Music Genre Classification

Mulitmedia Content analysis

Wen-Hai Tseng
RE6124027
Nation Cheng Kung University
RE6124027@gs.ncku.edu.tw

I. 程式執行環境

當談到音訊處理中的特徵提取和分類時，傳統方法和深度學習方法是兩個主要的取向。傳統方法通常涉及特徵提取器（例如 MFCC）和分類器（例如 SVM）的組合，而深度學習方法則更傾向於直接使用深度神經網絡，如 CNN。

本次 Music Genre Classification 將提取特徵並使用分類器（例如 SVM）的經典方法與在音訊表示（Melspectrogram）上使用 CNN 來提取特徵和分類的深度學習方法進行比較。

II. Visual Feature

音樂類型分類的方法主要包括特徵提取、深度學習和混合方法。在特徵提取中，從音頻信號中提取特徵如音調、節奏和頻譜等，然後利用機器學習算法將這些特徵映射到不同的音樂類型上。深度學習則利用深度神經網絡等技術，直接從原始音頻數據中學習特徵表示和音樂類型之間的映射關係，無需手動提取特徵。此外，混合方法結合了特徵提取和深度學習等方法，充分利用它們各自的優勢，以更準確地進行音樂類型分類。

A. 梅爾倒頻譜係數（Mel Frequency Cepstral Coefficents）

梅爾倒頻譜係數（MFCC）是基於人類聽覺特性設計的音頻處理技術，廣泛應用於語音識別等領域。其處理流程包括：首先進行預加重以增強高頻部分，接著將信號分割成多個幀並進行加窗處理；之後對每幀進行快速傅里葉變換（FFT）獲得頻譜，再透過梅爾濾波器組模擬人耳對頻率的感知；然後對濾波器的輸出取對數，以轉換其動態範圍；最後通過離散餘弦變換（DCT）減少特徵間的相關性。此外，為了捕捉語音信號的動態變化，常加入一階或二階差分係數。MFCC 不僅能有效提取語音特徵，更貼近人類的聽覺反應，使其在語音識別等應用中極為有效。

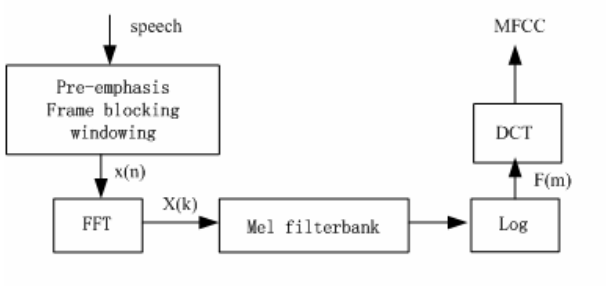

Fig.1 MFCC 特徵處理流程.

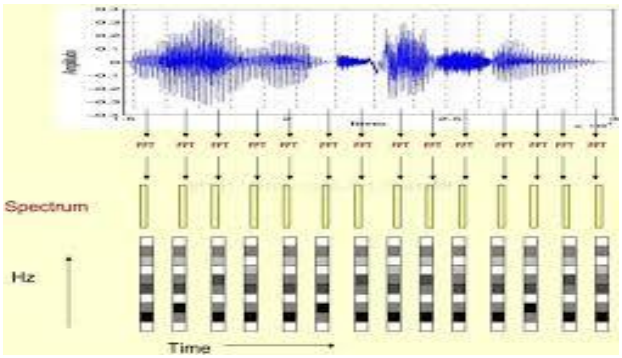

Fig.2 提取梅爾倒頻譜係數.

B. 梅爾頻譜（mel spectrogram）

梅爾頻譜（Mel Spectrogram）是一種聲音信號的頻譜圖，它按照梅爾尺度顯示聲音如何隨時間變化，更接近人類的聽覺感知。生成梅爾頻譜首先需將原始聲音信號分割成短時間內的幀，並進行加窗處理以減少邊緣效應。接著對每幀進行快速傅里葉變換（FFT），將時域信號轉換成頻域信號。隨後，這些頻譜經過梅爾濾波器組處理，根據梅爾標度設計一組梅爾濾波器如公式 1，將線性頻率轉換為梅爾頻率。最後對濾波器的輸出取對數能量，並將結果繪製成時間與梅爾頻率的二維圖表。梅爾頻譜因其有效表達聲音信號的時間和頻率信息並符合人類聽覺特性，被廣泛應用於語音識別、音樂分析等領域。

