Classification based on identified clusters in Spotify Dataset

Parnab Das
parnab.das2@mail.dcu.ie
School Of Computing,
Dublin City University
19211289

Souradip Goswami
souradip.goswami2@mail.dcu.ie
School Of Computing,
Dublin City University
19210273

Aritra Dutta
aritra.dutta2@mail.dcu.ie
School Of Computing,
Dublin City University
19211293

***Abstract*—With the digital manoeuvre, evolution has happened to the way we used to listen music and access it. Music streaming over the web is a more preferred way now, which routes to the need of an efficient Music Information retrieval system that well classifies and organises the music data. Classification of music can be done based on many abounding features, but in this paper we circumscribe our research only on acoustic features of 100,000 music tracks in the Spotify dataset. In this paper, we perform a qualitative comparison of different classification techniques based on the distinguished clusters, that we procure while clustering the data with heuristic and probabilistic approaches. Dimensionality reduction of data is done exerting PCA, to extract and consider the relevant principal components that accounts for most varied influences in the data. And then the effectiveness of the acoustic features in each classification are compared using classifiers like SVM, Random Forest and Naive Bayes algorithms. Predictability scores from the confusion matrix of the classifiers clearly indicates Naive Bayes algorithm to be best approach for classifying acoustic features of music.**

I. INTRODUCTION

Music Information Retrieval has been an important research aspect in present day scenario to delve into, with the advent and adoption of digitized music streaming services. As the media archives are growing exorbitantly on a daily basis, an intelligent adaptive searching technique to retrieve information in more efficient way is the need of the hour. Out of all techniques, feature classification is one that develops associated links between different features, extracted from the song to build up a commonality for better browsing functionality. These techniques are further aggravated and employed for building up marketing tools (for example song recommendation system based on customized retrieved information traits). Among the giant players of music streaming services, Spotify is currently leading the chart with a 36% market share of music-streaming subs, empowering and upending the music industry around the globe. Spotify was developed in 2006 by Daniel Ek, former CTO of Stardoll, and Martin Lorentzon, co-founder of TradeDoubler in Stockholm, Sweden. [14].

Unfoldment of diversified cultures around the globe throughout the time has made a wider impact on the music industry, stemming to the evolution of more music genres into the ears of the peer. As people listening to more genres now, classification of music features is a more of researched topic now, where most researchers are diving into. As music is evolving and the combinatorial links of traits between different genres are increasing, artists/singers are drawing motivation from other genres and reflecting upon their exertions. [11]

In this research paper, we try to attain the music feature clustering and classification problem by evaluating and comparing different approaches adopted in various research works. We use the Kaggle dataset [3] for our feature classification, consisting of audio features for 232,725 tracks. The Audio features are being extracted from the tracks through an API call made to the Spotify web API, containing audio features that are further explained in the section III. We analyzed those already extracted feature vectors to do a comparative analysis of how different classifier algorithms classifies using the numerical features of a song.

II. RELATED WORK

Many researches have been directed on music feature classification till date, and to linchpin further, these have been on abounding features extracted from various sources whether it is text based features from song lyrics, content based features from audio signals or reference features counting title and composer details. In this paper, we narrow down our focus to only acoustic features of 232,725 song tracks from Spotify. [3]

In this research paper by Li, Ogihara and Qi [11], they have proposed a new feature extraction method for music classification, DWCHs. Classifiers used for multiclass classification in this paper are SVM1, SVM2, Multiclass Proximal Support Vector machine(MPSVM), Gaussian Mixture Modelling(GMM), Linear Discriminant Analysis(LDA), k-nearest neighbors(KNN), and distinguishes SVM as the best algorithm for content based feature classification. We proceed to diversify on this claim with an approach to adopt reduction methods to reduce the multi modal acoustic features and compare other classifiers like Naive Bayes, Random Forest with SVM to account the predictability.

