The 2016 Conference on Computational Linguistics and Speech Processing
ROCLING 2016, pp. 387-401

基於多模態主動式學習法進行需備標記樣本之挑選用於候用校長評鑑之自動化評分系統建置

A Multimodal Active Learning Approach toward Identifying Samples to Label during the Development of Automatic Oral Presentation Assessment System for Pre-service Principals Certification Program

孫泓敬　Hung-Ching Sun
國立清華大學電機工程學系
Department of Electrical Engineering
National Tsing Hua University
s103061558@m103.ntnu.edu.tw

李祈均　Chi-Chun Lee
國立清華大學電機工程學系
Department of Electrical Engineering
National Tsing Hua University
cclee@ee.nthu.edu.tw

摘要

主動式學習 (active learning)在機器學習領域中越來越受到重視，因為它可以用來優化訓練的過程，讓結果更好[13]。主要的概念是假如學習演算法可以在學習的過程中選擇比較決定性的資料點而不是挑選全部資料來做學習。接著根據對於模型而言具有代表性的資料點做挑選，將會對於學習的效果更有幫助，獲得更佳的結果。換句話說，透過觀察已知的標記資料，主動地挑選未標記的資料，並藉此獲得比挑選全部資料或是隨機抽樣資料的監督式學習方式更高的準確率以及更少的資料量。

對於任何監督式學習 (supervised learning)來說，假如想要促使學習系統表現的更好，則需要大量的被標記的資料來做訓練。但是，在這些被標記的資料中，可能會存在著對於學習系統有著負面影響的資料，從而降低學習效果與準確率。在這篇論文中，我們將會應用主動式學習的概念在系統學習的過程上，藉此來分辨資料對於系統的好壞；並測試主動式學習在訓練過程中的實際效果[9]。

Abstract

Active learning is becoming more and more important in machine learning that can optimize the learning process [13]. The main concept is that if learning algorithm can choose the most informative data points from which it learns, instead of choosing all of them, it will perform better with less training. In other words, we recursively select the unlabeled data instances by observing the known labeled data instances to obtain higher recognition accuracy while using smaller amounts of data instances, i.e., a subset of all of the dataset or random choose data when training the supervised learning system. [9]

For any supervised learning, if you would like to make the system perform well, it had to be trained on lots of labeled instances. But, in these labeled instances, there might be some worthless instances which affect the learning system and raise your training cost. So, we used the active learning concept during training process to discriminate whether the data instance is good for the learning system or not. In this work, we would like to know that the concept of active learning to select the training data, will work or not.

關鍵詞：主動式學習、資料選取、多模態訊號處理、機器學習

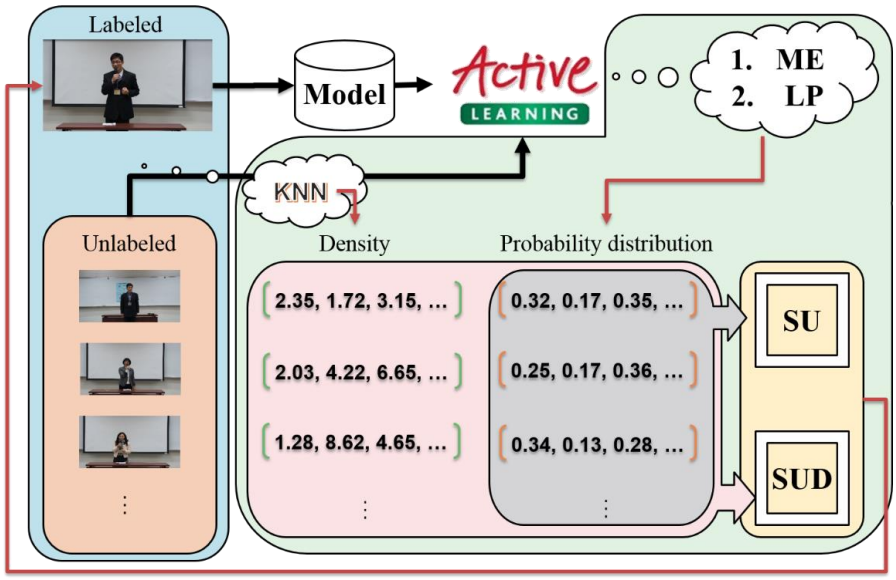
Keywords: active learning, data selection, multimodal signal processing, machine learning

