Feature extraction and genre-classification using customized kernel for Music information retrieval

Mr. Karthik V[a], Dr. Savita Choudhary[b]
[a]Assistant Professor, Department of Information Science and Engineering, Ramaiah Institute of Technology, Bengaluru, India
[b] Associate Professor, Department of Computer Science and Engineering,Sir M. Visvesvaraya Institute of Technology, Bengaluru, India

AbstractMusic feature extraction and genres form a natural way to consolidate audio and they share related rhythm and texture. We will be building a customizedfeature extraction genre classification model using customized kernel in support vector machine that will use features representing timbre, rhythmic and pitch analysis of the audio. We train various classifiers like k-Nearest neighbor, Support vector machine, Logistic Regression, Neural Network on the GTZAN dataset provided by MARYSAS. We are able to get good accuracy using Customized kernel and ensemble voting classifier and support vector machine on both 10-genre and 4-genre classification.
Keywords: Genreclassification, Support vector machine, feature extraction GTZAN dataset.

1. Introduction

With the advent of digital music, it has become very important to group music files for various tasks like search-retrieval and recommender systems. Manual annotation of such a huge dataset is an impossible task, and hence" Automatic Music Genre Classification" has been a widely studied research topic in the field of Multimedia Information Retrieval (MIR).Automated Music Genre Classification has been studied by numerous researchers and still remains a challenging topic in Music Information Retrieval (MIR) community due to the fuzzy nature of features and the ambiguity associated with human perception of genres. Ground breaking work in this field was performed by Tzanetakis, etc. al in [1], by using acoustic features through audio analysis done on a dataset consisting of 1000 audio files. This dataset, now provided by MARSYAS, has been widely used in approaching this problem. This work has also proposed various content-based features which we are using in our approach as well.

Materials and Methods

1. Dataset: The 1000 songs are considered for feature extraction, exploring and exploiting the different methods of audio data. The labeled data in the range of [-1,1] and sampling frequency average is considered as 4.2Hz.the work carried out with the dataset of GTZAN from MARSYAS audio data is distributed into several genres like hip hop, classical, pop, metal, rock, reggae, blues, disco, jazz and country. File length is 30seconds and 22050 Hz and it is 16-bit sample and used 67% for training and 33% for testing.

2. Music data processing for machine Learning: (i) Time domain features, (ii) Frequency domain features, (iii) Time-Frequency domain features. In Time domain features analog signals are edited and manipulated by computers. Analog digital conversion process sample and quantize the analog songs to get digital signals, once the digitized process is done next step is to do with framing and bundle together with a bunch of samples, for example frame1: sample1=128 frames are overlapped, and it is perceivable audio chunk 1 sample = 44.1khz = 0.0227ms, human can receive sound with 1 sample <10ms, duration of frame(df) = 1/sampling rate * total number of samples in frame. Sample rate is nothing but duration of single sample. 1/44100*512 = 11.6ms.

   Next step is to compute features then to aggregate using mean, median, Gaussian mixture model were we get feature value/vector/matrix and these are the snapshots for the complete duration of audio signals and pipeline include Figure 1.

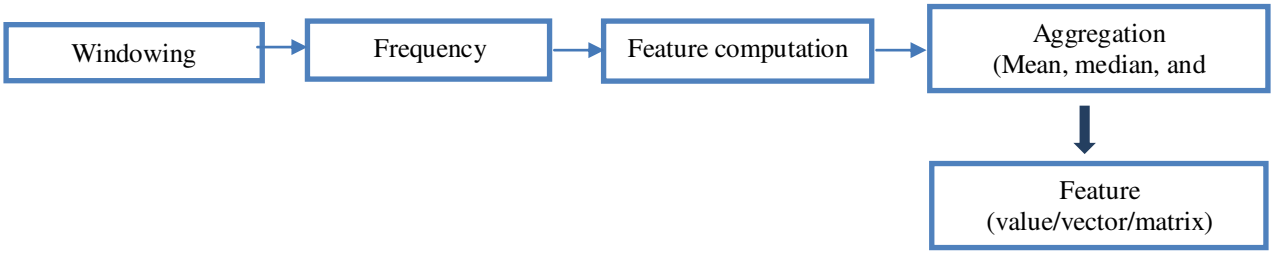


Figure 1. pipeline for feature extraction method

We extracted the features corresponding to three domains that answer different characteristics of an audio file. These are: Timbre, Rhythm and Dynamic Pitch. Since the audio file is continuous, we needed to segment each audio into shorter clip to get the changes in characteristics over time. Since the texture or rhythm of a song is prevalent for a short duration, our choice of segmentation frame had to be small. We chose to split the audio into 30 sec analysis window that will be classified into the different genres. Each analysis window is further divided into frames of 250 milliseconds. Feature computation takes place at the frame level firstly. We used various content-based features as suggested by [1] and added other combinations of features to improve the accuracy.

