to build reliable classifiers on other audio datasets, I have examined its performance in differentiating four unique heart sounds and found it to be equally effective. The final heart sound classifier achieved a 98% test set accuracy.**

I. INTRODUCTION

Audio classifiers categorize acoustic input based on inherent sound characteristics. Many classifiers have been created on the UrbanSound8K [1], a 10-class public dataset of 8732 city sound files. The highest performing among these classifiers achieved a 79% accuracy [2]. I sought to build a classifier with a higher accuracy on this dataset. Subsequently, I aimed to create a generalizable pipeline that enables building high performance classifiers on other audio datasets. A generalizable pipeline could be integral to the development of high-fidelity audio classifying machines for all applications.

II. METHODS

Through employing Python and the librosa library [3], audio features including Mel-frequency cepstral coefficients (MFCCs), Chroma, Mel, Contrast, and Tonnetz were extracted from the UrbanSound8K dataset. These components were used to train various machine learning classifiers using the Sci-kit learn library [4]. Classifiers included in the study were Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Classification and Regression Trees (CART), Naive Bayes (NB), and Support Vector Machines (SVM). These classifiers were further tuned by applying dimensionality reduction, pre-processing, hyper-parameter optimization, and data-augmentation techniques [4-5].

Accuracy was the primary metric used to quantify classifier performance. In the assessment phase, source data was randomly split (80:20 ratio) into training and testing sets. K-fold (k=10) cross validation was then employed on the training set, and testing was performed on the held-out test set. To assess the generalizability, the resulting best-performing pipeline on the UrbanSound8K dataset was applied to a four-class heart sound dataset [6].

III. RESULTS AND DISCUSSION

Figure 1 shows that the KNN classifier achieved the highest baseline 10-fold cross-validation accuracy rate (mean ± S.E.M; 82.7 ± 0.02%) of all of the classifiers in the study. The figure also shows that the KNN classifier performed similarly when provided the full audio feature-set as compared to when given only MFCCs.

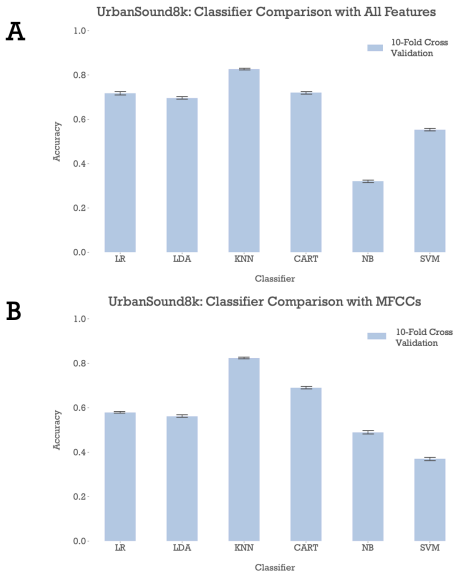

Figure 1A: *Baseline classification accuracy of six classifiers with the full feature set provided (MFCCs, Chroma, Mel, Contrast, Tonnetz). Classifiers included Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Classification and Regression Trees (CART), Naïve Bayes (NB), and Support Vector Machines (SVM).*

Figure 1B*: Baseline acuracy of the same classifiers with only MFCCs as input.*

Thus, I chose to employ MFCCs as the sole input to the KNN classifier before attempting pre-processing, hyper-parameter optimization, and data augmentation steps to improve its accuracy. Table 1 summarizes improvements achieved after implementing the above techniques. A 99.36 ± 0.09% cross validation accuracy was achieved on the training set by applying these three additional tuning steps. The final pipeline earned a 99.47% test set accuracy.

Baseline KNN with all features	82.66 ± 0.43%
+Only MFCCs (Dim. Red)	82.37 ± 0.45%
+Data Preprocessing	89.00 ± 0.44%
+Hyperparameter Optimization	94.60 ± 0.21%
+Data Augmentation	99.36 ± 0.09%
Test Set Performance	**99.47%**

Table 1*: The increases in performance enabled by including dimensionality reduction, pre-processing, hyper-paramater optimization, and data augmentation in the final UrbanSound8K KNN classification pipeline enabled achieving a 99.47% test set accuracy.*

Figure 2 shows the confusion matrix of the tuned KNN pipeline. The majority of the false detections observed were between the jackhammer and drilling classes.

