**基於深度信念網絡的音樂情緒分類算法研究**

<http://bangqu.com/97e3Gv.html?fbclid=IwAR3uF3Ng4h9edhiRSA35n8QT1t34IFJWX7b1M1uvpQZyok9DweIE9SEGpag>

近年來互聯網技術快速發展，很多音樂平臺都爲用戶提供了在線數字音樂服務，以方便用戶獲取音樂。但隨着音樂數據的激增，根據音樂的名稱、藝術家等文字信息的傳統分類已不再能滿足人們的需求。因此急需引入新的音樂信息來爲聽衆提供更智能的音樂選擇類別。音樂是傳達情感的藝術，情緒感受是音樂最重要的本質特徵之一。一項針對音樂網站Last.fm的研究表明，情緒標籤是用戶最常打的標籤之一[1]。基於情緒將音樂分類的方式，從音樂感受出發能爲聽衆提供更大的方便。

傳統的基於情緒的音樂分類算法一般分爲3個部分：信號預處理、信號特徵提取、分類算法識別。其中提取的音樂特徵需要基於音樂和心理學的知識手工選擇。這種方法很難得到好的特徵表示，對特徵的優化也將非常困難。除此之外，傳統的機器學習方法利用的是淺層的網絡結構，一般只含有一個非線性特徵層，已無法滿足解決複雜的自然信號如語音信號、自然語言等問題。

近年來，深度學習在人工智能領域取得了較好的成果。它模仿人腦處理信息的機制，採用多層結構依次訓練網絡來從原始數據中提取複雜的結構表示。其中Hinton在2006年提出了深度信念網絡（Deep Belief Network,DBN）,該網絡可以自動學習特徵而不依賴人工提取，在很多領域都表現出了較好的效果。[2]針對講話者區分、音素檢測等任務，使用了加入卷積的深度信念網絡來處理聲學信號，其性能要優於傳統的聲學特徵，梅爾倒譜系數（MFCC）；[3]中結合DBN強大的特徵提取能力和高斯混合模型-隱馬爾可夫模型（GMM-HMM）序列進行建模的優越性，在大詞彙量的語音識別任務中取得了很好的效果。

本文對深度信念網絡進行了研究，並針對其在應用上的優缺點結合卷積神經網絡中的卷積層的思想，提出了一種新的基於情緒的音樂分類算法。本論文通過實驗將傳統MFCC特徵方法和DBN方法進行對比，並對優化後的算法進行實驗測試。實驗表明，加入卷積後的深度信念網絡較原始網絡有更好的分類性能。

1. 傳統音樂情緒識別

音樂信號具有高維度高冗餘的特點，因此在傳統的音樂情緒識別方法中通常需要先進行一系列的預處理，然後再提取合適的音樂特徵，最後將提取到的特徵進行統計訓練，利用分類算法識別音樂情緒類別[4]。傳統方法的一般流程如圖4所示。

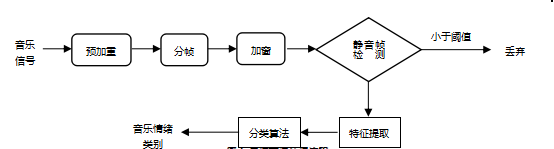


圖4.傳統方法處理流程

1. 預加重。音樂信號的高頻部分一般能量較低，需要使用濾波器濾除低頻干擾，來更易獲取高頻頻譜。這個階段就是預加重，信號的方程式爲： ，其中是預加重係數
2. 分幀。音樂信號是一種非平穩信號，但由於人體發音器官的相關原因，音樂信號又具有短時平穩的性質。爲此，需要將音頻截成一段段比較平穩的片段。截取的長度需要既能保證片段的平穩、又能方便處理。
3. 音樂信號中時常會出現一些大峯值信號，使信號不夠平滑難以處理。所以還需要對分幀的信號進行加窗操作來平滑信號，加窗信號爲：。一般常用的窗函數有漢寧窗、海明窗、布萊克曼窗等。本文在實驗中使用了海明窗進行處理。海明窗函數爲：

http://i2.bangqu.com/j/news/20190321/97e3Gv1553140804912s296Z.png

1. 靜音幀檢測。音樂信號中有時還會存在不具有信號參數特徵的靜音幀，影響識別的結果。該部分可以設置基於音頻信號短時能量的閾值來檢測靜音幀。若能量低於設定的閥值，則將其去除。本論文中，設定閥值爲0.1[5]。
2. MFCC特徵提取。提取特徵的好壞直接關係到最終輸出的結果。MFCC特徵是聲學任務中最常用的特徵之一。人類的耳蝸相當於一組濾波器，能夠通過過濾不同頻段、帶寬等特徵來區分不同的聲音。MFCC特徵就是在模仿這種能力。音樂信號的頻率與梅爾倒譜系數的關係如下：