一、緒論

近年來，機器學習已經在科學、工業、金融、疾病預防上發展越來越蓬勃，透過設計和分析並從中獲得規律，來預測未知數據可能的分布以及未來可能出現的情況。由於機器學習需要大量的資料來輸入，假如我們能夠嘗試對於這階段來做優化，找出更具有資訊性與代表性的資料數據，刪除特例、離群值(outliers)，這樣不但可以讓整個訓練效果提升、降低訓練的時間，更可以透過篩選去抽出值得被標記的資料去做標記，免去標記所有資料而產生出的巨幅成本[7]。在這篇論文中，我們將會基於主動式學習的概念來做為資料篩選的核心。根據部分已知的資料來篩選出對於系統最有價值的未標記資

料，也就是變異度大的資料，並將這些未標記的資料給予專家來標記。讓我們不再毫無目的性的輸入所有資料或是隨機性的輸入，而是循著讓學習系統的學習效果更快更有效率的方向前進[12]。

在這裡我們將會使用在主動式學習方式中常見的方法 Sampling by Uncertainty (SU)，以及其改善過後的方法 Sampling by Uncertainty and Density (SUD) [16]，並基於標籤傳播演算法、最大熵模型來建構在主動式學習法上[2]。前篇論文[3]中，我們使用此資料庫來建置一個校長演講自動化評量系統，希望透過此系統來輔助評分校長對於候用校長的演講評分上能夠更加客觀。我們根據這些演講內容，分別以聲音與影像部分著手，透過人類行為訊號處理的方法[8]來擷取出個別的特徵作為我們此篇論文所使用的資料庫。為了建置這個校長演講自動化評量系統，我們需要對於每個演講進行評分。個別演講需要至少兩位資深校長來進行評分，而每個演講被評分的平均時間長度大約三分鐘，對於每年度 200 筆演講來說，整個評分過程相當耗費人力與時間，因此希望透過主動式學習，來有效減少需要被評分的演講。



圖一、實驗流程

實驗內容的整體核心架構如圖一，我們設計出一組實驗來驗證並分析主動式學習對於機器學習上的幫助，並分析資料篩選過程與其對應篩選資料的特色。在一個連續性的評分中，將經過主動式學習的方式所篩選的資料與完全沒有篩選過的資料，分別算出與

實際評分校長評比分數的相關性。對於聲音與影像以及將兩者合成後三種結果來比較選擇前後其相關性高低。

二、資料蒐集

（一）、 校長演講儲訓計畫資料庫

在這裡我們蒐集 103 年度的校長儲訓計畫(NAER)資料作為聲音與影像特徵的來源。總共分為四個班級，大約包含 200 名的候用校長。在這 200 名候用校長中，只有 186 個演講擁有至少兩名評分校長來對他們做評分。評分校長們會根據他們的表現細分出七個方面做評比[10], [11]，如下：

1. 內容：演說內容是否符合主旨(0-20)
2. 架構：架構是否分明、井然有序(0-20)
3. 用字遣詞：用字對於聽眾是否適合(0-20)
4. 態度得體：服裝、儀態(0-10)
5. 發音標準：咬字(0-10)
6. 語調：音量適切、語句流暢(0-10)
7. 時間：控制得宜與否(0-10)

我們選擇將這七大項目的評比做加總來做為最後訓練模型時候的標記。在錄製的器材上，我們採用高解析度的索尼(Sony)攝影機並配備外部指向性麥克風，為了讓整體影像都能夠捕捉到整個演講者的上半身，我們將攝影機做固定；而錄製環境則是位於一間教室裡，演講者使用手拿式麥克風連接擴音器來進行演講。