Time Domain

The extraction of features is done in either time domain or frequency domain analysis. In Time Domain any audio file is stored as time varying waveform where we have discrete samples representing the signal. For a file having sampling rate of 22050Hz, we store 22050 samples,each sample being 16-bit integer. The time domain analysis leads to various features like amplitude of the signal, the mean amplitude(energy) or the zero-crossing rate.

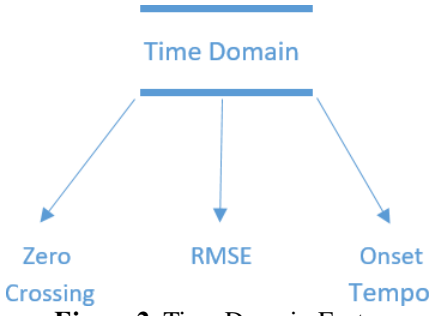


Figure 2. Time Domain Features

Frequency/Spectral Domain

Apart from the time domain representation, the audio file can also be represented as sum of sinusoidal waves of different frequencies. Such a representation where we plot the frequency intensity over the entire bandwidth is called the spectral domain representation of the audio file. To get the frequency spectrum of any signal, we use Fast Fourier Transform and get the coefficients corresponding to each frequency bin. Using the frequency spectrum, we can analyze the primary frequencies of any signal. We use features like spectral centroid, spectral roll off in the frequency domain analysis.

2. There are three kinds of features that are used for Music Genre Classification.

a. Timbral Features

Timbral texture features give the changes in frequency spectrum in a signal over time. When the music is a mixture of sounds, the changes in texture is a very useful differentiating factor. Such features are widely used in tasks like Speech-Music dis- crimination, speech recognition. When the rhythm and pitch is not sufficient to classify the music file, the timbral features are observed to be the key feature. Percussion instruments are generally having differentiating timbral features. In order to compute the timbral features, we first take the short time Fouriertransform (STFT). Before we understand the timbral features in detail, let us look at the basics of short time Fourier transform.

Short time Fourier transform [23], [29] is applied on a continuous signal by first breaking them into shorter signals like frames. To maintain the continuity, generally the window function is overlapped over one another. Convoluting with some window function like Hamming Window or Hanning window tends to reduce the noise after the Fourier transform. The equation for short time Fourier transform is as follows:

$$\text{Short Term Fourier transform}\{x(a)\}(\tau,\omega) = X(\tau,\omega) = \int_{-\infty}^{\infty} \boldsymbol{x(a)w(a-\tau)e^{-iwa}}\text{dt}$$

Where x(a) is the time varying audio signal, w(a) is the window function like Hanning/Hamming window X (τ, ω) is the frequency domain signal and spectrogram(x(a)) (τ,ω) =|X(τ,ω)|^2

b. Mel-Frequency Cepstral Coefficient (MFCC)

Mel-Frequency cepstral coefficients are the most widely used timbral audio feature and finds its use in almost all speech recognition and speech classification problems. In order to compute MFCCs, first, the signal is divided into multiple short frames. For each frame, the short time Fourier transform is computed. The frequency spectrum is then plotted on a Mel-scale that consists of 2 filters - linear up to frequency range of 1KHz and logarithmic after that.

c. Spectral Centroid

Spectral centroid is a way to measure the brightness of a sound. It is the mean frequency component in the frequency spectrum of any audio signal. For any frame of 250ms, we get the Spectral centroid using below equation: Spectral Centroid = ∑n = ∑Mn=1kF[M] k=1F[M] Where F[M] is the amplitude corresponding to the Mth frequency bin.