Another paper by Li, Ogihara, Peng, Shao, and Zhu on muic clsutering with features from various information sources [12], identifies coequal artists based on features from different sources, that are integrated by clustering techniques to perform bimodal learning. The approach has been more on comparing the constraints noticed while performing multimodal and unimodal clustering methods, to magnify the disagreements made per constraints on the adopted ground information of various sources for clustering.

In some research works, feature clustering is contemplated to be a mighty substitute to feature selection for reducing the dimensionality of data. A paper by Inderjit, Subramanyam and Rahul [6] focuses on the same by proposing an information-theoretic divisive algorithm for feature/word clustering, that reduces the objective function value/computational cost to capture the optimal feature clustering. The algorithm designed in this paper is claimed to be minimizing the ”within cluster Jensen-Shannon divergence” while simultaneously maximizing the ”between cluster Jensen-Shannon divergence”. From the empirical evidences of the experimental comparisons of the clustering algorithms, it is also claimed that feature clustering is an effective technique for building smaller class models in hierarchical classification. For our case, though we do not look over hierarchical classification techniques to classify our data, but we build our models on the radicality of feature clustering aspects for reducing the dimensionality of data. [6]

This research work aims to provide a comparative study between (1) different clustering techniques, (2) traditional machine learning classifiers such as SVM, Random Forest and Naive Bayes to classify music features. Relative importance of dimensionality reduction and predictability power of the different classifiers for music features are also being discussed in this study.

III. Dataset

In this work, we make use of Spotify dataset of 232,725 tracks encompassing 18 attributes, among which 14 signifies to the features of the song such as popularity, acousticness, danceability, etc,. We pick out random sample subset of 100,000 observations from the dataset for our further study. The dataset was created by scraping audio features for a track by making Web API calls [1] to Spotify and the response being recorded in JSON format, further converted to CSV format and uploaded in Kaggle. [3]

IV. Exploratory Data Analysis

We perform our analysis on Spotify Dataset [3], that comprises of 232,725 observations and 18 characteristics. Data has both numerical and categorical variables, among which we get rid of categorical variables and annotated ones like artist name, track name, track id, popularity, time signature, duration, key, mode and genre, as we think these to be of less significant features for our classification problem. We generate a random sample of 100,000 observations for our analysis, also the columns of newly generated sample dataset has no null/missing values present, and now the dataset has only float values. Acousticness is highly correlated with energy and loudness, and we take care of multicollinearity in data with the employment of a dimensionality reduction tool known as PCA (Principal Component Analysis). PCA reduces these large set of correlated and non-correlated feature set to a smaller less correlated set of features, called principal components, that still retains most of relevant information about the data. [13]

V. Methodology

This section deals with the flow of methodological approaches adopted while performing feature classification of songs, effectuating various classifier algorithms based on clustering overviews on the feature set. To mention the algorithms, we have exerted k-means clustering V-A and model based clustering techniques to tag each observations in a cluster. Bringing in the praxis of PCA, to reduce dimensionality and validate the retainment of the songs' features by inclusion of the principal components, only and mainly to identify most of the variance in the data.

A. ***K-means clustering***

After removing the necessary features that have negligible impact we proceed with the clustering and classification of the feature set. Our primary aim is to find the optimal number of clusters based on which we will perform the classification. We start off with K-means clustering which is a partition-based clustering method where the data points are divided into non-overlapping subsets using cluster centroid. The k-means method involves two basic steps: • Allocation: assigning each observation to the closest cluster centroid. The process is started by generating centroids randomly. • Update: updating the cluster centroids by computing the mean of the points assigned to the corresponding cluster. The steps are repeated until no points are moved between groups.