在校長儲訓計畫中，透過對於候用校長的口頭演講作為他們的期末測驗，而他們的評分表格則會用來作為系統學習的標記。總共八位評分校長分別對應四個班級，而每位評分校長大約對 30 至 40 位不重複的候用校長作評分。在這篇論文裡，只針對最後總分部分作為學習系統的標記，並且將分數另外作正規排序法作為另一種分數標記的向度。

（二）、 資料庫標記與正規排序法

二元標記法在目前訓練機器學習系統去辨識客觀性評比上是一種常見的標記方式，優點在於比較極端的行為預測上有較好的分辨效果。大部分機器學習在學習標記上都是從比較極端的分類部分著手研究自動化系統的可能性，對於中間模糊地帶則是依靠著強化訓練模型或是對於特徵做更高維度的編碼來提高那部分的辨識率。在此篇研究中，我們將各評分校長評比完成的 186 組演講分數做平均，然後加以排序並提取高分、低分組各 20%來標記成「高」與「低」，而剩下的中間模糊地帶則標記為「中」。

由於每位評分校長評分上的分部差異皆不盡相同，而導致在選擇高低分組別時候會出現偏差，所以在這裡我們引入正規排序法[5]來降低因為評分的範圍不同而影響各組別選取資料的偏差。正規排序法是一種對於訊息提取與標記正規化上常用的方式，首先他會對個別評分者所評分數做排序並從小到大依序從 1 標記它們到結束，接著對這些標記的分數做總個數的正規化(除以分數的總個數)。最後將評比同一組別的兩位評分者分數在將其分成高低組之前，正規化後取平均值。透過使用原始分數的標記與正規排序化後的標記來作為整個系統的學習標記。

三、實驗理論

（一）、 短時高密度特徵擷取法

1. 短時高密度聲音特徵擷取法

我們使用高維度的特徵擷取方法來對於校長演講中的聲音部分擷取特徵，這種特徵擷取分析方式在很多過去的研究上，用來表示演講在不同面向上的特徵，配合上影像訊號的特徵對於複雜的辨識上有著高分辨率。除此之外，也可以用於複雜且全面性的人類行為模型建構、情緒辨識等部分上[1], [6]。一般來說，計算聲音特徵的方法是依照演講者的停頓為一個段落的方式(segmentation per utterance)將演講的聲音檔案分段，接著對於每份聲音檔案按照固定音框為單位並重疊一半音框大小來計算出低階描述特徵 LLD

(low level descriptors)。之後再對整個結果接續計算出其在統計學上的各種基本函數值，產生出一組高維度的向量特徵。

與之前不同的地方是，這篇論文不僅僅只對於整個音檔計算統計學數值，而是使用移動式音框的方式(sliding window approach)對於單一聲音檔案按照 200ms 為單位的音框並重疊其 50%音框大小來移動，即 100ms 來計算高維度的統計函數值，以此方式來嘗試捕捉聲音在統計數值上面更詳細的時變性。

總結來說，聲音短時高密度聲音特徵擷取法是對於三分鐘的演講來執行，步驟如下：

1. 使用語音活性檢測 VAD(voice activity detection)，來進行聲音檔案的自動化切割。
2. 透過移動式音框的方式，對每組句子產生維度 171 維的聲音低階敘述特徵向量。
3. 透過移動式音框對聲音低階敘述單元計算統計函數，產生 8861 維的特徵向量。
4. 合併同一演講中個別不同長度的切割後音檔，對於每組演講產生出維度 8861 維的聲音特徵矩陣。

由於每組聲音檔案計算後的長度不同，所以我們將會對於這些聲音特徵做更進一步的編碼，將其轉換成一組固定長度的聲音編碼特徵。實驗的特徵擷取是採用 opensmile 計算工具來完成[4]。

