The 2016 Conference on Computational Linguistics and Speech Processing
ROCLING 2016, pp. 244-254
© The Association for Computational Linguistics and Chinese Language Processing

標記對於類神經語音情緒辨識系統辨識效果之影響

Effects of Label in Neural Speech Emotion Recognition System

吳東翰 Tung-Han Wu
國立中山大學資訊工程學系
Department of Computer Science and Information Engineering
National Sun Yat-sen University
ajason6208@gmail.com

陳嘉平 Chia-Ping Chen
國立中山大學資訊工程學系
Department of Computer Science and Information Engineering
National Sun Yat-sen University
cpchen@mail.cse.nsysu.edu.tw

摘要

本研究主要目的是探討平衡訓練語料與不平衡訓練語料對於語音情緒辨識的影響。此外由於情緒標記的主觀性可能帶來誤差，因此在此論文中也探討在輸入狀態下資料標記錯誤與否對於系統未加權平均辨識率(Unweighted Accuracy, UA)之影響。實驗工作主要是設計一個類神經網路之語音情緒辨識系統，並使用 INTERSPEECH 2009 Emotional Challenge 中所釋出之 FAU-Aibo 情緒語料庫，以作為辨識率之基準。實驗結果顯示，在假設訓練語料正確標記時，資料平衡與資料不平衡時的未加權平均辨識率分別為 39.6% 與 34.6%；在容許訓練語料錯誤標記時，資料平衡與資料不平衡時的未加權平均辨識率分別為 41.8% 與 35.7%。因此在利用類神經系統作為辨識工具時，若能考慮訓練語料錯誤標記的因素，並適當的提供標記錯誤參數，系統之未加權平均辨識率將可以明顯改善。

關鍵詞：情緒辨識、情緒辨識資料庫、類神經網路、系統平均辨識率

1 緒論

人與人在互相交流時語言溝通就扮演了一個很重要的角色，人們會透過語言來互相交流訊息。基於此理念下開發出了許多人機介面 (Human-computer Iteration, HCI) 相關的產品，如早期的 IBM、Microsoft 系統的語音輸入等，能用語音取代傳統鍵盤打字，而近年來自動語音辨識 (Automatic Speech Recognition, ASR) 已經進步到能夠準確的辨識出語者的語音，不僅能轉換成文字，還能進一步轉化成聲音來與使用者互動，例如 Google 的語音搜尋以及蘋果公司的 Siri 為人們所熟知的代表作之一。人與人在溝通時除了考慮語言在字面上本身所表示的意思之外，還會去考慮當時語者說話時的情緒，但是現今即使是功能完善的 Siri 也無法針對人類的情緒來給出最好的回應，因此在自動情感語音辨識 (Speech Emotion Recognition, SER) 上仍有相當大的發展性。此外，在 2014 年 4 月由日本軟銀集團以及法國 Aldebaran Robotics 公司共同研發的機器人 Pepper 就是一台可以表達情緒以及辨識人類情緒的人型機器人，並且已經應用於日本軟銀集團旗下的分行進行服務，可見情緒辨識也被廣泛應用於各種不同的產品上。

早在 1980 年代，研究發現在情緒上存在著普遍能夠識別出的特徵，而這些特徵主要與情緒的發聲模式有關，因此開創了使用聲學統計特徵進行情感分類的先河 [1]。情感語音辨識屬於情感運算的一部分，在 1997 年, Picard 所著一書為最早提到情感運算辨識的起頭，此書定義了情感計算 [2]，他從一個資訊工程研究者的角度來說明情感運算的應用以及重要性。接下來的幾年內學者對於情感上的分類做了許多研究，並探討情緒對於生理及心理所產生的影響以及變化。本篇論文所使用的 FAU-Aibo [3] [4] 資料庫為 INTERSPEECH 2009 Emotional Challenge 中所指定的基準情緒語料庫，該語料庫改善了傳統語料庫所欠缺的部分。例如：資料量太小、情感的表達不夠自然、實驗結果和他人所做的研究無法有一個好的比較基準等，基於以上原因我們選用了該語料庫做為我們實驗的依據。

