2021 IEEE 7th International Conference on Big Data Intelligence and Computing (DataCom) | 978-1-6654-2071-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/DataCom53700.2021.00008

Supervised learning and unsupervised learning on music data with different genres.

1st Chunang Liu
liuchunang2021@163.com
Jining, China

2nd Zehan Chao
czh.ucla@viax.org
Beijing, China

***Abstract*—As the digital music datasets evolve fast in the past decades, researchers start to focus on the machine learning tasks on music datasets with music content analysis. Both supervised learning and unsupervised learning have become increasingly popular as searching and recommending are the major tasks for digital music datasets. In this work, we are interested in exploring automatic classification of music with different genres. In this paper, we visualize and embed Spotify music in 3 dimension space with principal component analysis techniques; classifying different music genres via MFCC related algorithms and various machine learning algorithms. We use the method of PCA to determine the relationship between each feature of those songs in the data including 154932 songs and utilize different visualization techniques to obtain insight into the dataset. Thus, we could get the similarity of the songs by analyzing the two clusters of points on the graph. Additionally, we use numerous supervised learning techniques to perform classification on those songs in ten genres and evaluating their performance. Experimental results show the feasibility of automatic management of music databases and the potential to improve.**

***Index Terms*—PCA, music classification, machine learning**

I. Introduction

Nowadays, as the living standard of citizens has increased a lot, some art forms like music have become an integral part of people's lives. Different genres of music have different features and give people distinctive feelings. For instance, rock and roll is full of passion and it could make people excited and energetic while country music is very soothing and relaxing which could make people calm down. These features decide whether the kind of music is popular or not to some degree. Our work focuses on classifying different music genres via MFCC related algorithms.

Whether using labeled datasets is the main distinction between the supervised learning [31] and unsupervised learning [4]. In short, supervised learning uses labeled input and output data while an unsupervised learning algorithm does not. Unsupervised learning is suitable for linear category structures rather than nonlinear category structures and it is proven to be multifaceted in that performance varies with task conditions. Supervised learning models tend to be more accurate than unsupervised learning models yet they require people to label the data appropriately. Intentional unsupervised learning is more regular in comparison to incidental unsupervised learning while their accuracy is almost the same. However, the acquisition and application of intentional unsupervised learning are more laborious than incidental one [20].

Automatic music classification is a fundamental problem for music indexing, content-based music retrieval, music recommendation and online music distribution. Numerous methods have been developed over the years to efficiently classify music information, yet the hurdles remain, for instance, with large music database the warehouses require exhausting and time-consuming work, particularly when categorizing audio genre manually [28].

For classification also, various techniques have been tried by the researchers. Threshold-based techniques [3], [5], Neural Network [12], [2], [7], clustering [25], [17] have been deployed for song classification. Instrumentals are classified according to the type of instruments whereas genre-based classification has been achieved for song [13].

In this work, we aim at providing more comprehensive understanding of relationship between different features of songs based on PCA related methods and classification of different genres of music based on MFCC related methods. The research questions considered are:

- What is the trend of a feature of songs when another feature changes?
- How can the relationship between different features of songs be plotted by different graphs?
- How well does each model of classification perform?
- What are the most significant features of different genres of songs?

In this article, we will use data of ten genres of music including *blues, classical, jazz, disco, pop, hip-hop, rock, reggae, metal* and *country*. We will use the method of PCA. PCA(principal components analysis) is a standard tool in multivariate data analysis to reduce the number of dimensions while retaining as much as possible of the data's variation. PCA investigates the first few components containing the majority of the data's variation are explored rather than those original variables. The visualization and statistical analysis of those variables could help us to find the similarities and

978-1-6654-2071-6/21/$31.00 ©2021 IEEE
DOI 10.1109/DataCom53700.2021.00008

differences between samples and which factors of origin contribute to the new variables mainly [15].

II. Related Work

A. PCA

Nowadays, we are often in the face of huge data with numerical variables, thus, it is pretty hard for us to analyse the data. At this time, we need to use some tools to help us. The principal component analysis is one of them. Principal component analysis (PCA) is a mathematical algorithm that could help us to reduce the dimensionality of the data we want to analyse while retaining most of the variation in the data set. By identifying directions, which are called principal components, it accomplishes this reduction. By using a few components, each sample can be represented by relatively few numbers rather than using all of those variables. We can visually assess similarities and differences between samples and determine whether samples can be grouped by plotting those samples. There are many applications of PCA, for example, applications in computational biology [10], size eduction [14], human face recognition [21] and so on.