2. 短時高密度影像特徵擷取法

這篇論文使用短時高密度軌跡方法來計算影像的特徵值[14]，在這裡的資料點單位代表的是影像幀數 (15Hz $\approx$ 66ms)。與其他影像特徵擷取不同的地方是，這個方法是對於影像幀數進行高密度的取樣，而不是尋找影像的關鍵特徵點。基本上，這個演算法首先對於個別幀數進行高密度特徵點取樣，接著刪除不需要的特徵點，包含了根據自相關函數計算而無法隨著時間追蹤的特徵點(可能肇因於沒有移動)，或是太過頻繁移動的特徵點(錯誤的特徵點追蹤)。對於演講的影像部分，最重要的地方在於身體的移動、手勢的動作等，所以我們擷取了與其最相關的特徵來使用，如下：

- MBH_{xy}：基於光流(optical flow)對於 x 和 y 方向計算相對移動，以減少受到相機晃動等干擾，並量化為直方圖(motion boundary histogram)

- Traj：將 x 和 y 方向的高密度移動訊息軌跡正規化[15]

這篇論文中，我們還採用基於加速穩健特徵 SURF(speed up robust features)和隨機抽樣一致性 RANSAC(random sample consensus)兩種方法來改進攝影機位置估算法。透過新的算法，可以在計算特徵值之前移除由於攝影機的移動所造成的誤差而產生出錯誤的追蹤軌跡。總結來說，短時高密度影像特徵擷取法對於軌跡計算出以下兩種特徵值：

1. TRAJ：每 15 個幀數(frame)計算 x 和 y 的移動訊息軌跡，維度為 30 維。
2. MBH_x 和 MBH_y：個別維度為 96 維的 x 和 y 方向相對應的移動軌跡邊界直方圖。

（二）、 段落層級特徵值編碼

透過短時高密度特徵擷取法所產生的特徵中，聲音部分是根據每 200ms 為一個音框單位重疊 100ms 時間長度來移動計算，最後得到長度為 8861 維度；影像部分則是對於每 66ms 來計算，最後得到長度為 222 維，並且降維成 111 維(15 維 Traj、48 維 MBH_{xy})。根據不同演講的時間長度，會計算出不同長度的特徵結果，所以在這邊我們將會使用兩種編碼方式來編碼，將短時小單位的高密度特徵轉變成段落層級的固定長度向量表示方法：K 類分群方式的詞袋模型 k-means bag of word encoding (BOW)以及費雪向量編碼 Fisher-vector encoding (FV)。

1. 基於 K 類分群方式之詞袋模型編碼

首先我們對於輸入的所有資料的特徵值作個別的隨機取樣，並透過 K-means 分類方式來對於全部抽樣後的資料，訓練出類似字典(dictionary)或是密碼書(codebook)。接著，再根據訓練好的密碼書來將我們的輸入特徵值分配到其對應最相近的密碼或是字，計算距離的方式是使用歐式距離(Euclidean distance)。之後根據每個口頭演講被分配完後的結果，再對他們作直方圖，並使用標準分數正規化(z-score normalization)來正規化個別演講。最終對於個別演講會得到一個長度為 K 的聲音與影像的特徵值。