現今世界中許多分類的問題通常並不是單純的 0 或 1 的問題， Tsoumakas 等人統整了近幾年來複數標記的需求不斷應用於許多大型的應用 [5]，例如：音樂分類 [6]、語意分類 [7] 等，而我們所研究的情緒分類也屬於複數標記的類型。在分辨人類的情緒時，我們常常不能百分之百的篤定結果，自動情感語音辨識 (Speech Emotion Recognition, SER) 其中一項困難的點在於情緒的分類是一種很主觀的資訊，常常同一句話在不同的聆聽者下會有不同的結果，而為了解決上述這種情況我們嘗試在類神經

網路 (Neural Network) 架構下透過模擬標記錯誤的方法來消除主觀意識所造成的誤差，藉此來改善情緒辨識下的辨識率。

本論文主要分為四個部分，第一部分為緒論；第二部分為研究方法，首先先介紹了實驗時的輸入特徵，接下來是針對資料進行語者正規化的前處理以及偏斜資料進行平衡，最後再加入我們考慮資料標記錯誤的方法來完成實驗，第三部分為實驗結果分析，第四部分為對整體實驗的結論以及未來展望。

2 研究方法

2.1 基準特徵集

本論文所使用的特徵為 INTERSPEECH 2009 Emotional Challenge [3] 所採用的基準特徵集如表 1 所示，包含 16 個低階參數(Low-Level descriptors, LLDs)與其 delta 和 12 個泛函，此特徵集於隱藏式馬可夫模型與線性支持向量機分類器上得到了該挑戰的基準實驗結果分別為 35.9% 以及 38.2%。在聲學特徵上所採用了包含聲韻、頻譜、聲音能量等特徵，所選擇的16個低階參數為過零率(Zero Cross Rate, ZCR)、能量方均根(Root Mean Square, RMS)、音調頻率(pitch frequency)、諧音噪音比(Harmonics to Noise Ratio, HNR)、梅爾倒頻譜係數(Mel Frequency Cepstral Coefficients, MFCCs) (1-12維) 等；12 個泛函為平均值(mean)、標準差(standard deviation)、峰度(kurtosis)和偏移態(skewness)、最大與最小值、相對位置(relative position)與範圍(range)以及另外兩個線性迴歸係數(linear regression coefficients)及其均方差(Mean Square Error, MSE)。最後經過一階係數差並經由12個泛函計算後，最後特徵即包含了 16×2×12=384 個特徵參數。

表 1: 基準特徵集

LLDs	Functionals
RMS Energy	mean
ZCR	standard devation
MFCC 1-12	kurtosis, skewness
HNR	extrmes:value, rel.position, range
F0	linear regression:offset, slope, MSE

2.2 語者正規化(CSHE)

當我們在擷取特徵時，可能會因為語者的不同而產生差異性，因此我們採用直方圖均衡法(Histogram Equalization, HE)做為我們的語者正規化方法(Cross-speaker histogram equalization, CSHE)。基於 Chiou 的做法 [8]，我們將直方圖轉換的公式 (1) 定義如下，其中 $X(x)$ 代表原始特徵分佈， $Y(y)$ 代表目標分佈， p 代表原始的特徵值， q 代表轉換過的特徵值：

$$\int_{x=-\infty}^{p} X(x)dx = \int_{y=-\infty}^{q} Y(y)dy \tag{1}$$

對於兩個分佈 $X(x)$ 與 $Y(y)$ ，我們的目標是將 $X(x)$ 轉換到 $Y(y)$。我們分別計算 $X(x)$ 與 $Y(y)$ 兩者的累積分布函數(Cumulative Distribution Function, CDF) ，再將原始特徵值轉換到目標分布上。因此我們會將多個訓練語者視為一個虛擬語者，在所有資料中我們就可以得到此虛擬語者的累積分佈函數 $c_Y(y)$ 做為目標分佈，而對於每一位訓練語者的資料