B. MFCC

Mel Frequency Cepstral Coefficients (MFCC) are a feature widely used in automatic speech and speaker recognition which were introduced by Davis and Mermelstein in the 1980s [8], and have been state-of-the-art ever since. Methodology of MFCC includes several steps. Initially, we are supposed to frame the signal into short frames, for each frame calculate the periodogram estimate of the power spectrum. Then we need to apply the mel filterbank to the power spectra, sum the energy in each filter. After that, we are expected to take the logarithm of all filterbank energies and take the DCT of the log filterbank energies. In the end, we keep DCT hao 2-13, discard the rest.

There are some related applications of MfCC, for example, MFCC has been used for designing a text dependent speaker identification system. In addition to this, MFCCs are also increasingly finding uses in music information retrieval applications such as genre classification, audio similarity.

MFCCs have both pros and cons. As the frequency bands are positioned logarithmically in MFCC, it approximates the human system response more closely than any other system. Nevertheless, MFCC values are not very robust in the presence of additive noise, and so it is common to normalize their values in speech recognition systems to lessen the influence of noise.

There are also some related or similar methods of MFCCs:

Perceptual Linear Prediction -PLP

This technique is based on the short-term spectrum of speech. It combined several engineering approximations to select characteristics of human hearing and approximates auditory spectra by an autoregressive all-pole model. PLP uses engineering approximations for three basic concepts from the psycho-acoustic of hearing: spectrum critical band spectral resolution, the equal-loudness curve and intensity power low [16]. Like MFCC, PLP employs an auditory based warping of the frequency axis derived from the frequency sensitivity of human hearing.

Linear Predictive Coding - LPC

Linear predictive analysis of speech was introduced in the late 1960s and has become the predominant technique for estimating the basic parameters of speech [22]. Based on a highly simplified model for speech production, LPC could give us both an accurate estimate of the speech parameters such as pitch, formants and spectra. It tries to reproduce the human speech production mechanism. Plus, all the vocal tract parameters are represented in a set of LPC coefficients. The number of coefficients is generally 10 to 20. It is fast, simple and its the capability to extract and store time varying formant information so that it is widely used by people.

C. Music Classification

Unsupervised clustering is widely applied in automatic music classification. It could avoid the constraint of a fixed taxonomy in that clustering data in an un-supervised way can make a classification emerge from the data themselves based on objective similarity measures. This could in turn tackle the problem that the classification suffers from ambiguities and inconsistencies as it has been seen earlier [27].

In the unsupervised approach, an audio title is represented by a large number of features that could be seen in the previous section, and a similarity measure is used to compare titles among one another. It takes the advantage of the similarity measure to organize the music collection with clusters of similar titles [27].

Supervised classification is the other essential way of music classification. It supposes that a taxonomy of genres is given and people try to map a database of songs into it by machine learning algorithms. For this method, labeled data is needed and it attempts to form automatically relationships between the features of the training set and the related categories [27].

The supervised approach to music genre classification has been studied extensively. Here are a large number of commonly used supervised machine learning algorithms.

Thresholding is the simplest method of image segmentation. In our music classification, it provides a measure of the certainty or uncertainty of a prediction. It also provides additional granularity over just predicting the class label that can be interpreted. From a gray scale image, binary images can be created by Thresholding. Binary images are produced from color images by segmentation. Segmentation is the process of assigning each pixel in the source image to two or more

classes. The usual result is several binary images when there are more than two classes. In image processing, Thresholding is used to split an image into smaller segments, or junks, using at least one color or gray scale value to define their boundary. The benefits of obtaining a binary image are that it reduces the complexity of the data and simplifies the process of recognition and classification [3].

Neural networks could help us cluster and classify. We could think of them as clustering and classification layers on top of the data we store and manage, and when they have a labeled dataset to train on, they will classify the data. The inspiration of Neural Network is a simple model of how the brain works in nature: a neuron "fires" if the stimuli received from other neurons exceed a certain threshold. Network training can be understood as the process of minimising a loss function by iteratively adjusting the weights such that the deviation of the actual network output from the desired output is minimised. Popular choices for the loss function are the sum of quadratic deviations or a measure of the entropy [12].