http://i2.bangqu.com/j/news/20190321/97e3Gv15531408052718tJZj.png

從公式中可以看到，對於頻率小於1000Hz的聲音信號，梅爾濾波呈線性相關,頻率大於1000Hz時則成對數關係。MFCC提取過程如圖7所示：信號經過預處理後，通過快速傅里葉變換（FFT），求得各幀信號的頻譜信息；然後通過平方運算得到每一幀的功率譜；再設計一組梅爾濾波器，對上一步得到的幅度譜濾波。計算每個濾波器組輸出的對數能量；最後對上一步的輸出結果做離散餘弦變換（DCT）後即可得到MFCC特徵。

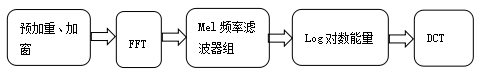


圖7.MFCC特徵提取過程

提取信號的MFCC特徵後，可將特徵輸入到分類算法中進行模型訓練，來學習情緒標籤和音樂特徵之間的聯繫。常見的分類算法有：支持向量機算法（SVM）、K近鄰算法、決策樹等。

1. 基於深度信念網絡的情緒分類
2. 深度信念網絡

深度信念網絡是Hinton在2006年提出的一種能自上而下自動學習特徵的多層神經網絡。它由多個受限玻爾茲曼機（RBM）堆疊而成。其結構如圖8所示,每相鄰兩層組成一個受限玻爾茲曼機。自下而上，前一個RBM的隱層即爲下一個RBM的顯層。前一個RBM的輸出即爲下一個RBM的輸入，以此類推逐層進行訓練，直至最後一層[6]。

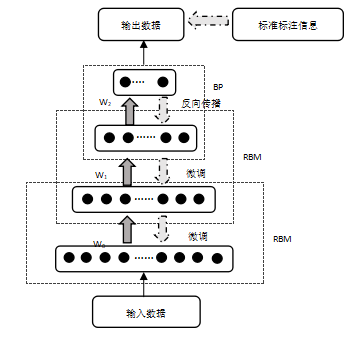


圖8.DBM結構

通過訓練學習即可得到各層的鏈接權重和節點偏置，完成網絡初始化。再採用反向傳導算法（BP）自上而下有監督的對深度信念網絡進行微調，來克服易陷於局部最優和訓練時間長的缺點。儘管深度信念網絡表現出了強大的特徵學習能力，但由上圖原理可知，網絡的訓練需要大量的樣本數據，且會生成更大數目的參數權值，一方面基於情緒的音樂分類問題缺少大量的樣本數據，另一方面大量參數的生成會耗費巨大的時間，不利於實際應用。

1. 基於卷積DBN的音樂情緒分類

由上文可知深度信念網絡在數據訓練中存在一些無法克服的缺點。要想解決此問題就需要嘗試減少神經網絡訓練的參數數目，常用的技術就是權值共享方法。權值共享是卷積神經網絡中卷積層採用的一種思想。與普通的神經網絡全連接、參數不共享不同，卷積神經網絡具有局部連接、權值共享的特點。所謂的權值共享，即指一層中多個節點的連接共享相同的一組參數。共享參數的可行性主要來源於生活中很多對象都有局部相關性，例如音樂中每一幀信號並不是孤立的，正是連續各幀共同作用，纔將情緒傳達給聽衆。因此，將卷積神經網絡與DBN相結合，利用權值共享技術能顯著減少參數的數量，還能反映音樂傳達情緒的本質，具有重要的意義。

與傳統機器學習方法不同，DBN是將信號的原始時域特徵作爲輸入，並通過逐層訓練自動學習輸入信號的特徵。常用的時域特徵有短時平均幅度差、短時能量、短時自相關係數等。本文采用短時能量作爲深度神經網絡的原始輸入特徵。短時能量能夠反映聲學信號的能量隨時間變化的過程。通過短時能量特徵能明顯區分清濁音段。設一段音樂信號爲x(n)，則其短時平均能量爲：

http://i2.bangqu.com/j/news/20190321/97e3Gv1553140807037h5wK5.png

但原始的時域特徵維度很高，存在大量的冗餘和噪音。因此需要對輸入的訓練數據進行降維處理。本文將採用主成分分析（Principal Component Analysis，PCA）的方法。主成分分析，是考察多個變量間相關性的一種多元統計方法，研究如何通過少數幾個主成分來揭示多個變量間的內部結構。