$$D_x = \{x_1, ..., x_n\} \tag{2}$$

我們也可以由 D_x 算出對應的累積分布函數 $c_X(x)$，最後再經由直方圖均衡法求得對應的特徵值。圖 1 表示正規化的流程。

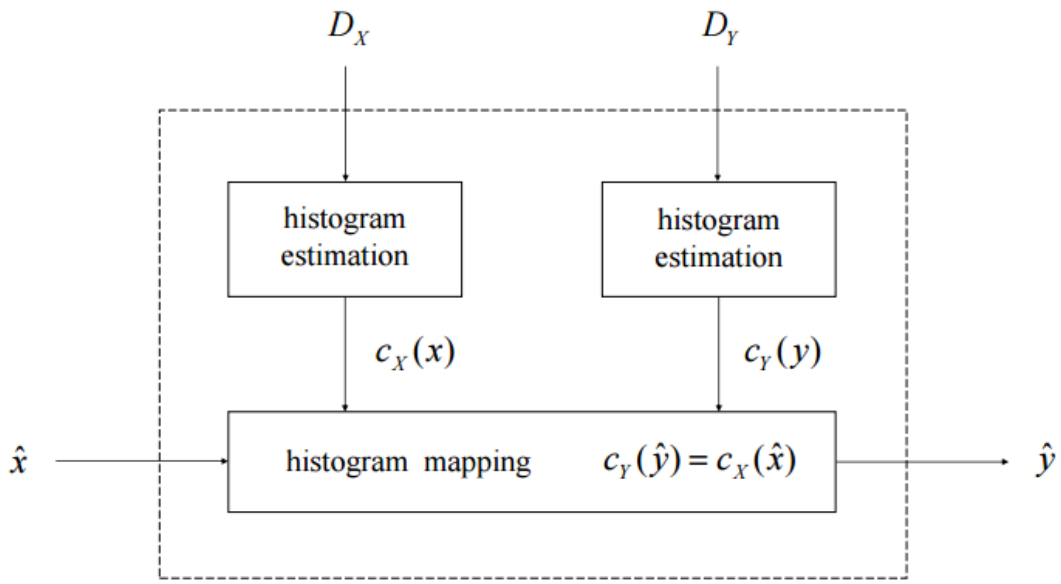

圖 1: 語者正規化流程圖

2.3 資料平衡化

本實驗所採用的資料平衡化的方法為少數類過採樣技術(Synthetic Minority Oversampling Technique, SMOTE) [9] 來對資料進行平衡。首先我們隨機選出一個少數類別之資料樣本 X_i ，並透過最近鄰居法(K-Nearest Neighbors, KNN)來產生 K 個鄰近的樣本，並從這K個樣本中隨機選出其中一筆資料 X_i' 後計算出兩樣本之間的差值，最後再隨機乘上一個介於 0 到 1 之間的數值來產生新的樣本，其公式 (3) 如下：

