Music Genre Classification: A Comparison Between Continuous Features and Discrete Features in An Imbalanced Dataset

1.0 Introduction

Music genre classification is an area of research that has been studied for decades. It is an automatic tool to identify which genre a piece of music belongs to with the information extracted from the music data in a fast and effective way (Correa & Rodrigues, 2014). Despite the difficulty of the task, it can simplify people’s work and provide important value in pure research and commercial application (McKay & Fujinaga, 2006).

Armentano et al. (2016) demonstrate a point that humans identify music genres by comparing features that we can understand and interpret, such as the cultural background and musical knowledge that we know. It means that humans accomplish classification by only using high-level information, even though there may be other features provided to us, we only utilize the one that could possibly give us the most accurate result. This shows that not all features might be useful when classifying genre. In addition, it is found that the dataset has an imbalanced distribution of labels, which might influence the performance of the classification. Thus, the hypothesis for this paper is that using only one type of features with balanced training data would increase the overall performance of classifiers. The paper will develop a music genre classification on only the continuous feature and a combination of both continuous and discrete features with or without sampling to compare performances between them.

To achieve this goal, I will first discuss the related work done by other researchers. Next, the paper will analyze the features that we extracted and select a few classifiers to participate throughout the classification. After the implementation, the results will be provided along with a discussion to demonstrate the overall performance and shortages for this experience.

2.0 Literature Review

Dannenberg et al. (1997) were one of the first that demonstrated the use of machine learning can create effective music classifiers rather than "hand-coded approaches". They used 13 low-level features and compared the performance between naive Bayesian, linear, and neural network classifiers. The authors discovered that the Bayesian classifier had the best and stable result with above 90% accuracy.

Another paper by McKay and Fujinaga (2004) discussed the importance of dividing sets of features into different levels. To do so, the total of 109 musical features extracted from symbolic recordings are analyzed regarding their weighting. The paper used a combination of neural networks and k-nearest neighbour as classification techniques. The outcomes are similar to Dannenberg et al. (1997) with a success rate of 98% for 3 root genres and 90% for 9 leaf genres. However, a more recent study by Armentano et al. (2016) identified that using only high-level features increase the likelihood of obtaining better performance for music classification. Armentano outperformed McKay and Fujinaga (2004) with the same taxonomy in their dataset by only taking 3 features into the algorithm.

3.0 Data Analysis

In this section, I will analyze the entire dataset by the given features and labels. The dataset in this paper is obtained from both Bertin-Mahieux et al. (2011) and Schindler & Rauber’s (2012) report. It has been divided into training, validation, and testing set.

3.1 Feature Analysis

There are three types of features in the dataset, which are textual, continuous, and discrete. Each type of feature will be discussed regarding the degrees of importance for this particular case and remove those features that may potentially reduce the performance of the classifiers.

3.1.1 Textual Data

The first type of features is the text-based features, which are "title" and "tags" containing multiple words in each instance. However, Words in testing instances may not occur in the training instances, it might influence the performance of classifiers. Text-based features are also not included in this paper's hypothesis. For the purpose of this report, it is decided to ignore this type of features.

3.1.2 Continuous Data

Continuous data includes "loudness", "tempo", "duration" and all of the 148 audio features. This type of features provide unique value that separates them from every instance, each instance's relationship with labels can be measured by distances. Hence, it is assumed that this type of features would contribute to the most accurate result.

3.1.3 Discrete Data

"Time signature", "key", and "mode" are discrete data, these data might not be good indicators because of the limited value that they offer. For instance, the "mode" feature provides binary value to differentiate between labels. However, there are eight labels in total, it cannot provide a clear view to classifiers to predict the correct answer. Besides, the distribution of this feature is imbalanced where "1" appears twice often than "0" (see Table 1). Therefore, it is reasonable to believe that discrete features may reduce the performance of classification.