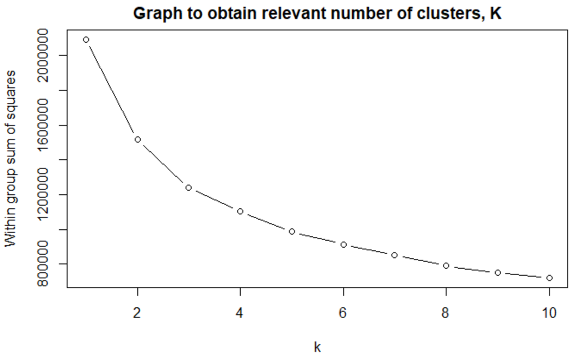


Figure 1: Plot of within group sum of squares to determine optimal no. of clusters by Elbow method

The main aim is to minimise within the cluster sum of squares in such a way that the distance between each observation and the centroid in the cluster is low. We plot Fig. 1 the within group sum of squares and try to figure out the elbow that determines the optimal number of clusters. Though the elbow is not quite distinct, but it seems to be somewhere between two and three. Therefore, we find out the Calinski-Harabasz index to ensure that our assumptions from the previous plot is correct. The Calinski-Harabasz index (CH) is given by:

$$CH = BSS(n-k)/WSS(k-1) \tag{1}$$

where BSS → between sum of squares and WSS → within sum of squares.

A large value of the index indicates low within the cluster variability and large between the cluster variation. We plot the CH index vs the number of clusters and the largest value corresponds to two clusters.

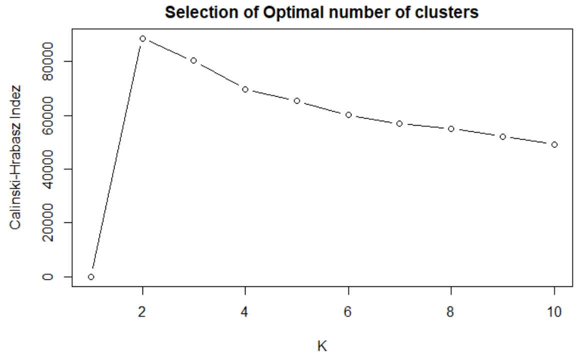


Figure 2: Calinski-Harabasz Index to determine optimal no. of clusters

Though we are almost confident that optimal number of clusters are two, but we wanted to create our model with both two and three clusters to check the accuracy for both and see whether our analysis is correct.

B. *Model based clustering*

In Model based clustering, a probability based approach is embraced to create the clusters rather than a heuristic approach in k-means clustering to assign centroids for clusters, presuming each group data to be represented in the form of Gaussian density function. [9]

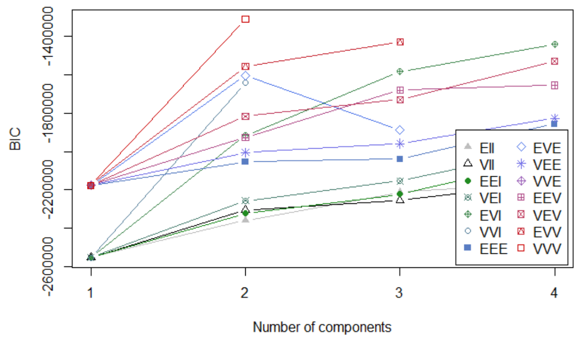
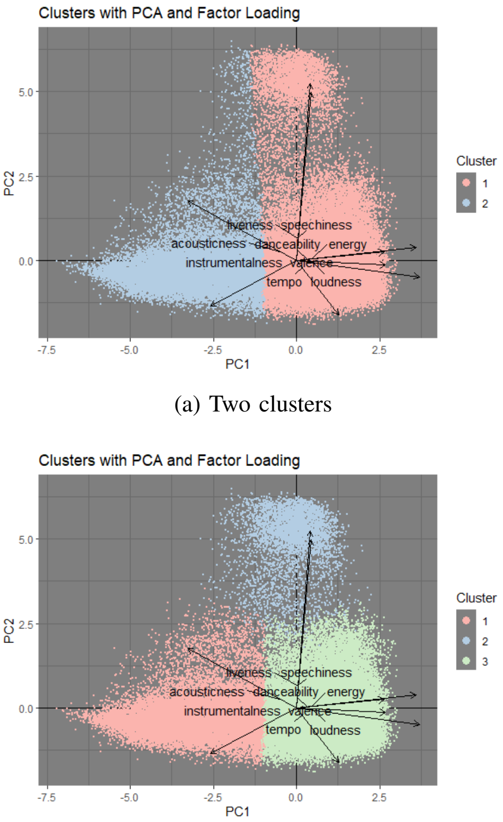