Neural networks are superior to other methods in those correlations between the input variables can be learned by them, and they can also incorporate information from quality variables (e.g. the return code of a certain algorithm) and there is no need for input variables to be filled for each event [12].

Clustering is a type of unsupervised machine learning which aims to find ho- mogeneous subgroups such that objects in the same clusters are more similar to each other than the others. The purpose of clustering is descriptive, that of clas- sification is predictive (Veyssieres and Plant, 1998). Since the goal of clustering is to discover a new set of categories, the new groups are of interest in themselves, and their assessment is constitutional. In classification tasks, nevertheless, a vital part of the assessment is extrinsic, since the groups must reflect some reference set of classes [25].

Clustering of objects is as ancient as the human need for describing the salient characteristics of men and objects and identifying them with a type. Thus, it embraces a variety of scientific disciplines: from mathematics and statistics to biology and genetics, each of which uses different terms to describe the topologies formed using this analysis. From biological 'taxonomies', to medical 'syndromes' and genetic 'genotypes' to manufacturing "group technology" - the problem is identical: forming categories of entities and assigning individuals to the correct clusters within it [25].

III. PCA and Matrix Factorization

Principal Component Analysis (PCA) is an unsupervised learning technique and it is usually used to reduce the dimension of the input high-dimension data while minimizing the loss of information. Aside from eigenvector-based factorization (SVD), non-negative matrix factorization (NMF) has many desirable properties. It is recognized that NMF provides a continuous non-negative solution to the K-means clustering and also a solution to the spectral clustering [11]. In general, non-negative Matrix Factorization seeks to decompose the original matrix into two lower-dimension matrices but restrict them to be non-negative:

$$\min_{W,H} ||A - W \cdot H||_F \qquad \text{s.t. } W \in \mathbb{R}^{m\times r}_{\geq 0}, H \in \mathbb{R}^{r\times n}_{\geq 0} \quad (1)$$

Here we focus on the Frobenius norm approximation and many other losses could be defined to approximate A with W and H. This method is broadly used when the dataset, A, comes as non-negative signals. The applications include but are not restricted to topic modeling, image processing [23], and even astronomy [24].

Although this minimization problem is not convex for (W, H) at the same time, the objective is convex for each variable when fixing the other. Therefore an approximation of the optimal (W, H) could be found using alternating direction minimization(ADMM). An iterative algorithm is provided in Lee, Seung's work [19].

$$h^{t+1}_{i,j} = h^t_{i,j} \frac{(AH^T)_{i,j}}{(W^T WH)_{i,j}} \quad (2)$$

$$w^{t+1}_{i,j} = w^t_{i,j} \frac{(AH^T)_{i,j}}{(WHH^T)_{i,j}} \quad (3)$$

where $h_{i,j}$ stands for the (i, j) entry of H and $w_{i,j}$ for the (i, j) entry of W. The iterating algorithm above is also proven to be convergent hence guaranteed to find at least locally optimal solutions of Problem 1.

Now we have the approximation $A \approx WH$, and for each column of A, we have:

$$\mathbf{A_i} \approx \sum_{j=1}^{k} \mathbf{w}_j h_{j,i} \quad (4)$$

In the equation above, $\mathbf{w}_j$ stands for the j^{th} column of W, Thus, each vector $\mathbf{x_i}$ is approximated by a linear combination of the top music features, W, providing a dimension-reduced approximation of the original music-features matrix. The main difference between NMF and other matrix factorization problems is that there are no negative entries in the factorized components, which means only additive combinations among different bases are applied. Therefore, NMF is also believed to be capable of learning a parts-based representation [18].

IV. Experiments

A. Music PCA

To start the experiment, we obtain a music database consisting of 154932 songs with 16 features including *acousticness, danceability, duration, energy, instrumentalness, key, liveness,*

loudness, mode, speechiness, tempo, time signature, valence, target, song title, and artist.Then we use the method of PCA to reduce the demonsionality, and visually assess similarities and differences between samples and determine whether samples can be grouped by plotting those samples. Plus, we also collect a music data contains 2994 songs in ten musical genres. We use the method of MFCC to classify and analyse them. Both supervised and unsupervised classification is used in the experiment and different methods lead to different results.