$$\mathrm{Mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (1)$$

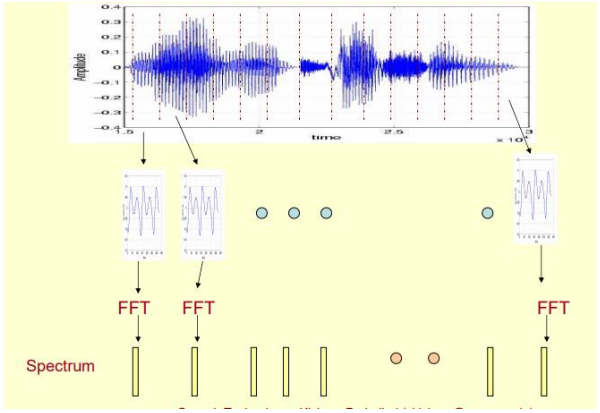

Fig.3 提取梅爾頻譜

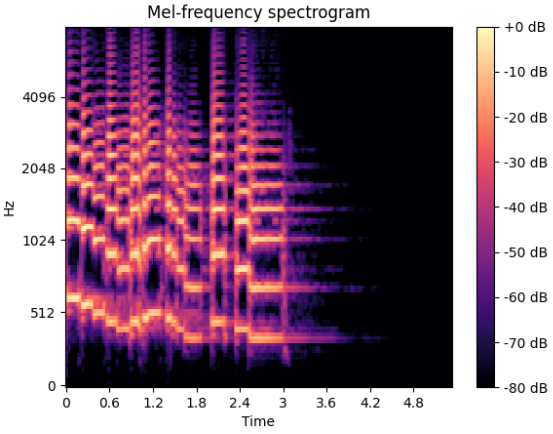

Fig.4 梅爾頻譜

III. Music Genre Classification 演算法

音樂類型分類可以使用不同的方法進行訓練和分類。一種方法是利用傳統的機器學習方法，例如使用從音頻信號中提取的 MFCC（Mel-Frequency Cepstral Coefficients）特徵，然後應用支持向量機（SVM）或隨機森林（Random Forest）等算法進行分類訓練。另一種方法是使用深度學習技術，例如使用 Mel 頻譜圖（Mel Spectrogram）作為輸入數據，通過深度神經網絡模型如卷積神經網絡（CNN）或循環神經網絡（RNN）進行訓練和分類。這兩種方法各有優缺點，前者需要手動提取特徵，而後者則能夠直接從原始數據中學習特徵表示，但需要更多的數據和計算資源。

A. *Logistic Regression*

邏輯回歸（Logistic Regression）是一種廣泛應用於統計和機器學習領域的分析方法，主要用於解決二分類問題。這種模型基於邏輯函數（或稱 Sigmoid 函數）。邏輯回歸的主要優勢在於其輸出的解釋性強，提供了預測概率，對決策制定非常有用。模型的參數通常通過最大似然估計方法學習，以使觀察到的樣本數據在模型參數下的聯合概率最大。然而，邏輯回歸假設數據是線性可分的，對非線性關係的處理效果有限，且在處理多類別問題時複雜性較高。儘管存在這些限制，由於其模型直觀和計算效率高，邏輯回歸在許多實際應用場景中仍是受歡迎的選擇。

$$\sigma(z) = \frac{1}{1+e^{-z}} \quad (2)$$

B. *ElasticNet*

ElasticNet 是一種結合了 Lasso 和 Ridge 回歸特點的線性回歸模型，廣泛應用於統計學和機器學習領域，特別適合處理特徵多、存在多重共線性的數據集。它通過在損失函數中添加 L1 和 L2 正則化項來進行模型訓練，其中 L1 正則化有助於進行變量選擇，自動排除不重要的變量，而 L2 正則化則有助於處理共線性問題，降低模型變異性。ElasticNet 的參數 α 和 λ 提供了調整 L1 和 L2 正則化比重的靈活性，使得模型可以在純 Lasso（α=1）與純 Ridge（α=0）之間平滑過渡。選擇這些參數通常需要透過交叉驗證等方法，雖然這增加了計算負擔，但 ElasticNet 在高維數據分析中的表現及其靈活性使其成為許多專業領域首選模型。不過，加入正則化的模型在解釋性上可能會有所減弱。

$$\text{Minimize}\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\left(\alpha\sum_{j=1}^{p}|\beta_j|+\frac{1}{2}(1-\alpha)\sum_{j=1}^{p}\beta_j^2\right)\right\}' \quad (3)$$