Figure 3: Plot of BIC vs No. of Components in Model Based Clustering

It chooses the best number of components (clusters) and the covariance parameterisation based on the Bayesian Information Criterion (BIC), For our data, model based clustering estimated the best model to be with two components (i.e. clusters) and named the model as VVV, that interprets the two components to be ellipsoidal with varying volume, shape and orientation. [2]

C. *Principle Component Analysis (PCA)*

We use PCA to reduce the dimensionality of our data and extract the relevant and important song features only, that represent as a set of new orthogonal variables (principal components). It also simplifies the description of the dataset and analyse the structure of observations to uncover facts around data and recognize the trends. [4] PCA computes the principal components, procured from the linear combinatorial groups of features that accounts for varied influences in the data (i.e. variation of characteristics in the data). We trace back those influences as shown in the PCA plot Fig. 4 to find out what produces the most difference between the clusters in specific principal components.



(a) Two clusters

(b) Three clusters

Figure 4: PCA plot

Examining the magnitude and direction of coefficients of the original features, we infer that first principal component has large positive associations with acousticness, energy and large negative associations with loudness and valence. While second principal component has positive associations with liveness and negative association with speechiness. The PC6 explains

91.281% of the summative variance, whereas PC7 explains total variance of 95.67% and PC8 explains 98.72%. But we proceed with the first six principal components for our further analysis, as it explains over 90% total variance in the data.

VI. Algorithmic classifiers

PCA explicates the direction of maximum variability, but it does not ensure maximum separation between the classes. So to ensure the maximum separation between the classes, we perform classification on principal components, extracted from the PCA for two and three clusters V-C. In this case, we consider principal components 1,2,3,4,5 and 6 for classification using various algorithms that are debated on further. Though there are many classification algorithms up for grabs now, but concluding on which classification is superior to which is contentious. The performance, complexity and accuracy of an classification algorithm depends on the enactment and nature of the dataset. So to decide upon the best classifiers among Random Forest, Naive Bayes, Support Vector Machine we verify the applicability of those classifiers implementing the most common holdout method. [5] In holdout method, dataset is divided into two sets training set and validation set, training set of data is used to train the model and the validation dataset to qualify the accuracy/ predictive power of the classifier. In the subsequent sections, we further anatomize the interpretations and relevance of the indicated classifiers.

A. *Random Forest*

Random forest classifier is a wise choice as classifier for classification problems where the high performance is a need compared to interpretability. [10] As we have huge dataset of features that being reduced to six principal components, high performance is a must for classification of musical feature components. Another reason for adjunction of Random Forest in our analysis is it's splitting mechanism based on bagged decision trees. These bagged decision trees are further decorrelated with the introduction of splitting of random subset of features than inclusion of all features of the model to classify. Adding on to the averaging of variances along all the trees ensures a low bias and moderate variance model [10]. We train the model as a function of cluster points/classes against all the six principal components extracted from PCA V-C and perform the classification mechanism exerting the Random Forest algorithm. Model is based on two number of variables tried at each spit, which is generally calculated as square root of the no. of features/variables i.e, 7 in our case. Out-of-bag error estimates the prediction error of the model, which is 0.14% for our model for two clusters compared to 0.21% of model for three clusters that well defines around 99.86% and 99.79% accuracy. Out-of-bag error initially drops down as number of tree grows and remains constant after, so we are unable to improve the error after 200 number of trees as shown in Fig.5.