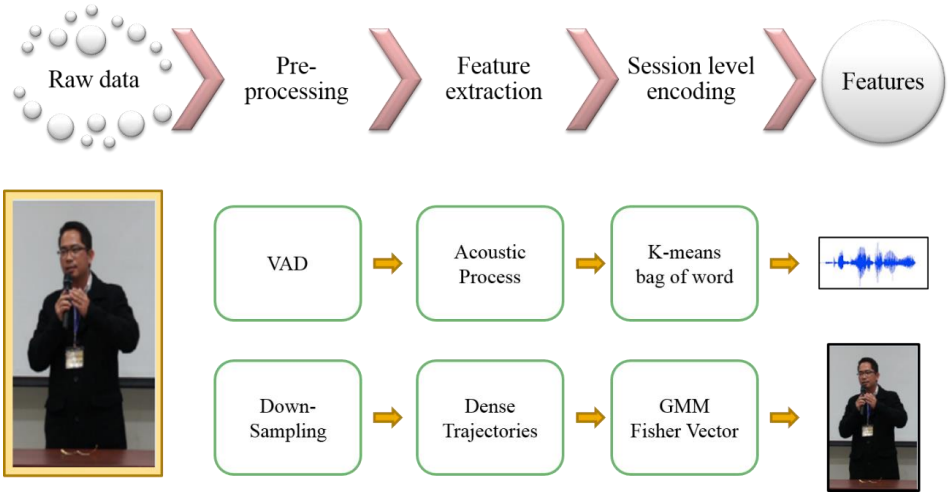
2. 費雪向量編碼

費雪向量編碼被證明在影像辨識上的結果優於 BOW，它擁有的優點在於生成模型以及判別模型上，而且相較於 BOW 只有一階的統計層，它則分成兩階的統計層來對於特徵值作編碼[19]。首先對於高密度單位的影像特徵值隨機抽樣，並將抽樣後的個別特徵值進行費雪核(Fisher Kernel)的計算，此方法是結合了生成式模型、判別式模型的優點，用來計算兩種不同的資料分布間的相似性。接著假設一組似然函數(likelihood)，其梯度向量表示：

$$G_{\lambda}^{X} = \nabla_{\lambda} log\, u_{\lambda}(X)$$

$u_{\lambda}(X)$ 代表了 X 的機率密度函數(pdf)，接著透過高斯混合模型 Gaussian mixture model(GMM)訓練輸入特徵值，並產生出機率密度函數，接著取出高斯混合模型的標準差、平均值分布作為機率密度函數的輸入。最後對於這兩項計算上列的公式的梯度向量，並計算費雪信息矩陣(FIM)的一階、二階導數，得到最後的費雪向量結果，再對其使用 L2 正規化[18]。

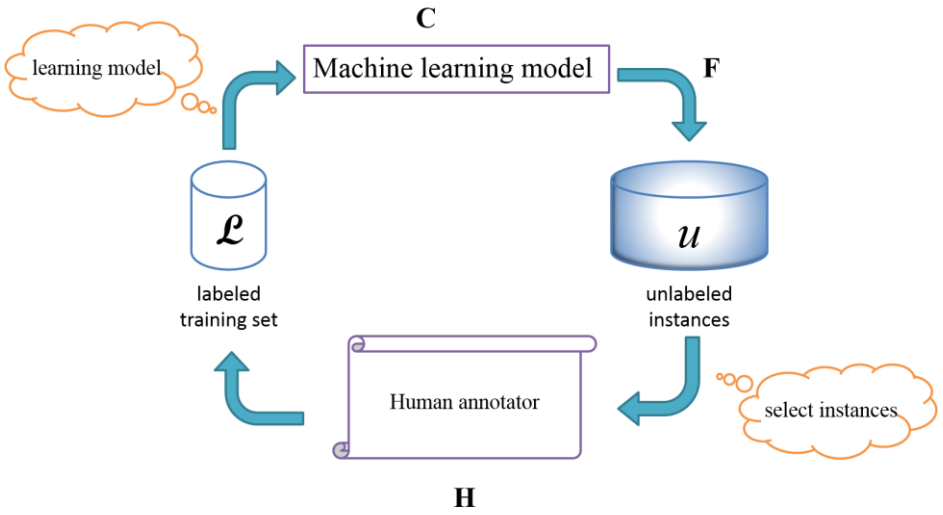
最後比較兩種編碼在聲音、影像的結果後，聲音部分使用 BOW 的編碼方式，而影像則是使用 FV 的編碼作為最後機器學習的輸入特徵。以下為特徵擷取與編碼流程圖：



圖二、特徵值擷取與編碼流程

（三）、　主動式學習法

大部分的情況下沒有標記的資料非常多，而且每筆資料給予專家標記的成本都很高。所以我們透過每次學習系統所回饋的資訊來推導出符合何種性質的資料，對於接下來系統的學習有所幫助，再將這些數據資料給專家進行標記，最後將標記好的資料加進訓練樣本裡進行系統訓練。主動式學習法由五種元素所構成，分別為：L 為已標記資料樣本、μ 為未標記資料樣本、C 為機器學習使用分類器、F 為分類器在學習中透過此信息函數來回饋信息、H 為專家標記，實際操作流程如圖三。