d. Rhythmic Features

Rhythm is a very natural way for humans to identify genres among different music files. The rhythm could be estimated through various measures like beat, tempo, rubato and gives us the periodicity of the frequency components for any music file. In order to take the rhythmic characteristics for our classification model, we estimated the onset tempo. - Estimated Tempo: The tempo of any music file is estimated based on the average beat per minute of the most dominant frequency in that file. First the onset autocorrelation is plotted for the entire duration and the first periodicity, thus observed is marked asthe estimated tempo of the track

e. Pitch/Dynamic Features

Pitch and other dynamic features for any audio file is useful in identifying variations in the sound over time. Features like zero crossing rate and energy tend to indicate the noisiness or the entropy in the signal. Short time energy is a time domain feature computed by taking the mean square of amplitudes of the samples within a frame.

2.1. Low Energy: Another way to interpret the short time energy over the entire sample is to compute the low energy for the music file. As discussed above, the short time energy tends to give the mean energy of an individual frame. To aggregate this for the entire sample, low energy is defined as the percentage of frames that have RMS energy less than the mean RMS energy of all the frames.

2.2. Zero Crossing Rate: As evident from the name itself, zero crossing rate of a signal is the number of times the signal changes sign from positive to negative or vice versa. We get the zero-crossing rate for any frame as per below equation. Although the pitch related features are predominantly used for distinguishing music against speech content, the aggregation of these features with others have shown to improve the results of music genre classification.

3. Feature Aggregation

In order to aggregate the features for the entire window of 30 seconds, we used the gaussian mixture assumption as per [4]. Each feature is estimated as random variable from a Gaussian mixture. Hence, mean and variance are the typical values that are used for each feature. For representing the Mel-Frequency cepstral coefficients, we had to use 12 coefficients for each frame. And we had 1200 frames for the entire window, which resulted in a very huge feature dimensionality. With our aggregation technique, now we represent MFCCs as 12 mean MFCCs and 12*12 covariance matrix for the different coefficients resulting in 156 features.

the features for our classification task. In this section, we explain the complete pipeline from input music file to output genre label that we obtain for the given dataset. We can divide our method into Window selection, Segmentation and noise reduction,Feature computation and aggregation by frames,Model creation and training on feature vectors, and Genre classification on trained model. The Workflow diagram for the above method is shown in Figure 3.

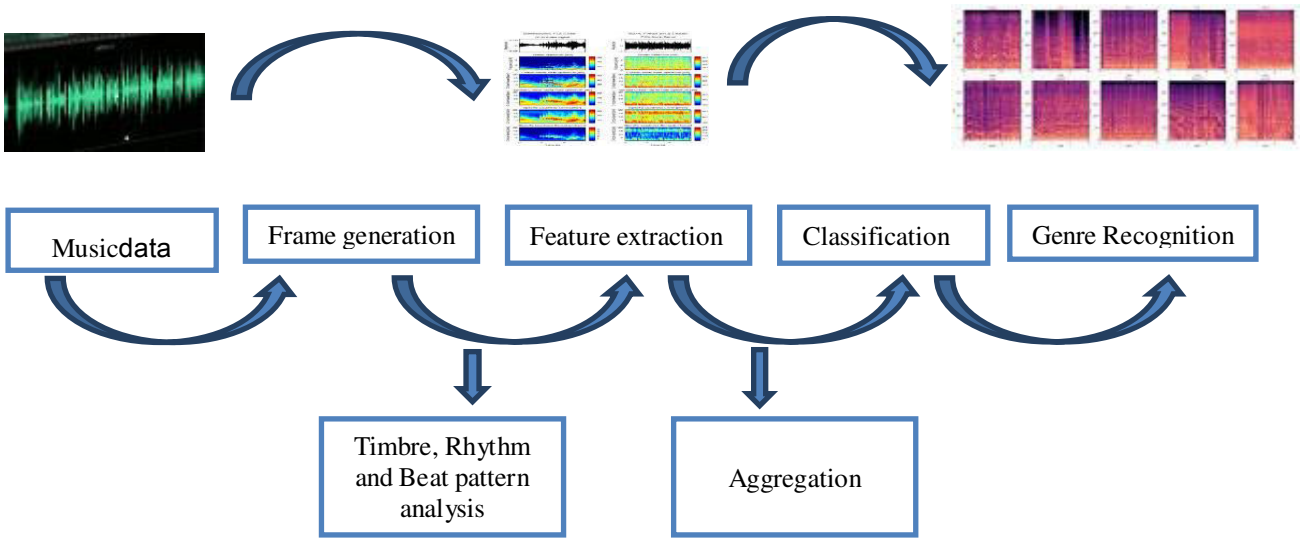


Figure 3. Workflow Diagram

3.1. Window Selection

The first part of the feature extraction process is to select audio window to extract its features. We fix our length of audio to be 30 seconds. This helps us give a uniform length for all audio tracks and avoid biasing the audio towards a particular genre by virtue of it being longer or shorter compared to other tracks.