在訓練數據經過降維處理後，輸入DBN中進行訓練，來自動學習數據的特徵信息，最後將輸出數據通過深度學習中常用的Softmax分類函數進行數據處理，完成分類任務。具體基於DBN的音樂情緒分類結構如圖9所示。

本文針對原始深度信念網絡的缺點提出的結合卷積神經網絡的DBN與原始DBN不同。它將卷積神經網絡中的卷積層和下采樣層共同加入到了DBN中，如下圖10所示，在神經網絡的第二層有6個節點，在第三層具有2個節點，如果使用全連接的話需要訓練的參數爲12個，由於採用了權值共享並加入了下采樣需要訓練的參數變爲了3個。卷積層的輸出會作爲下一層受限玻爾茲曼機的隱藏層節點。

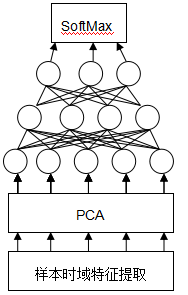


圖9.基於DBN的音樂情緒分類

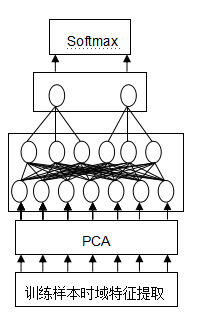


圖10.基於卷積DBN的音樂情緒分類

1. 對比實驗分析

實驗共使用三種方法進行音樂情緒的分類。第一組使用MFCC係數作爲輸入特徵；第二、三組使用原始的時域特徵作爲輸入，並分別採用DBN和加入卷積的DBN學習特徵。爲了保證實驗的可比性，都將使用softmax作爲分類算法。

1. 數據集

由於自建基於情緒的音樂分類數據集需要花費大量的時間和資金，且需要一定的音樂和心理學知識，因此本論文采用了美國德雷塞爾大學音樂和娛樂技術實驗室（Music and Entertainment Technology Laboratory ，MET-lab）的MoodSwings數據集。該數據集以一種協同式在線遊戲的方式收集用戶對音樂情緒的打分，並運用一些遊戲規則和激勵方式來儘量保證標記的客觀性。遊戲是目前比較科學的情緒表徵空間，V-A空間進行的[7]，如圖11，V維度是愉悅程度，代表情緒給人感覺的正或負；A是活化程度，代表音樂的能量高低。

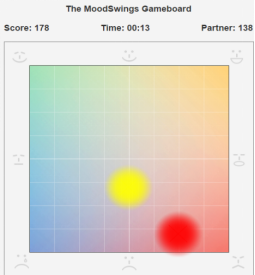


圖11. V-A標記遊戲

每組實驗樣本都採用MoodSwing數據庫中歌曲片段，隨機選擇800首歌曲作爲訓練樣本，其餘200首歌曲作爲測試樣本。樣本歌曲的情緒被分爲4個類別。

1. 基於傳統方法和基於深度信念網絡音樂情緒分類對比

第一組實驗採用MFCC係數作爲輸入特徵。使用30ms幀長、20ms窗口的海明窗對音樂進行分幀處理，然後使用Matlab下工具箱MIRToolBox對每一幀的音頻片段提取N維的MFCC特徵值，系統架構如圖12：

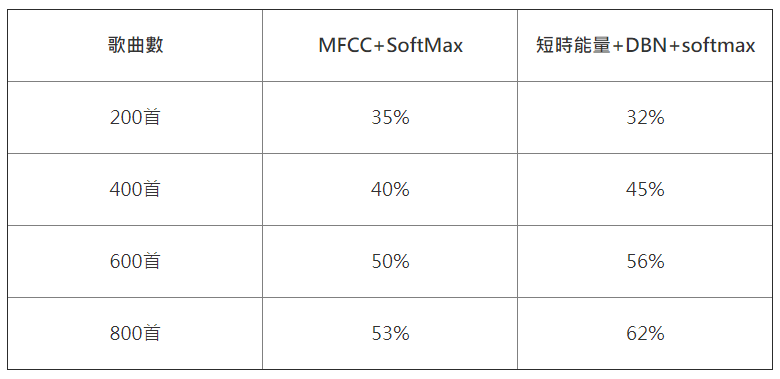
http://i2.bangqu.com/j/news/20190321/97e3Gv15531408096329HJ77.png