圖三、主動式學習流程圖

主動式學習的主要核心在於如何設計 F 來擷取學習模型所回饋的資訊，使其能夠從 μ 未標記資料樣本中選擇出最佳的資料，經過不斷的疊代以及回饋的資訊讓其被標記後能夠有效提升準確率。

1.　不確定性採樣法

直觀的來說，這個方法就是尋找對於目前學習模型來說，哪些未標記資料可以提供最多的信息量，讓學習模型可以更全面性的學習其缺乏的地方。透過尋找「最不確定」的資料，我們稱為不確定性採樣法(Sampling by Uncertainty)。對於擁有較大不確定性的資料來說，意味著根據經過訓練的學習模型上，其提供對於每項資料的可信度裡，該筆資料擁有比其他資料更具有資訊性的性質，對於目前系統來說變異度比較大，希望透過

學習此資料來提高模型的辨識率。相反的，可信度較高的資料反而無益於模型辨識率的提高，因為這些資料基本上已經在學習模型裡完善的學習過了。在每筆資料數據裡，可以得到對於每種標記的預測機率分布，接著根據這些機率分布來計算其信息熵作為前面所提到的 F(信息函數)。信息熵的計算公式如下：

$$\mathrm{H}(\mathrm{x}) = -\sum_{y \in Y} P(y|x) \log P(y|x)$$

在這邊的$\mathbf{P(y|x)}$代表的是資料數據對於標記的預測機率分布，H 則是我們所計算出來的信息熵。接下來我們將介紹在這篇論文中，用來計算信息熵的兩種學習模型演算法。

（1）、 標籤傳播演算法

標籤傳播演算法是一種半監督式的機器學習方法[17]。核心概念是，相似的資料點或鄰近的資料點應該具有相同的標記；也就是說，資料點的標記會根據他們的鄰近程度來傳播。已標記的資料如同一種資料源一樣，傳播到附近的未標記資料。透過標籤傳播演算法算出每筆資料對於各種標記的機率分布，再對其機率分布計算信息熵，以實現不確定性採樣法的實驗架構。經過測試後採用參數為$\sigma = 0.3$ (RBF kernel)的標籤傳播學習模型來實現，可以得到最佳的結果。

（2）、 最大熵模型學習法

邏輯回歸法(logistic regression)基本上就是最大熵模型學習法在對應學習類別為兩類的情形，當其擴展到三類別甚至多類別的學習時候，就是我們所熟悉的最大熵模型(Maximum Entropy Model)。最大熵模型學習法的原理是一種計算機率模型的準則，在學習許多不同的隨機變量統計模型時候，在滿足全部的已知條件下且不對於未知情況作主觀的假設，此時的機率分布相對的比較平均，預測的風險相對來說最小；也就是說，

對所有可能的機率分布中挑選出最客觀的機率分布。因為此機率分布保留最不確定性的資訊，也就是熵最大的情況，所以稱為最大熵模型學習法。

2. 不確定性密度採樣法

不確性採樣法的目的在於找出靠近決定邊界(decision boundaries)附近的未標記資料點，而且假設這些資料點有著最大不確定性，但是這個方法實際上還是有一些問題存在。通常來說，資料庫中會出現一些特異的資料點，可能具有一定程度的信息量(信息熵很高)，但是卻對於訓練系統的學習上比較沒有幫助，我們稱這些特異的資料點為離群值或極端值[20]。

為了避免離群值的情況出現，而且我們希望挑選出來的未標記資料點，不但具有資訊性(信息熵高)，而且還要具有代表性(密度大)。所以我們透過一種密度計算方式，來計算未標記資料點附近有幾個與其相似或相鄰，密度越高則該資料點屬於離群值的機會就越小。該密度計算方式，我們是基於最近鄰近法 KNN(K-Nearest-Neighbor)加上餘弦相似性法(cosine similarity)來實現密度計算。