The paper is going to find the relationship between different features of a song. We collect the data named $songAttributes\_1999-2019.csv$ in December 2020 from Kaggle [9]. The dataset contains 154932 songs with different features from 1999 to 2019. The data illustrates information of different features of songs, including *acousticness, danceability, duration, energy, instrumentalness, key, liveness, loudness, mode, speechiness, tempo, time_signature, valence, target, song_title*, and *artist*. We only use the numerical features rather than the features of words such as artist and song titles.

We applied the pandas Dataframe package to process the collected data. We use the model of linear regression to plot a graph between any two features and we could see the relationship between them directly. For example, we can use the value of *danceability* as X variable and use the value of valence as Y variable. When we put this into the procedure, we could get a line that is upward sloping, which means when the *danceability* of a song increases, the valence of the song will also increase. After that, we use histograms to analyse the relationship between different features of a song so that we could get a tendency of a feature when the other one changes. Finally, we use the method of PCA to reduce the dimensionality of these 16 features and get a three dimension graph. Then, we could see the relationship between the three variables directly.

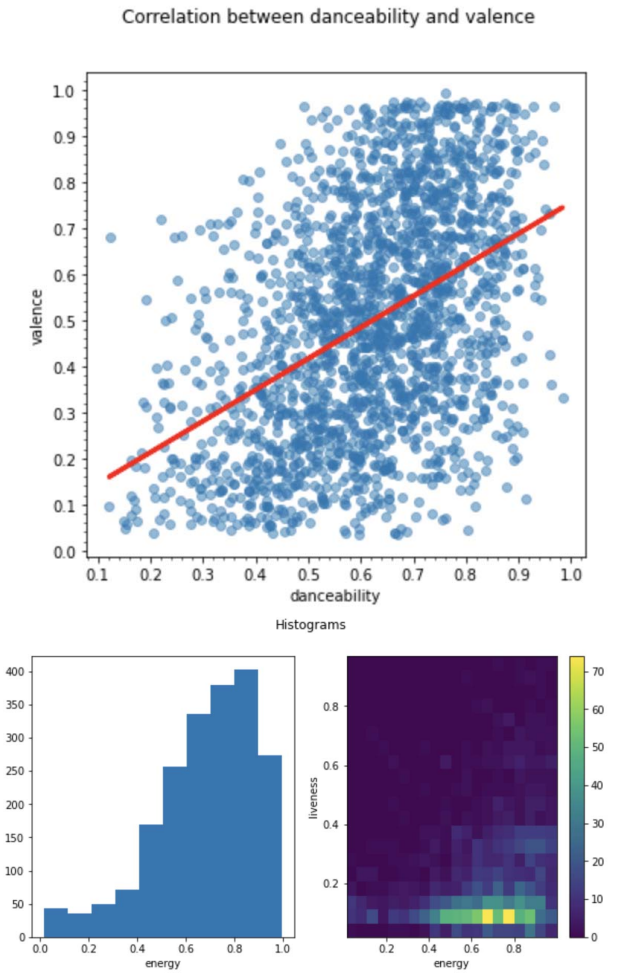

Fig. 1: the first plot is Linear Regression, which illustrates the relationship between danceability and valence; the second plot is the histogram of energy, it illustrates the distribution of energy; the third plot is a heatmap, it illustrates the magnitude of Liveness under the condition of different energy

B. Music Classification

Apart from using PCA to analyze features of a song, we also use the method of MFCC to classify music genres. We collect ten genres of music including *blues, classical, jazz, disco, pop, hip-hop, rock, reggae, metal* and *country*. After we collect the data of those ten genres of music, we import that music which have the same rate and each music contains the music signal as an array of length 661794. The raw signal is too large and unprocessed as being applied as input data. So we apply the MFCC technique to extract the intensity metric from each small time window of the music. With MFCC we are able to convert the sound signals into signal intensity, with size 2994. Next, we apply different classification tools to the signal intensity, and analyze their performance. We have numerical models for music classification such as a nearest neighbour, decision tree, neural net, random forest and so on.

Different models have different ways to classify music into different genres. We record the performance of each algorithm. Precision-Recall is a useful measure of the success of prediction when the classes are very imbalanced. Precision attempts to answer the question that what proportion of positive identifications was actually correct while recall attempts to answer the question that what proportion of actual positives was identified correctly. F1 score is defined as the harmonic mean of precision and recall and it is a measure of a model's accuracy on a dataset.