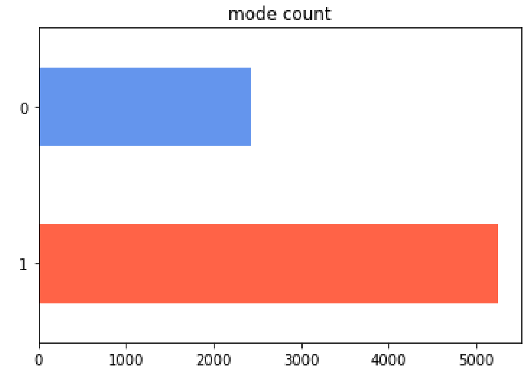


Table 1: distribution of instances in "mode"

3.2 Label Analysis

There are 8 different classes in the dataset, table 2 presents the distribution of classes for all instances in the training data. It shows that the training dataset has imbalanced labels. The most frequent class has 1326 more instances than the fewest class. Classifiers would get biased towards the majority class. In this case, "folk" and "classic pop and rock" have a high chance to be misclassified to replace other classes. Therefore, a sampling process is considered with the use of the imbalanced-learn library. I used the undersampling, which is a technique to drop some of the instances in training data that belong to the high-frequent classes (Weiss et al, 2007). By doing so, each class is at the same level of distribution. The classification process will show whether undersampling can improve the overall performance.

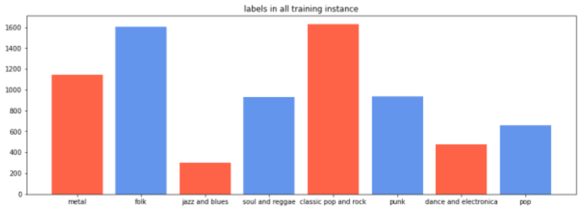


Table 2: label distribution in all training instance

4.0 Classifier

This section will introduce the classifiers that have been selected in this paper.

4.1 Multinomial Logistic Regression (MLR)

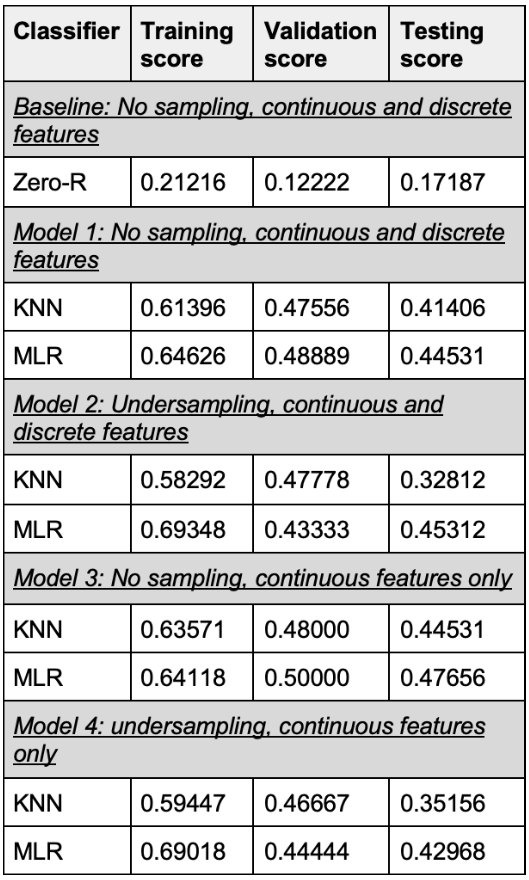
MLR is a simple extension of logistic regression, which allows users to predict a class that is more than two categories using multiple independent variables that can be either binary or continuous (Starkweather & Moske, 2011). It works as a series of logistic regression to estimate the likelihood of a label. This approach has an assumption stating that "the choice of one category is not related to the choice of another category" (Starkweather & Moske, 2011). It could improve the performance in an imbalanced dataset, which matches our case very well.

4.2 K-Nearest Neighbour (KNN)

KNN is an approach that measures distances, it operates by treating each instance as a point and finds the classes of a number of k closest to a test instance (McKay & Fujinaga, 2004). KNN is used as a benchmark in relation to MLR.