3.2. Segmentation and Noise Reduction

Once we extract the 30 seconds audio, we then separate the frames from them. We segment each audio into frames of 250 milliseconds. The overlapping frames are segmented using hamming window over 100 milliseconds which is use to smoothen the noise. Hamming window [9] is a typical sliding window used in signal theory to smoothen the energy distribution of a signal and overcome the noisy variations in the signal.

3.3. Feature's computation and aggregation by frames

Now we have our frames, we apply the methods mentioned in Section 5, and get the features out of the audio frame-by-frame. After extraction, we aggregate the features for the entire 30 second window by frames, using a Gaussian assumption. The assumptions are that every feature is a Gaussian random variable picked for different frames.

Thus, we obtain the following mentioned different types of features.

a) New features - MFCC (12 mean and 12*12 covariance matrix)
b) spectral centroid (mean and standard deviation)
c) spectral roll-off (mean and standard deviation)
d) onset tempo (mean estimated tempo)
e) zero crossing rate (mean and standard deviation)
f) short time energy
g) low energy mean

3.4. Model creation and training on feature vectors

Using the above process, we obtain a 164-dimensional feature vector which we then use to train our various classification models. A number of classification techniques have been studied in literature for various different multi-modal data. For our given data and for our particular task, we studied and found out the classifiers which has proven to be a step ahead than all other types of classifiers. Here, we describe each of the classifiers and their training procedure for model creation.

1. Support Vector Machine

Support Vector Machine is a supervised learning technique in which the goal is to come up with a single decision boundary to classify all the data points in to the given classes. Given a set of labeled data points, in the training process the goal is to come up with a good enough decision boundary which will minimize the error in classification of data points.

2. Support Vector Machine Formulation

Support vector machine works by estimating the support vectors (Figure 4) which are the data points closest to the decision hyperplane, separating the classes. The error is calculated by summing the distance of the mis-classified examples from the decision hyperplane. Since ours is a multi-class labeling problem, and support vector machine finds hyperplane for classifying two classes. Thus, we come up with two approaches for converting the two class SVM classification to multi-class classification.

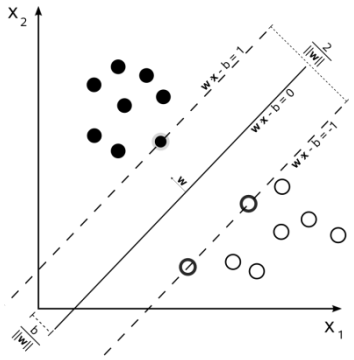


Figure 4: Support Vector Machine classification

4. Results and discussion

a. Model Classification for 10 Genres

Referring to our dataset we applied our models to 10 genres classification namely blues, classical, country, disco, hip-hop, metal, jazz, pop, reggae and rock. From observation we found out that support vector machine outperformed every other classification model in our experiment with 66.48% accuracy. Accuracy for all models is shown in Table 1.

Table 1. 10 Genre Accuracy

10 Genre Accuracy	
Models	Accuracy
Support Vector Machine	66.48%
Ensemble	62.48%
Logistic Regression	59.39%
Neural Network	47.81%
K-Nearest Neighbor	47.51%
Naive Bayes	47.14%