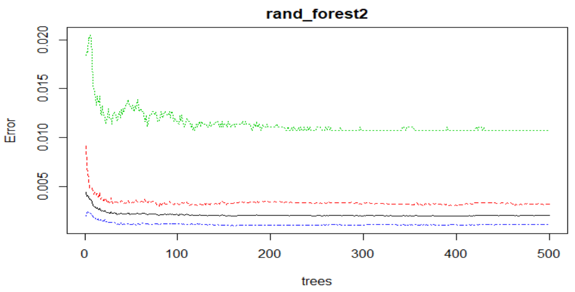
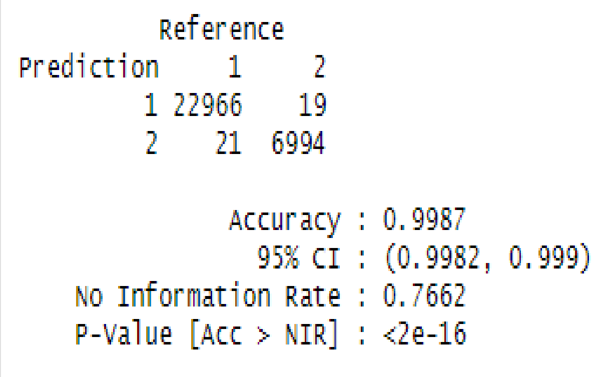


Figure 5: Random Forest error plot

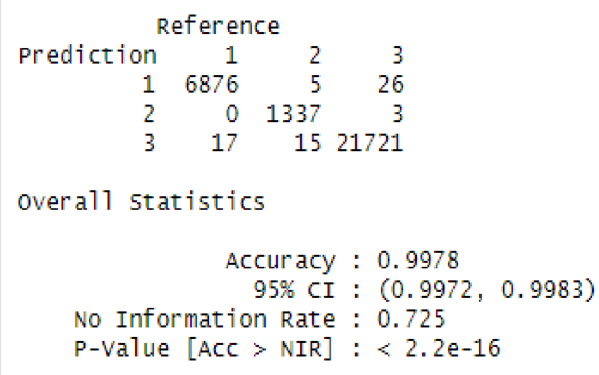
In conformity with the confusion matrix, we analyse that for model of two clusters, predictions are good when predicting for class 1 and for the model of three clusters, predictions are good when predicting for class 1 and class 3 compared to class 2 based on estimation done on out-of-bag-data as shown in Fig.6.



```
          Reference
Prediction     1     2
         1 22966    19
         2    21  6994

               Accuracy : 0.9987
                 95% CI : (0.9982, 0.999)
    No Information Rate : 0.7662
    P-Value [Acc > NIR] : <2e-16
```

(a) Two clusters



```
          Reference
Prediction     1     2     3
         1  6876     5    26
         2     0  1337     3
         3    17    15 21721

Overall Statistics

               Accuracy : 0.9978
                 95% CI : (0.9972, 0.9983)
    No Information Rate : 0.725
    P-Value [Acc > NIR] : < 2.2e-16
```

(b) Three clusters

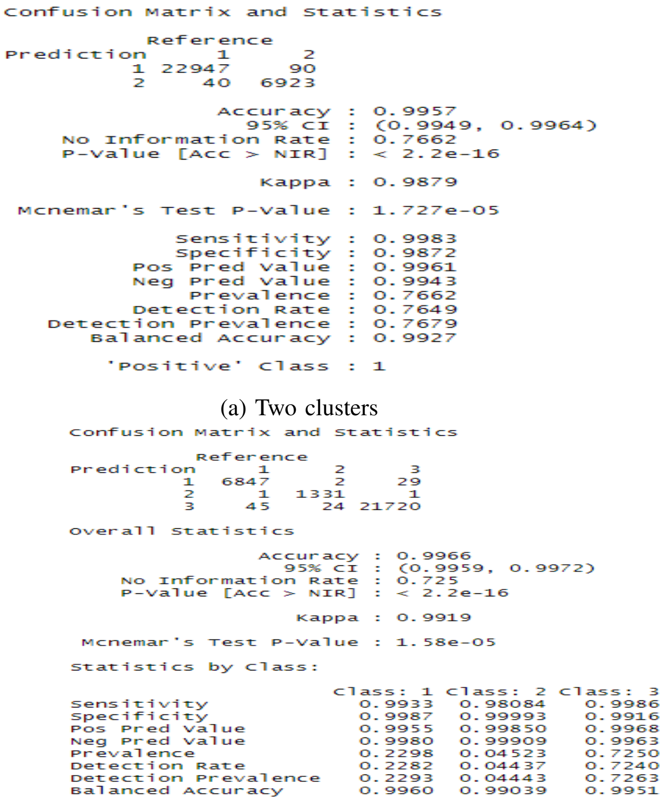
Figure 6: Confusion Matrix and Statistics of Random Forest Algorithm