四、實驗設計與結果

聲音部分，透過使用短時高密度聲音特徵擷取法對於每個演講產生出音框數不一，但是均為 8861 維的特徵值。接著再透過 K 類分群方式的詞袋模型對其作編碼，最後對於每個演講皆可得到對應一組 2000 維的最後編碼特徵。影像部分，透過使用短時高密度影像特徵擷取法對於每個演講產生出 111 維的特徵值。接著進行費雪向量編碼，對應每個演講者將產生出個別為 56832 維的輸入特徵值。

（一）、 實驗設計

在這部分，我們將會設計一個對於總分部分的連續性分數實驗預測。對象包含聲音、影像兩大部分來執行，並計算個別使用主動式學習後的效果以及對兩者進行多模態特徵融合模型訓練(Multimodal fusion)。最後我們將個別計算後的斯皮爾曼相關性做為最後結果(在情緒辨識上，經常會將此結果做為最後評判系統學習好壞的指標)。支持向量回歸模型使用的參數為$\epsilon = 0.15$來得到最好結果。多模態特徵融合則是將聲音、影像兩部分最後預測的分數結果，做平均值來計算新的斯皮爾曼相關性。

（二）、 實驗結果

以下為聲音、影像個別對應原始分數、正規化分數的最佳相關性結果，並且將最佳值標記為紅色。其對應表示參數：傳播標籤法(LP)，最大熵模型學習法(ME)，不確定性採樣法(SU)，不確定性密度採樣法(SUD)

表一、聲音、影像之主動式學習結果

Active Learning Spearman results					
	Spearman: Raw				
	LP_SU	**LP_SUD**	**ME_SU**	**ME_SUD**	**baseline**
Audio	0.494	0.505	**0.528**	0.495	0.443
Video	0.353	**0.368**	0.357	0.362	0.343
Fusion	0.515	0.538	**0.551**	0.529	0.485
	Spearman: Rank				
	LP_SU	**LP_SUD**	**ME_SU**	**ME_SUD**	**baseline**
Audio	0.512	0.525	**0.532**	0.501	0.483
Video	0.365	0.387	0.347	**0.393**	0.336
Fusion	0.542	**0.566**	0.536	0.560	0.516

以上結果都顯示出，使用主動式學習方法後對於準確率有一定程度的改善。這也可以解釋說，對於全部資料的方式進行訓練可能會受到離群值的影響，導致學習結果變差。在表一中，可以看到最大熵模型的效果較佳，而且可以發現聲音部分最好的結果在不確定性採樣法上，而影像則剛好相反。影像不論在何種分數標記(原始、正規化)、使用的不確定性採樣密度方法(標籤傳播、最大熵模型)，皆可以發現相關性結果都是正的。證明在加入密度這項因素下，其結果會變得更好，更可以反映出影像資料比較集中的特性，同時也可以推斷密度法可以有效的去除掉離群值。在未來擁有更大量的資料情況下，效果應該會更加明顯。聲音部分在標籤傳播標記方法中，反映出其基於此方法的情況下，對於密度法的效果比較顯著；而在最大熵模型中則有著比較差的效果。原因可能在於，標籤傳播的概念也是透過將標記的作一種傳播源來擴散出去，自然會受到密度的影響，所以標籤傳播方法對於密度法的向性比較好。同時也反映出聲音資料比較分散的特性，因為資料分布較為分散，所以先透過標籤傳播的方法對於距離上的分布作信息

熵的計算，之後加上密度法對於在高維角度的密度計算，從而強化了對於最後訓練支持向量回歸模型的資訊提供。

五、結論

藉這次實驗證明，主動式學習的方式的確有用，在方法 2 的效果可以有效的降低系統所需要的資料量，而且結果更好。大約只需要 70%的未標記資料即可有效的讓模型進行學習，這對於追求訓練資料數量減少的目標上有很大的幫助，也證明了上述的實驗流程設計有一定的可行性。透過有效的降低資料被標記的數量，來減少整個標記的成本與時間的耗費。最重要的是，使用主動式學習法可以有效的找出對於目前的模型來說，變異度最大的(對於模型資訊性最大)未標記資料來進行標記。