5.0 Result and discussion

This section shows the classification performance of the two selected classifiers, MLR and KNN, and one baseline classifier, Zero-R. Standardization is used during the preprocessing phase to organize the numerical data, it has been tested that works better than Normalization and One Hot Encoding in this dataset. Regarding the scope of this paper, detail description or comparison would not be mentioned. I set up four models according to the imbalanced dataset and features hypothesis. Table 3 records the accuracy score that each scenario produced in training, validation and testing dataset.



Classifier	Training score	Validation score	Testing score
Baseline: No sampling, continuous and discrete features			
Zero-R	0.21216	0.12222	0.17187
Model 1: No sampling, continuous and discrete features			
KNN	0.61396	0.47556	0.41406
MLR	0.64626	0.48889	0.44531
Model 2: Undersampling, continuous and discrete features			
KNN	0.58292	0.47778	0.32812
MLR	0.69348	0.43333	0.45312
Model 3: No sampling, continuous features only			
KNN	0.63571	0.48000	0.44531
MLR	0.64118	0.50000	0.47656
Model 4: undersampling, continuous features only			
KNN	0.59447	0.46667	0.35156
MLR	0.69018	0.44444	0.42968

Table 3: accuracy score comparison

5.1 Training, Validation, Testing dataset

Based on the result from table 3, the average scores in the training dataset are around 60% to 70%, but it is 30% to 50% in the validation and testing dataset. Training dataset scores higher than the others; it implies that all of the models may be overfitted. This situation occurs when a model learns the detail in the training data too well. Overfitting a model will have poor performance on new data, in this case, the validation and testing data. Therefore, the models produced in this paper may not work well in future predictions.

5.2 Baseline Classifier

Baseline classifier is taking the most simplistic view of data in the classification method to test whether the other classifiers have achieved the minimum requirements. This report uses Zero-R classifier, which simply determines every instance as the most frequent class in the dataset. As table 3 showed, the performance of the Zero-R classifier for both validation and testing set is below 0.2, this is two times smaller than other classifiers that I implemented in the paper. Thus, both MLR and KNN in 4 different scenarios reach the minimum performance requirement.

5.3 Benchmark Classifier

As KNN is selected as the benchmark classifier in this study, each model has been tested based on both KNN and MLR. It is observed that MLR has an overall higher performance than KNN, the only exception is the validation dataset in model 2 in which KNN performed 4% better. The MLR model also is faster than KNN since KNN takes time to find the optimal k before training the data.

5.4 Full Dataset and Undersampling

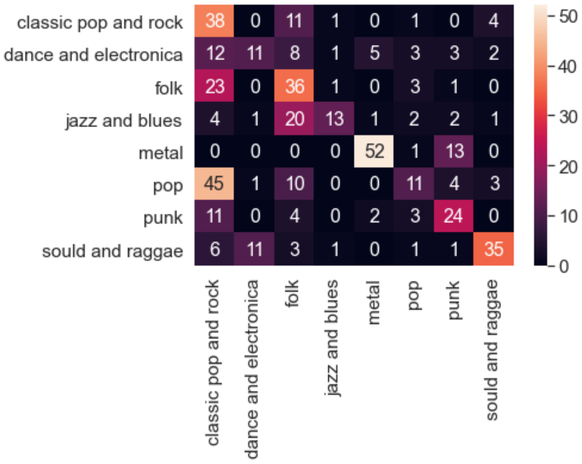
Although there are some scores in the training set showing that models with undersampling methods performed slightly higher than using all instances, the overall outcomes are worse when sampling correction applied. The reason behind this could be: the number of instances from training dataset is reduced by more than three times, it may possibly skip some of the representative instances and lead to inaccurate final results. It is also worth mentioning that the sampling correction has a lower effect on KNN than MLR since the scores in KNN are more stable.

5.5 Continuous and Discrete Features

The comparison of continuous and discrete features can be done by model 1 versus model 3 from table 3 since we stated that the full dataset works better than the partial dataset, so we can focus on model with the full dataset. It is discovered that with only continuous features, the accuracy score from all set will all be slightly higher than using both types of features. It implies that not all features may be relevant to a selected classifier or dataset, which proves that our hypothesis on feature selection is meaningful. So it is vital to evaluate and choose the most applicable features that bring value to the classification.