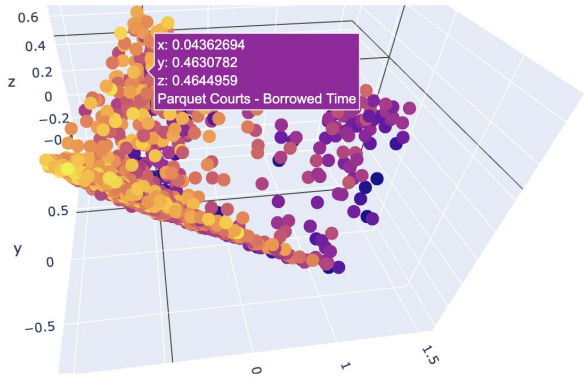

Fig. 2: The result of PCA: two clusters of points with different colours represent two groups of music while each group contains music with similar properties

	Precision	Recall	F1 Score	Accuracy
SVM [29]	0.599	0.603	0.595	0.610
GPC [30]	0.390	0.129	0.0796	0.135
ANN [1]	0.627	0.626	0.620	0.630
KNN [6]	0.633	0.618	0.610	0.620
DT [26]	0.393	0.397	0.392	0.400

TABLE I: Performance of different task.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1score = \frac{2 * (Recall * Precision)}{Recall + Precision} \tag{7}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{8}$$

A true positive is an outcome where the model correctly predicts the positive class, eg. classifying *hiphop* into the genre of *hiphop*. Similarly, a true negative is an outcome where the model correctly predicts the negative class, eg. classifying *jazz* into the *jazz* genre. A false positive is an outcome where the model incorrectly predicts the positive class, eg. classifying *jazz* music into the genre of *hiphop*. A false negative refers to an incorrect indication that something is not present when it really is, eg. classifying *hip-hop* music into the genre of *jazz*. We apply several common machine learning algorithms to the music classification and record their performance in terms of precision, recall, F1 score, and accuracy. The table below summarizes the scores.

V. Discussion

In this paper, we visualize and embed Spotify music in 3 dimension space with principal component analysis techniques; classifying different music genres via MFCC related algorithms and various machine learning algorithms. During the process of the experiment, we are able to find the advantages and disadvantages of those classification models. SVM is more effective in high dimensional spaces while it is not suitable for large data sets, and SVM works relatively well when there is a clear margin of separation between classes. In our experiment, the prediction scores are about 0.6, this is a relatively precise score in comparison with other models' results. However, it does not perform very well when the data set has more noise i.e. target classes are overlapping. In comparison to other models, decision trees require less effort for data preparation during pre-processing: A decision tree does not require normalization of data and does not require scaling of data as well. However, the Decision Tree algorithm is inadequate for applying regression and predicting continuous values. In the experiment, the prediction scores of it all do not exceed 0.4. The gaussian process directly captures the model uncertainty. As an example, in regression, GPC directly gives you a distribution for the prediction value, rather than just one value as the prediction. Nevertheless, they lose efficiency in high dimensional spaces – namely when the number of features exceeds a few dozens. GPC in our experiment has a bad performance, F1 score of it even as low as 0.0796. Nearest Neighbor does not learn anything in the training period. It does not derive any discriminative function from the training data. In other words, there is no training period for it. It stores the training dataset and learns from it only at the time of making real-time predictions. This makes the KNN algorithm much faster than other algorithms that require training. Also, new data can be added seamlessly which will not impact the accuracy of the algorithm. On the other hand, KNN does not work well with the large dataset: In large datasets, the cost of calculating the distance between the new point and each existing point is huge which degrades the performance of the algorithm. Plus, it does not work well with high dimensions: The KNN algorithm doesn't work well with high dimensional data because with a large number of dimensions, it becomes difficult for the algorithm to calculate the distance in each dimension. As for MLP Classifier, it can be applied to complex non-linear problems and works well with large input data, and same accuracy ratio could be achieved even with smaller data. It also has drawbacks, computation of ANN are difficult and time-consuming, and the proper functioning of the model depends on the quality of training. Both ANN and KNN in the experiment have scored around 0.62, which are great performances.

We also have something for improvement. We could increase the number of models used in our experiment especially those models which could bring about more precise result. We could modulate the exponent of our models to find a more appropriate figure of probability. In this way, we could improve the classification performance of the models used in our experiment.