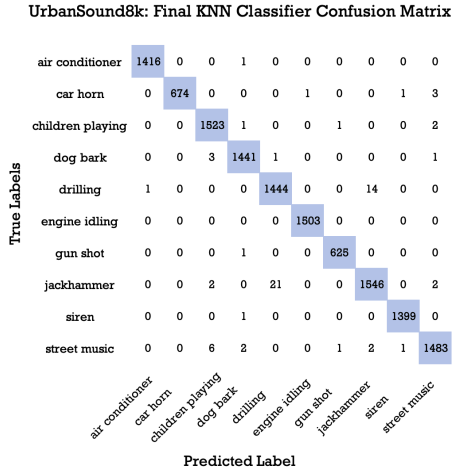

Figure 2: *Confusion Matrix of the final UrbanSound8K KNN pipeline*

In order to assess the pipeline's generalizability, the same procedures were applied to a separate dataset – a four-class heart sound dataset. A KNN classifier was the highest performer on the new dataset (Table 2), despite CART and LR classifiers performing better initially* (Figure 3).

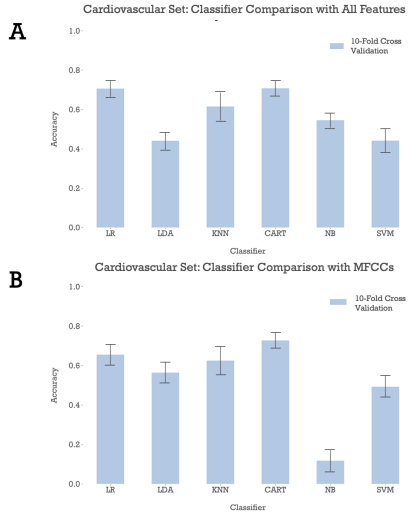

Figure 3A: *Baseline cardiovascular sound classification accuracies when the full feature set is provided.*

Figure 3B: *Baseline acuracies of the same classifiers with only MFCCs as input.*

Table 2 (following page): *The stepwise boosts in accuracy achieved by implementing dimensionality reduction, pre-processing, hyper-paramater optimization, and data augmentation in the final heart sound KNN classification pipeline.*

Baseline KNN with all features	61.56 ± 7.07%
+Only MFCCs (Dim. Red)	62.56 ± 7.10%
+Data Preprocessing	71.78 ± 4.61%
+Hyperparameter Optimization	71.78 ± 4.61%
+Data Augmentation	98.48 ± 0.09%
Test Set Performance	**97.791%**

IV. DISCUSSION

Optimized KNN pipelines were the highest performers on both datasets in this study. As shown by Figure 1 and 3, the KNN classifiers performed equally well with either the full audio feature set as input or with MFCCs as the sole input. This finding has several advantages. Not only is feature extraction faster with one input, but scaling is also more efficient and overfitting is minimized. The program's memory footprint is lowered as well. As shown in Table 1 and 2, preprocessing input data, optimizing classifier hyper-parameters, and using augmented datasets were essential for achieving a high accuracy.

The above pipeline was also employed on additional audio datasets in order to further examine its generalizability [6, 7]. The pipeline achieved >90% cross-validation and test-set accuracies on these datasets as well (data not shown).

V. CONCLUSION

The above results illustrate that a ML pipeline including pre-processing, hyper-parameter optimization, and data augmentation steps can be used to build state-of-the-art audio classifiers on multiple datasets with MFCCs. This approach enables developing models with high performance and low variance while avoiding overfitting. In the future, I will explore the effectiveness of this pipeline for semi-supervised learning applications.

VI. REFERENCES

[1] J. Salamon, C. Jacoby and J. P. Bello, "A Dataset and Taxonomy for Urban Sound Research", 22nd ACM International Conference on Multimedia, Orlando USA, Nov. 2014, pp. 1041–1044.

[2] J. Salamon, and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," *IEEE Signal Processing Letters*, Nov. 2016, pp. 279 – 283.

[3] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar et al, "Librosa: Audio and Music Signal Analysis in Python." *Proceedings of the 14th Python in Science Conference*, July 2015, pp. 18-25.

[4] F. Pedregosa , G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel et al, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*. Nov 2011, pp. 2825-2830.

[5] B. McFee, E. Humphrey, and J.P. Bello, "A software framework for musical data augmentation." *International Society for Music Information Retrieval Conference*, Aug 2015, pp. 248-254.

[6] P.J. Bentley, G. Nordehn, M. Coimbra, S. Mannor, R. Getz, "The PASCAL Classifying Heart Sounds Challenge 2011," [online] Available: www.peterjbentley.com/heartchallenge/.

[7] G. Clifford, C. Liu, B. Moody, D. Springer, I. Silva, Q. Li, and R. Mark. "Classification of normal/abnormal heart sound recordings: The PhysioNet/Computing in Cardiology Challenge 2016," *Computing in Cardiology Conference (CinC) - IEEE*, Sep. 2016, pp. 609-612.