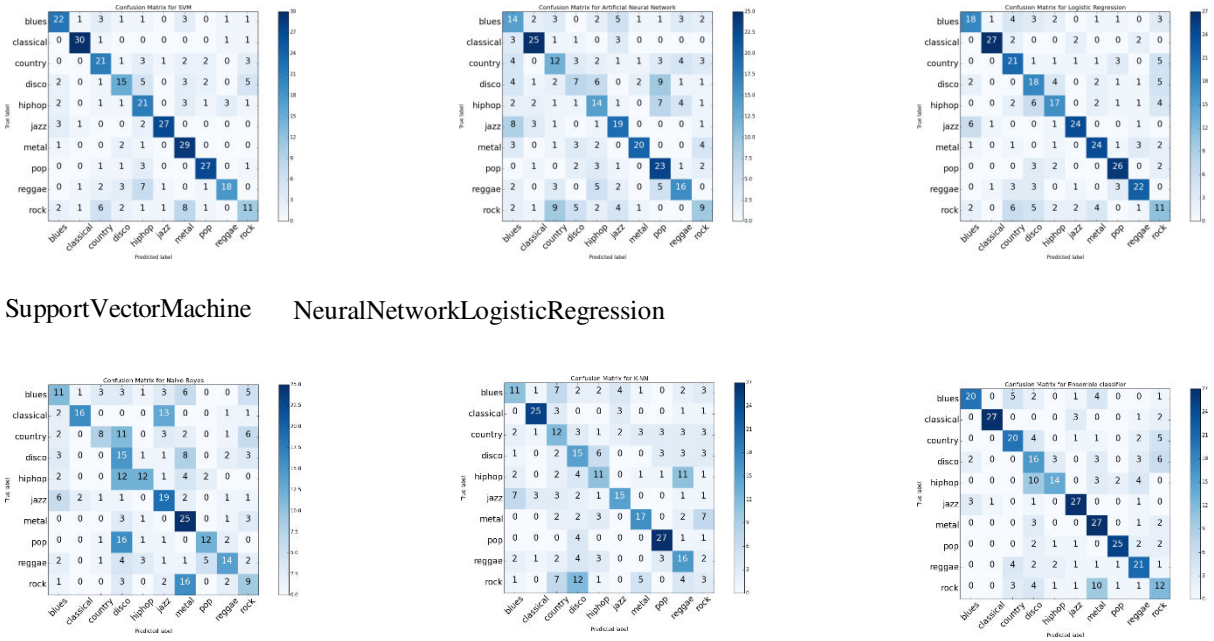


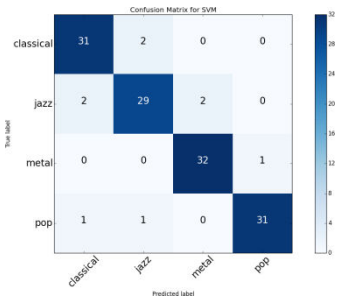
Figure 5: 10 Genre Classification

1. Model Classification for 4 Genres

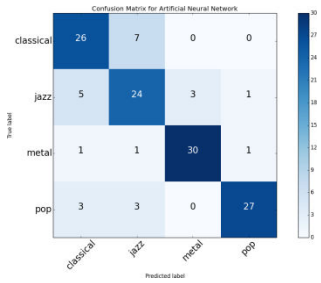
The four-genre classification classify audio in classical, jazz, metal and pop genres. The accuracy of this classification was very high and we achieved close to 90 percent accuracy in SVM. Other models also performed really well in 4 genre classifications. The results for 4 genre classifications are shown in Table 2.
Above confusion matrix shows the accuracy of models on 4 genre classifications.

Table 2.4 Genre Accuracy

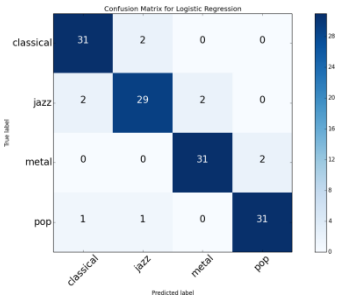
4 Genre Other Features	
Models	Accuracy
Support Vector Machine	79.39%
K-Nearest Neighbor	76.66%
Logistic Regression	78.78%
Neural Network	74.24%
Ensemble	81.36%
Naive Bayes	78.78%