REFERENCES

[1] James A Anderson, Edward Rosenfeld, and Andras Pellionisz. *Neurocomputing*, volume 2. MIT press, 1988.

[2] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

[3] K Bhargavi and S Jyothi. A survey on threshold based segmentation technique in image processing. *International Journal of Innovative Research and Development*, 3(12):234–239, 2014.

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

[5] Alexander Chudik, Kamiar Mohaddes, M Hashem Pesaran, Mehdi Raissi, and Alessandro Rebucci. A counterfactual economic analysis of covid-19 using a threshold augmented multi-country model. *Journal of International Money and Finance*, page 102477, 2021.

[6] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[7] Yasuhiro Date and Jun Kikuchi. Application of a deep neural network to metabolomics studies and its performance in determining important variables. *Analytical chemistry*, 90(3):1805–1810, 2018.

[8] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

[9] Daniel DeFoe. Data on songs from billboard 1999-2019, 2020.

[10] Karthik Devarajan. Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS computational biology*, 4(7):e1000029, 2008.

[11] Chris Ding and Xiaofeng He. Principal component analysis and effective k-means clustering. In *Proceedings of the 2004 SIAM International Conference on Data Mining*, pages 497–501. SIAM, 2004.

[12] M Feindt and U Kerzel. The neurobayes neural network package. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 559(1):190–194, 2006.

[13] Arijit Ghosal, Rudrasis Chakraborty, Bibhas Chandra Dhara, and Sanjoy Kumar Saha. Music classification based on mfcc variants and amplitude variation pattern: a hierarchical approach. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 5(1):131–150, 2012.

[14] Richard P Good, Daniel Kost, and Gregory A Cherry. Introducing a unified pca algorithm for model size reduction. *IEEE Transactions on Semiconductor Manufacturing*, 23(2):201–209, 2010.

[15] Detlef Groth, Stefanie Hartmann, Sebastian Klie, and Joachim Selbig. Principal components analysis. In *Computational Toxicology*, pages 527–547. Springer, 2013.

[16] Hynek Hermansky. Perceptual linear predictive (plp) analysis of speech. *the Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[17] Md Rezaul Karim, Oya Beyan, Achille Zappa, Ivan G Costa, Dietrich Rebholz-Schuhmann, Michael Cochez, and Stefan Decker. Deep learning-based clustering approaches for bioinformatics. *Briefings in Bioinformatics*, 22(1):393–415, 2021.

[18] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[19] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[20] Bradley C Love. Comparing supervised and unsupervised category learning. *Psychonomic bulletin & review*, 9(4):829–835, 2002.

[21] Firoz Mahmud, Md Enamul Haque, Syed Tauhid Zuhori, and Biprodip Pal. Human face recognition using pca based genetic algorithm. In *2014 International Conference on Electrical Engineering and Information & Communication Technology*, pages 1–5. IEEE, 2014.

[22] FW Mounts. A video encoding system with conditional picture-element replenishment. *Bell System Technical Journal*, 48(7):2545–2554, 1969.

[23] Bin Ren, Laurent Pueyo, Christine Chen, Élodie Choquet, John H Debes, Gaspard Duchêne, François Ménard, and Marshall D Perrin. Using data imputation for signal separation in high-contrast imaging. *The Astrophysical Journal*, 892(2):74, 2020.

[24] Bin Ren, Laurent Pueyo, Guangtun Ben Zhu, John Debes, and Gaspard Duchêne. Non-negative matrix factorization: robust extraction of extended structures. *The Astrophysical Journal*, 852(2):104, 2018.

[25] Lior Rokach and Oded Maimon. Clustering methods. In *Data mining and knowledge discovery handbook*, pages 321–352. Springer, 2005.

[26] S Rasoul Safavian and David Landgrebe. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674, 1991.

[27] Nicolas Scaringella, Giorgio Zoia, and Daniel Mlynek. Automatic genre classification of music content: a survey. *IEEE Signal Processing Magazine*, 23(2):133–141, 2006.

[28] R Thiruvengatanadhan. Music classification using mfcc and svm. *International Research Journal of Engineering and Technology*, 5:922–924, 2018.

[29] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science; business media, 2013.

[30] Christopher KI Williams and Carl Edward Rasmussen. Gaussian processes for regression. 1996.

[31] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018.