在未來，短期目標上會尋找不同的信息熵求取方式來應用在不確定性採樣法與密度法，並研究何時停止主動式學習方法可以得到最好的效果，而非透過結果回推實際的挑選個數；同時也研究其他的主動式學習方式來強化整體架構。中期目標上，會對於主動式學習過程進行簡化，並應用在其他的資料庫。遠程目標裡，希望能夠設計出一個主動式學習系統，在完全不需要計算最後的相關性的基礎上，透過加入其他的方法讓我們可以統整出一種參數，並依照此參數來篩選整個未標記的資料庫予以專家標記，藉此減少大量標記時產生的成本耗費，並獲得更佳的模型訓練成果，以提供更有幫助的行為訊息來輔助專家做出評判。

參考文獻

[1] L. Baraldi, F. Paci, G. Serra, L. Benini, and R. Cucchiara. Gesture recognition in ego-centric videos using dense trajectories and hand segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 688-693, 2014.

[2] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra. A maximum entropy approach to natural language processing. Computational linguistics, 22(1):39-71, 1996.

[3] Shan-Wen Hsiao, Hung-Ching Sun, Ming-Chuan Hsieh, Ming-Hsueh Tsai, Hsin-Chih Lin, Chi-Chun Lee: A multimodal approach for automatic assessment of school principals' oral presentation during pre-service training program. INTERSPEECH 2015: 2529-2533.

[4] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM international conference on Multimedia, pages 1459-1462. ACM, 2010.

[5] H. D. Kim, C. Zhai, and J. Han. Aggregation of multiple judgments for evaluating ordered lists. In Advances in information retrieval, pages 166-178. Springer, 2010.

[6] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan. Emotion recognition using a hierarchical binary decision tree approach. Speech Communication, 53(9):1162-1171, 2011.

[7] I. Muslea, S. Minton, and C. A. Knoblock. Selective sampling with redundant views. In AAAI/IAAI, pages 621-626, 2000.

[8] S. Narayanan and P. G. Georgiou. Behavioral signal processing: Deriving human behavioral informatics from speech and language. Proceedings of the IEEE, 101(5):1203-1233, 2013.

[9] M. Prince. Does active learning work? a review of the research. Journal of engineering education, 93(3):223-231, 2004.

[10] D. S. Cheng, H. Salamin, P. Salvagnini, M. Cristani, A. Vinciarelli, and V. Murino, “Predicting online lecture ratings based on gesturing and vocal behavior,” Journal on Multimodal User Interfaces, vol. 8, no. 2,pp. 151–160, 2014.

[11] P. Salvagnini, H. Salamin, M. Cristani, A. Vinciarelli, and V. Murino. Learning how to teach from "videolectures": automatic prediction of lecture ratings based on teacher's nonverbal behavior. In Cognitive Infocommunications (CogInfoCom), 2012 IEEE 3rd International Conference on, pages 415-419. IEEE, 2012.

[12] G. Schohn and D. Cohn. Less is more: Active learning with support vector machines. In ICML, pages 839-846. Citeseer, 2000.

[13] B. Settles. Active learning literature survey. University of Wisconsin, Madison, 52(55-66):11, 2010.

[14] A. Tamrakar, S. Ali, Q. Yu, J. Liu, O. Javed, A. Divakaran, H. Cheng, and H. Sawhney. Evaluation of low-level features and their combinations for complex event detection in open source videos. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pages 3681-3688. IEEE, 2012.

[15] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pages 3169-3176. IEEE, 2011.

[16] J. Zhu, H. Wang, T. Yao, and B. K. Tsou. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, pages 1137-1144. Association for Computational Linguistics, 2008.

[17] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, Citeseer, 2002.

[18] F. Perronnin, J. S´anchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in Computer Vision–ECCV 2010. Springer, 2010, pp. 143–156.

[19] K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods." in BMVC, vol. 2, no. 4, 2011, pp. 8–19.

[20] Rousseeuw, Peter J., and Annick M. Leroy. *Robust regression and outlier detection*. Vol. 589. John Wiley & Sons, 2005.