C. *Decision Tree*

決策樹是一種在機器學習中廣泛應用於分類和回歸任務的模型，它通過從根節點開始，根據特徵選擇和條件進行分割來建構樹狀結構，目的是使得每個節點的數據盡可能純淨。這種模型的直觀性強，易於解釋，且不需要對數據進行預處理如標準化或歸一化。決策樹能有效處理數據間的非線性關係，但也存在缺點，如容易發生過擬合，尤其是樹較深時，並且對數據變動敏感，可能因小的數據變化而產生完全不同的樹結構。此外，由於決策樹在節點分割時採用貪婪算法，只考慮當前最優解，因此可能無法達到全局最優。

D. *Random Forest*

隨機森林是一種集成學習方法，通過結合多個決策樹來提高模型的預測精確性和穩定性。在訓練過程中，每棵樹都獨立於隨機選取的樣本和特徵進行建構，這種方法被稱為自助聚合（bootstrap aggregating）或 bagging。這樣的設計不僅增加了模型對數據的適應性，降低了過擬合的風險，還提升了對新數據的泛化能力。隨機森林在分類和回歸任務中都能表現出高度的精確性，特別適合處理大規模且特徵複雜的數據集。然而，其模型結構的複雜性使得解釋性較差，且在計算和記憶資源上的需求相對較高。

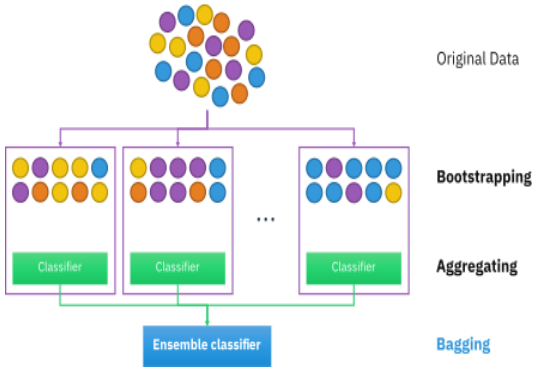

Fig.5 Bootstrap aggregating.

E. *Support Vector Machine*

支持向量機（Support Vector Machine，簡稱 SVM）是一種用於分類和回歸的監督式學習模型，其核心目標是在高維空間中尋找一個最佳超平面來最大化不同類別之間的間隔。SVM 利用核技巧將數據映射到更高維度空間，使得原本不可分的數據變得可分，常用的核函數包括線性核、多項式核、徑向基（RBF）核等。這種模型在中小規模數據集上通常能提供出色的分類效果，並且具有強大的泛化

能力。然而，SVM 在處理大規模數據集時可能會顯得效率低下，且其性能高度依賴於核函數與參數的適當選擇，這些參數的設定通常需要通過精細的交叉驗證來確定。

F. CNN

卷積神經網絡（Convolutional Neural Network，簡稱 CNN）是一種專為處理具有類似格狀結構的數據（如圖像）設計的深度學習模型，廣泛應用於圖像識別、視頻分析及自然語言處理等領域。CNN 通過卷積層使用多個卷積核自動從圖像中提取特徵，再結合非線性激活函數如 ReLU 來增強模型的非線性學習能力。此外，池化層用於降低特徵維度和增強不變性特性，而全連接層則負責將學到的特徵映射到最終的輸出類別。CNN 的優勢在於能夠自動學習顯著特徵並有效運用於多樣的視覺任務，但其也需大量計算資源且有過擬合的風險。

Fig.5 SHOT CHANGE DETECTION 流程

IV. 分類效能

首先將資料隨機打亂，然後按照 70%訓練和 30%測試的比例分割。接著，將音訊轉換成梅爾頻譜圖，並將這些頻譜圖分割成 1.5 秒長的窗口，每個窗口之間有 50%的重疊，從而形成一個具有樣本、時間、頻率和通道維度的資料集。之後，使用一個卷積神經網絡（CNN）進行訓練，或者將音訊轉換成梅爾頻率倒頻譜係數（MFCC）特徵，然後應用傳統機器學習方法。最後，你將在測試集上使用多數投票方法來評估模型的效能。

本次訓練在機器學習的訓練模型中，我使用了 Logistic Regression、ElasticNet、Decision Tree、Random Forest、Support Vector Machine，使用 cross_validation 設定 CV=5，這 5 種模型中在驗證集中準確度最高的是 SVM，並且在測試集上的表現也是最好，但是在驗證與測試上的表現落差有點高有可能過擬合的問題。