圖12.傳統音樂情緒分類架構

第二組實驗使用Matalab的MIRToolBox工具箱對每首歌曲的每一幀進行短時能量的提取，得到2000X1000維度的矩陣。然後利用PCA進行降維處理，得到1200X1000維的轉換後的數據，將這些特徵數據輸入到深度神經網絡進行訓練得到模型。

第一組實驗與第二組實驗的對比結果如下。爲了驗證歌曲數目對模型分類性能的影響，每次實驗利用N首歌曲當做訓練樣本。第一組中，使用全部N首進行模型的訓練。第二組中，採用2/3N的樣本用於無監督的訓練，1/3N的樣本用於有監督的微調。此次採用4層DBN，兩個隱藏層分別爲800個神經元節點和300個神經元節點，最後的輸出層有100個神經元節點。實驗結果如下表1所示：

表1.MFCC與DBN提取特徵對比



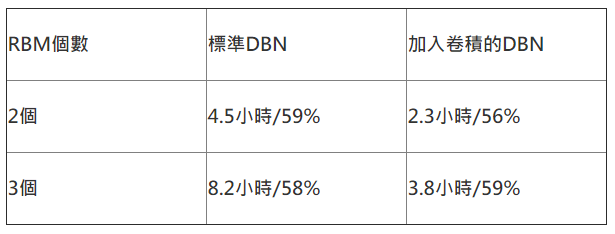
從數據中可以看到，兩種分類方法的分類準確度都隨着樣本數據的增長而增長。但是使用MFCC+softmax的方法的準確度增長速度逐漸減慢，而採用深度神經網絡的分類方法，仍在穩定的增長。這說明在大量樣本數量下，深度學習方法將比傳統方法分類效果更好。而且DBN可以使用無標籤數據進行訓練，這更符合實際場景數據的需要。

另一方面可以看到，在樣本數量非常少的情況下，使用DBN提取到的特徵性能並不如MFCC特徵。這也證實了深度學習網絡需要一定數量的樣本進行訓練才能更好的擬合測試樣本。

1. 加入卷積操作的DBN

上節實驗顯示使用原始DBN進行音樂特徵的提取，需要大量的樣本數據進行訓練，且需要花費大量的訓練時間。因此需要利用權值共享技術，將深度信念網絡結合卷積神經網絡來提高深度信念網絡的性能。本實驗在深度信念網絡中加入卷積層，來對音樂的情緒進行分類。 實驗結果如下表2：

表2.標準DBN和加入卷積的DBN對比



從數據可以看到，加入卷積層的DBN訓練時長遠小於原始DBN，這是權值共享帶來的好處。然而還應該看到，進行音樂分類的準確度並沒有顯著提高，但這有可能是參數選擇不當導致的後果。

加入卷積層來源於將卷積神經網絡和深度信念網絡結合的想法，沒有具體的理論支持和嚴謹的推導，所以沒有取得顯著地性能提升。但是這種加入卷積的思想是值得深入研究的。

1. 結論

基於情緒的音樂分類任務，是人工智能擁有情緒感知能力的初步嘗試，能夠爲用戶帶來新的娛樂體驗。本文針對音樂的情緒分類問題，分析研究了傳統的機器學習方法的性能以及深度信念網絡在音樂信號處理的能力，並將卷積神經網絡與深度信念網絡相結合，優化了深度信念網絡在訓練時需要大量數據樣本來生成大量參數，耗時耗力的問題。實驗顯示，加入卷積的深度信念網絡確實比原始網絡在訓練耗時上有較好的效果，但是也許因爲參數設置的問題，其準確率並沒有得到很大的提升，這也是之後工作需要進一步解決的問題。

1. 參考文獻
2. 郭俊. 社會化數字音樂服務設計研究——以Last.fm的用戶標籤行爲數據分析爲例[D]. 中山大學, 2012.
3. Deng L, Seltzer M L, Yu D, et al. Binary coding of speech spectrograms using a deep auto-encoder[C]//Interspeech. 2010: 1692-1695.
4. Mohamed A, Sainath T N, Dahl G, et al. Deep belief networks using discriminative features for phone recognition[C]//Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. IEEE, 2011: 5060-5063.
5. 王磊, 杜利民, 王勁林. 基於AdaBoost的音樂情緒分類[J]. 電子與信息學報, 2007, 29(9):2067-2072.
6. 張昕. 基於深度信念網絡和隱馬爾科夫模型的音樂自動分類算法研究[D]. 大連理工大學, 2015.
7. Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2014, 18(7):1527-1554.
8. Grimm M. Emotion estimation in speech using a 3D emotion space concept[J]. Robust Speech Recognition & Understanding, 2007, 16:381--385.