5.6 Error Analysis

As table 3 showed, model 3 provides the most accurate score compared to other models. A confusion matrix (see figure 1) and a classification report (see table 4) are built based on this model for the validation set using MLR to evaluate the errors. We observe that the "classic pop and rock" tend to confuse MLR as the precision is only 0.27, which means that many other music genres are mistakenly classified to be this genre. Because of this, the recall rate for other genres is low, especially for "pop" music since most "pop" have been identified as "classic pop and rock". In fact, this situation makes sense as these two types of music are quite similar in real life. To deal with this issue, a reasonable deduction of "classic pop and rock" instances may help. It is also useful to identify the features that can better differentiate those two classes.



	classic pop and rock	dance and electronica	folk	jazz and blues	metal	pop	punk	sould and raggae
classic pop and rock	38	0	11	1	0	1	0	4
dance and electronica	12	11	8	1	5	3	3	2
folk	23	0	36	1	0	3	1	0
jazz and blues	4	1	20	13	1	2	2	1
metal	0	0	0	0	52	1	13	0
pop	45	1	10	0	0	11	4	3
punk	11	0	4	0	2	3	24	0
sould and raggae	6	11	3	1	0	1	1	35

Figure 1: confusion matrix for model 3

	Precision	recall	F1-score
Classic pop and rock	0.27	0.69	0.39
Dance and electronica	0.46	0.24	0.32
Folk	0.39	0.56	0.46
Jazz and blues	0.76	0.30	0.43
Metal	0.87	0.79	0.83
Pop	0.44	0.15	0.22
Punk	0.50	0.55	0.52
Soul and reggae	0.78	0.60	0.68

Table 4: classification report for model 3

6.0 Conclusion and Future Work

The purpose of this study is to analyze if using one type of features with a balanced training set would enhance classification performance. The results conclude that only having continuous features will increase the accuracy score that the classifiers produced. Unfortunately, the current report limits the scope and doesn't go deep into each column to identify several features rather than using a whole type of features.

In addition, to deal with one of the biggest limitations of this dataset, which is the unequal distribution of the classes in the training data, the undersampling approach was applied. The results show that this method lowers the overall performance. For the future study, other sampling methods can be considered such as oversampling and SMOTE. By doing so, it is possible to identify methods that best suit the dataset. This is also applicable when selecting other classifiers since both MLR and KNN are tend to overfit these models. Thus, it is confident to believe that the classification project on this dataset will improve throughout the change.

References

Armentano, M. G., De Noni, W. A., & Cardoso, H. F. (2017). Genre classification of symbolic pieces of music. *Journal of Intelligent Information Systems: Integrating Artificial Intelligence and Database Technologies*, *48*(3), 579. https://doi.org/10.1007/s10844-016-0430-7

Bertin-Mahieux, T., Ellis, D. P.W., Whitman, B., & Lamere, P. (2011). The million song dataset. In Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)

Corrêa, D. C., & Rodrigues, F. A. (2016). A survey on symbolic data-based music genre classification. *Expert Systems With Applications*, *60*, 190–210. https://doi.org/10.1016/j.eswa.2016.04.008

Dannenberg, R. B., Thom, B., & Watson, D. (1997). A machine learning approach to musical style recognition. In Proceedings of the international computer music conference (pp. 344–347).

McKay, C., & Fujinaga, I. (2004). *Automatic genre classification using large high-level musical feature sets.*

McKay, C., & Fujinaga, I. (2006). *Musical genre classification: Is it worth pursuing and how can it be improved?* 101–106.

Schindler, A., & Rauber, A. (2012). Capturing the temporal domain in Echonest Features for improved classification effectiveness. InProceedingsofthe10thInternationalWorkshoponAdaptiveMultimedia Retrieval (AMR).

Starkweather, J., & Moske, A. K. (2011). Multinomial logistic regression. Retrieved from http://www.readbag.com/unt-rss-class-jon-benchmarks-mlr-jds-aug2011

Weiss, G. M., McCarthy, K., & Zabar, B. (2007). Cost-sensitive learning vs. sampling: Which is best for handling imbalanced classes with unequal error costs?. DMIN, 7, 35-41.