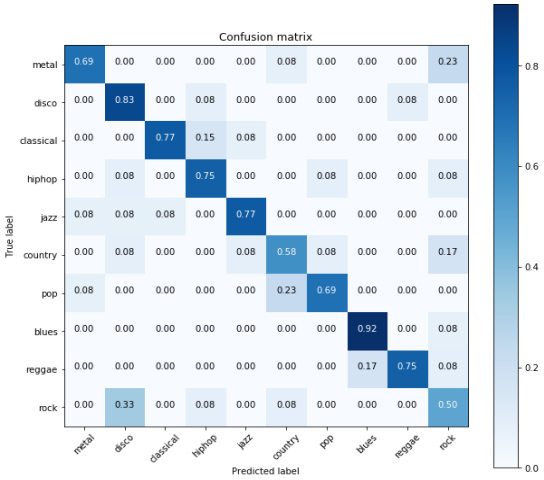

Fig.6 SVM 的混淆矩陣

本次訓練在深度學習中我採用了簡單的分類模型，模型訓練設定 epoch=150、optimizer 使用 Adam，loss 用 categorical_crossentropy，並且在訓練完成後準確度並有沒超過機器學習的方法，準確度只有 70.67%，但是使用 Majority Vote 的後處理將準確度提升到 73.67%。

	Val Accuracy	Test Accuracy
Logistic Regression	79.2%	72%
ElasticNet	75.2%	62.4%
Decision Tree	56%	51.2%
Random Forest	73.87%	72.8%
SVM	82.67%	72.8%
CNN	70.04%	70.66%(73.67%)

Table1. 6 個模型的準確度

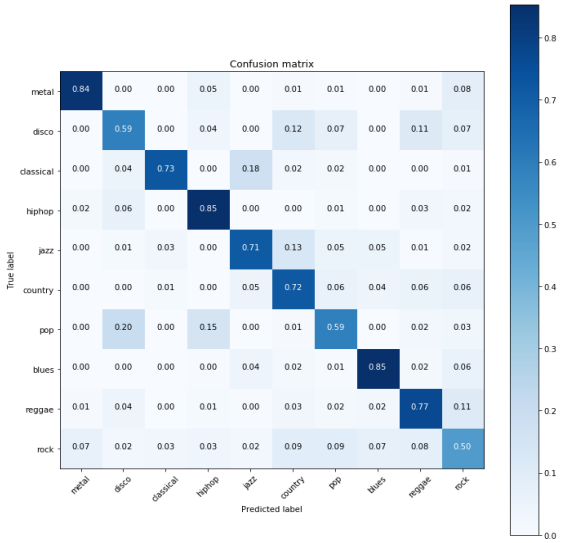

Fig.7 CNN 的混淆矩陣

根據提供的數據，各個機器學習模型在驗證集和測試集上的表現顯示出顯著的差異。支持向量機（SVM）和邏輯回歸在所有模型中表現最佳，其中 SVM 在驗證集上達到了 82.67%的準確率，測試集為 72.8%，顯示出較好的學習和泛化能力。邏輯回歸也表現穩定，準確率分別為 79.2%和 72%。隨機森林和卷積神經網絡（CNN）展示了良好的穩定性和競爭力，特別是隨機森林在驗證集和測試集上的表現非常接近。相比之下，ElasticNet 和決策樹的表現較弱，ElasticNet 在測試集上的準確率下降到 62.4%，而決策樹的準確率最低，顯示這兩個模型在當前應用中的泛化能力不足。總體而言，SVM 和邏輯回歸因其強大的學習能力和泛化性而脫穎而出，而 ElasticNet 和決策樹則可能需要進一步的優化來提高性能。

根從圖 6 和圖 7 的混淆矩陣可以看出，機器學習和深度學習模型在不同類別的表現上存在差異。如果在這些模型上進一步應用集成學習技術，還有可能提高整體的準確度。

REFERENCES

[1] https://github.com/Hguimaraes/gtzan.keras/tree/master
[2] https://zhuanlan.zhihu.com/p/351956040
[3] https://en.wikipedia.org/wiki/Bootstrap_aggregating
[4] https://zhuanlan.zhihu.com/p/459396415
[5] https://zhuanlan.zhihu.com/p/350846654

[6]https://medium.com/@fish90510/%E6%A2%85%E7%88%BE%E9%A0%BB%E8%AD%9C-19645e1bc75d

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not