Now we predict the features using the model on our validation data, and the accuracy for two clusters is 99.87%, where twenty misclassifications happened for class 1 and nineteen misclassifications happened for class 2. For three clusters, accuracy is 99.78%, where seventeen components in class 1 were misclassified as class 3, five components and

15 components in class 2 misclassified as class 1 and class 3, 26 components and 3 components in class 3 misclassified as class 1 and class 2 as shown in Fig.6. To confer on the variable importance in the model, we find that PC1 has the most contribution followed by PC5 against MeanDecreaseGini parameter, that captures how pure the nodes are at the end of tree without the respective principal components.

B. *Support Vector Machine*

SVM works on the concept of finding a hyperplane in an N-dimensional space that distinctly confines and classifies the data points. [8]To effectually find this hyperplane from many possibilities with the maximum margin, the algorithm exerts on maximizing the margin distance, so that the distance between data points of different classes are well apart. We find that prediction scores of SVM are also good in classifying principal components of our clustered musical features. SVM predicts the classes with an accuracy of 99.57% for two clusters and 99.66% for three clusters.



```
Confusion Matrix and Statistics

          Reference
Prediction     1     2
         1 22947    90
         2    40  6923

               Accuracy : 0.9957
                 95% CI : (0.9949, 0.9964)
    No Information Rate : 0.7662
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9879

 Mcnemar's Test P-Value : 1.727e-05

            Sensitivity : 0.9983
            Specificity : 0.9872
         Pos Pred Value : 0.9961
         Neg Pred Value : 0.9943
             Prevalence : 0.7662
         Detection Rate : 0.7649
   Detection Prevalence : 0.7679
      Balanced Accuracy : 0.9927

       'Positive' Class : 1
```

(a) Two clusters

```
Confusion Matrix and Statistics

          Reference
Prediction     1     2     3
         1  6847     2    29
         2     1  1331     1
         3    45    24 21720

Overall Statistics

               Accuracy : 0.9966
                 95% CI : (0.9959, 0.9972)
    No Information Rate : 0.725
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9919

 Mcnemar's Test P-Value : 1.58e-05

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            0.9933  0.98084   0.9986
Specificity            0.9987  0.99993   0.9916
Pos Pred Value         0.9955  0.99850   0.9968
Neg Pred Value         0.9980  0.99909   0.9963
Prevalence             0.2298  0.04523   0.7250
Detection Rate         0.2282  0.04437   0.7240
Detection Prevalence   0.2293  0.04443   0.7263
Balanced Accuracy      0.9960  0.99039   0.9951
```