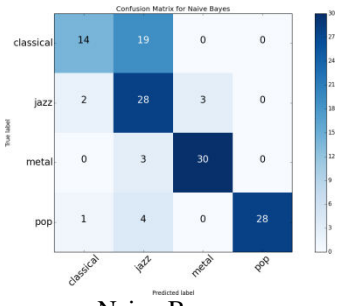
Support vector Machine



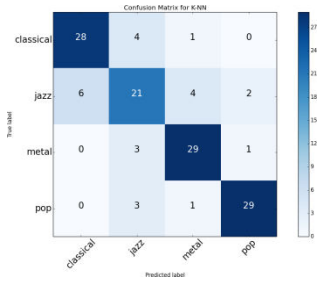
NeuralNetwork



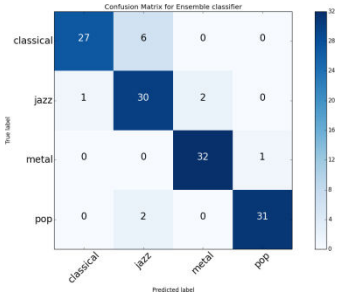
LogisticRegression



Naive Bayes



K-NN



Ensemble

Figure 6: 4 Genre Classification

5. Conclusions

From our experiments, we analyzed the accuracies of different machine learning models - Support Vector Machine (SVM), Decision Tree, Artificial Neural Network, Logistic Regression, k-NN and Ensemble based on Majority Vote. Based on our study, we were able to conclude that SVMs performs better in general to other classifiers. Logistic Regression and Ensemble method also performed quite well. If performance is a criterion to select a model, ensemble methods should be avoided. The best accuracy we achieved for 10-genre classification was 66.48% and the best accuracy we achieved for 4-genre classification was 81.66%

Acknowledgments

I thank Dr. Savita Choudhary for her guidance throughout this work. The material covered in work helped us in performing experiments on various models that were discussed. I would also like to thankDr Paramesha K for his help during the process.

References

Tzanetakis, George, and Perry Cook (2002).Musical genre classification of audio signals.IEEE Transactions onspeech and audio processing10.5 293-302.

McFee, Brian, et al (2015).librosa: Audio and music signal analysis in python. Proceedings of the 14th pythonin science conference.

Haggblade, Michael, Yang Hong, and Kenny Kao (2011).Music genre classification. Department of Computer Science, Stanford University

Camenzind and Goel Tom Camenzind and Shubham Goel (2013). Jazz Automatic Music Genre Detection. http //cs229.stanford.edu/proj2013/CamenzindGoeljazz Automatic Music Genre Detection.pdf

Lee, Chang-Hsing, et al. (2009). Automatic music genre classification based on modulation spectral analysis ofspectral and cepstral features. IEEE Transactions on Multimedia 11.4 670-682.

Schindler, Alexander, and Andreas Rauber (2015). An audio-visual approach to music genre classification through affective color features. European Conference on Information Retrieval. Springer International Publishing.

Howard, Sam, Carlos N. Silla Jr, and Colin G. Johnson (2011).Automatic lyrics-based music genre classification in a multilingual setting. Proceedings of the Thirteenth Brazilian Symposium on Computer Music. Vol. 34.

Maaten, Laurens's van der, and Geoffrey Hinton (2008). Visualizing data using t-SNE. Journal of Machine Learning Research 9. 2579-2605.

Kumar, D. Pradeep, B. J. Sowmya, and K. G. Srinivasa (2016). A comparative study of classifiers for music genre classification based on feature extractors. Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER), IEEE.

Abe, S. (2003) Analysis of Multiclass Support Vector Machines. International Conference on Computational Intelligence for Modelling Control and Automation (CIMCA 2003), 385-396.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2(2):955-974.

Kijsirikul, B. &Ussivakul, N. (2002) Multiclass support vector machines using adaptive directed acyclic graph. Proceedings of International Joint Conference on Neural Networks (IJCNN 2002), 980-985.

Vapnik, V. (1995) The Nature of Statistical Learning Theory. Springer-Verlag, London.

Vapnik, V. (1998). Statistical Learning Theory. Wiley-Interscience, NY.

Michalski, Ryszard S., Jaime G. Carbonell, and Tom M. Mitchell (2013). Machine learning: An artificial intelligence approach. Springer Science & Business Media.

Moffat, David, David Ronan, and Joshua D. Reiss (2015). An evaluation of audio feature extraction toolboxes.Proceedings of the 18th International Conference on Digital Audio Effects (DAFx-15), Trondheim, Norway.

Pedregosa, Fabian, et al. (2011)" Scikit-learn: Machine learning in Python." Journal of Machine Learning Research 12. 2825-2830.