(b) Three clusters

Figure 7: Confusion Matrix and Statistics of Support Vector Machine

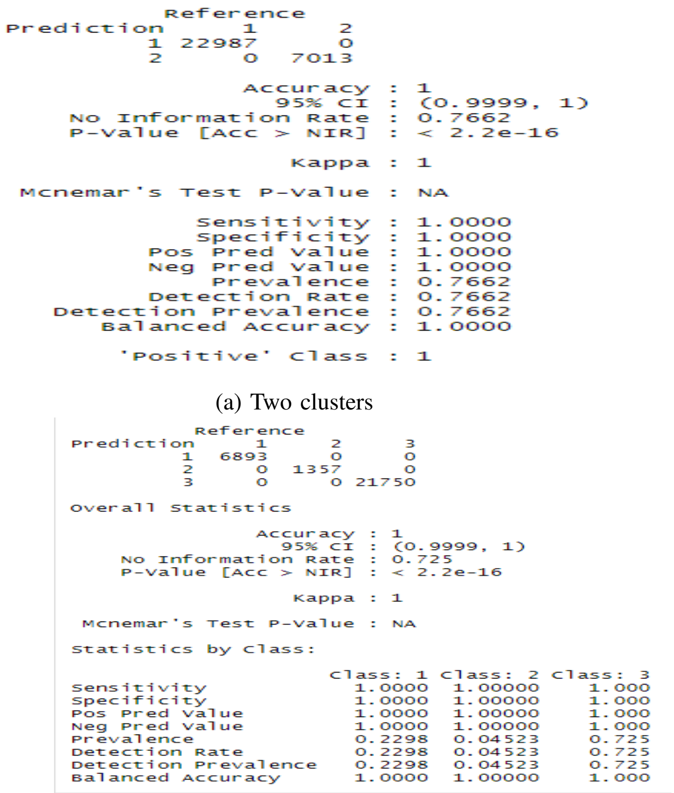
As in Fig. 7 , for two cluster scenario, misclassification in class 1 and class 2 is far more compared to Random Forest for two clusters, accounting forty components misclassified as class 2 and ninety components misclassified as class 1. Sensitivity for three classes are also good (class 1:99.33%, class 2:98.08%, class 3:99.86%) for three cluster classification using SVM. Though the prediction scores of predicting classes are well using SVM, but compared to Random Forest it performs less better as it misclassifies more in numbers in both for two clusters and three clusters scenario.

C. *Naive Bayes classifier*

Naive Bayes classifier is a probabilistic machine learning model, the crux of which is based on Bayes theorem2. [7]

$$P(A|B) = P(B|A)P(A)/P(B) \tag{2}$$

In case of Naive Bayes, the predictors/features i.e, principal components in our case are independent which accounts to better performance by the Naive Bayes classifier compared to other classification algorithms. To extrapolate the claim, we perform training on the model and capture the prediction scores. And the prediction accuracy for Naive Bayes algorithm reflects to be 100% with no misclassification of classes for both two clusters and three clusters as shown in Fig. 8



```
          Reference
Prediction     1     2
         1 22987     0
         2     0  7013

               Accuracy : 1
                 95% CI : (0.9999, 1)
    No Information Rate : 0.7662
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.7662
         Detection Rate : 0.7662
   Detection Prevalence : 0.7662
      Balanced Accuracy : 1.0000

       'Positive' Class : 1
```

(a) Two clusters

```
          Reference
Prediction     1     2     3
         1  6893     0     0
         2     0  1357     0
         3     0     0 21750

Overall Statistics

               Accuracy : 1
                 95% CI : (0.9999, 1)
    No Information Rate : 0.725
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            1.0000  1.00000    1.000
Specificity            1.0000  1.00000    1.000
Pos Pred Value         1.0000  1.00000    1.000
Neg Pred Value         1.0000  1.00000    1.000
Prevalence             0.2298  0.04523    0.725
Detection Rate         0.2298  0.04523    0.725
Detection Prevalence   0.2298  0.04523    0.725
Balanced Accuracy      1.0000  1.00000    1.000
```

(b) Three clusters

Figure 8: Confusion Matrix and Statistics of Naive Bayes Classifier

VII. Division of Work

The entire task has been divided into three specific parts and each part has been taken care-of by one of the group members. The github link for the project is : https://github.com/duttaa2/Spotify-Analysis. The details of the work performed by each individual has been provided below.

A. *Exploratory Data Analysis: (EDA)*

Analyzing the dataset by performing Exploratory Data Analysis defines the flow of the entire project. The analysis of the data involves finding the relationship of the attributes with one another, selecting the attributes which are of importance, checking for multi collinearity and thereby ensuring the model is well fitted. This part has been taken care-of by Parnab Das.

B. *Clustering:*

The K-means clustering has been performed to find the optimal number of clusters that can be generated from the data. Our entire approach has been done using both two and three clusters to check which one performs better. The model based clustering has also been performed to ensure the results of k-means are correct. This part has been taken care-of by Parnab Das.

C. *Principal Component Analysis: (PCA)*

PCA has been done on the data for dimension reduction and thereby extract those features from the data which are of utmost relevance. Out of the 9 principle components, 6 of them accounted for 90 percent of the information. Therefore, we have used these 6 components for further work. This part has been taken care-of by Souradip Goswami.

D. *Random Forest Classifier*

After the clustering has been performed and each data has been classified, the dataset has been divided into train and test for supervised learning. The first algorithm that is being used is the Random Forest which is an ensemble method involving multiple decision trees. This part has been taken care-of by Souradip Goswami.

E. *Naive Bayes Classifier*

The Naive Bayes is another supervised classification algorithm that is being used for training our dataset and then test for accuracy of the model. The Naive Bayes works on the principle of Bayes theorem and conditional variables. This Part has been taken care-of by Aritra Dutta.

F. *Support Vector Machines: SVM*

The SVM works on the principle of creating hyperplanes and classifying data based that on that plane in an n-dimensional space. This is quite a powerful classification algorithm that has been used to classify our data. This part has been taken care-of by Aritra Dutta.

G. *Report*

The Report has been written by every individual based on the task they have performed. The report has been merged using the Overleaf tool.

VIII. Conclusions

We performed a detailed study on the feature classification of Spotify music dataset [3] with the execution of various clustering methods and classification algorithms. The rationale for this extensive study on feature classification and clustering methods is to procure a more generative approach that qualifies and compares all the clustering methods and classifiers adopted in this paper. In pursuance of our discussions above, we conclude that probabilistic approaches in case of clustering (model based clustering techniques) and classification (Naive Bayes classifier) performs better with the utmost accuracy, when evaluated based on various metric methods.

Acknowledgements

References

[1] Get audio features for a track. https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/.

[2] Model based clustering essentials. https://www.datanovia.com/en/lessons/model-based-clustering-essentials/.

[3] Spotify tracks db. https://www.kaggle.com/zaheenhamidani/ultimate-spotify-tracks-db.

[4] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[5] Sidath Asiri. Machine learning classifiers. https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623.

[6] Inderjit S Dhillon, Subramanyam Mallela, and Rahul Kumar. A divisive information-theoretic feature clustering algorithm for text classification. *Journal of machine learning research*, 3(Mar):1265–1287, 2003.

[7] Rohith Gandhi. Naive bayes classifier. https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c.

[8] Rohith Gandhi. Support vector machine — introduction to machine learning algorithms. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47.

[9] Morgun Ivan. Model-based clustering and gaussian mixture model in r. https://en.proft.me/2017/02/1/model-based-clustering-r/.

[10] Julia Kho. Machine learning classifiers. https://towardsdatascience.com/why-random-forest-is-my-favorite-machine-learning-model-b97651fa3706.

[11] Tao Li, Mitsunori Ogihara, and Qi Li. A comparative study on content-based music genre classification. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 282–289, 2003.

[12] Tao Li, Mitsunori Ogihara, Wei Peng, Bo Shao, and Shenghuo Zhu. Music clustering with features from different information sources. *IEEE Transactions on Multimedia*, 11(3):477–485, 2009.

[13] Lexi V. Perez. Principal component analysis to address multicollinearity. https://www.whitman.edu/Documents/Academics/Mathematics/2017/Perez.pdf.

[14] Wikipedia contributors. Free music archive — Wikipedia, the free encyclopedia, 2020. [Online; accessed 9-April-2020].