$$X_{new} = X_i + (X_i' - X_i) \times \delta \tag{3}$$

2.4 標記錯誤模擬

首先我們先定義類神經網路架構如圖 2：

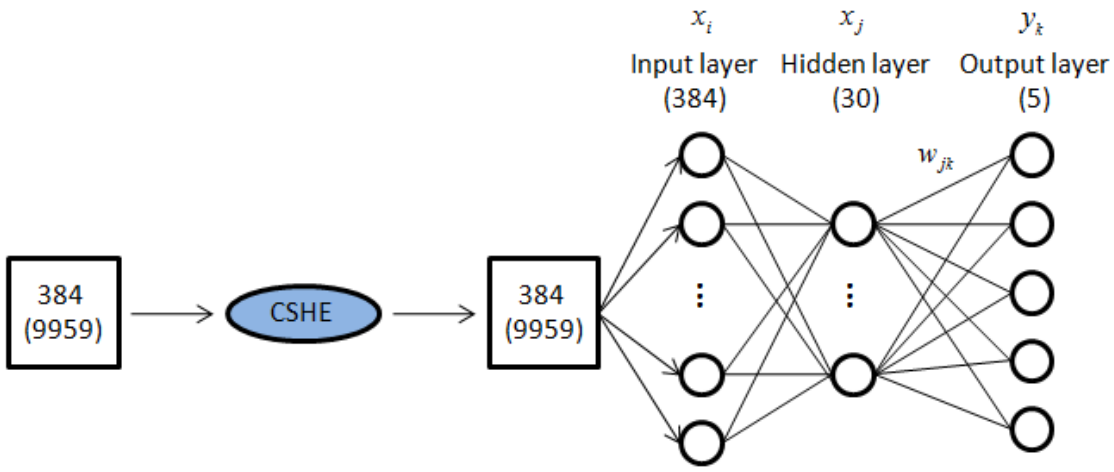

圖 2: 類神經網路架構

y_k 代表隱藏層到輸出層的激活函數輸出值(activation value)， a_k 代表神經元輸入的權重總合，如式 (4) 所示。此外，在多類的分類問題上，我們使用 softmax 函數做為輸出層的激活函數，對於一個 K 類的分類問題，其第 k 個神經元的輸出值為式 (5) 所示：

$$a_k = \sum_{j=1} x_j w_{jk} \tag{4}$$

$$y_k = \frac{\exp(a_k)}{\sum_{j=1}^{k} \exp(a_j)} \tag{5}$$

接下來假設有 K 個類別與 N 筆訓練資料的資料集 D。我們單看其中一筆資料 (t,x)，其資料似然度(likelihood)為式 (6) :

$$P(t|x,W)=\prod_{k=1}^{K} y_k(x,W)^{t_k} \tag{6}$$

而對於整個資料集 D 的資料似然度(likelihood)為式 (7) :

$$P(D|W)=\prod_{n=1}^{N}\prod_{k=1}^{K} y_{nk}(X,W)^{t_{nk}} \tag{7}$$

接下來我們對式 (7) 計算其負對數(negative logarithm) 之後，我們可以得到下式 (8)，其中 t 就是我們的正確標記，而 y 是我們預測出來的結果

$$E(W)=-\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\ \log(y_{nk})=\sum_{n} E_n(W) \tag{8}$$

此為交叉熵(cross-entropy)成本函數，其中t_n 為目標輸出值，y_n 為實際輸出值，而 $E_n(W)$ 則為該目標輸出值和實際輸出值之間的交叉熵。

$$y_{nk}=y_k(x_n,W) \tag{9}$$

最後我們要模擬錯誤標記時的情況，所以我們會假設訓練語料標記錯誤的機率是 ε，其中標記錯誤的方式又分為兩種類型進行探討，並將原本是以 One-Hot 來標記的 t_n 根據不同標記錯誤的類型來進行變換，其中 One-Hot 表示一個列向量中只有一個元素的值為 1 其餘的元素值為 0。

2.4.1 非彈性的標記錯誤(Hard Label Error)

t_n 會有 ε 的機率會隨機分到其他的類別上，並且依舊維持 One-Hot 的型式，這邊為了表示變化的情況，所以我們假設 ε =100% 的情況下，t 所產生的變化如下式 (10) ，矩陣的列表示第 n 筆資料，行表示第 k 個類別的機率分佈 :

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \tag{10}$$

2.4.2 彈性的標記錯誤(Soft Label Error)

t_n 會有 ε 的機率標記錯誤 ，但是此時的 t_n 不需要標記成 One-Hot 的型式，而是將錯誤率 ε 當成類別的錯誤率，再將 $\frac{\varepsilon}{4}$ 的機率平均分給其他四個類別。這邊為了表示變化的情況，所以我們假設 $\varepsilon = 40\%$ 的情況下我們所產生的變化如下式 (11) :

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0.1 & 0.1 & 0.6 & 0.1 & 0.1 \\ 0.6 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.6 & 0.1 & 0.1 & 0.1 \end{bmatrix} \tag{11}$$

經由對 t_n 進行改變之後，式 (8) 的 W 產生微小變化時，E 也會跟著產生微小變化，透過偏微分以及連鎖率(Chain Rule)可以將最上層(Top Layer)權重的梯度寫成下式 (12) :

$$\frac{\partial E_n}{\partial w_{jk}} = \frac{\partial E_n}{\partial y_{nk}} \frac{\partial y_{nk}}{\partial a_{nk}} \frac{\partial a_{nk}}{\partial w_{jk}} \tag{12}$$

接下來我們會針對式 (12) 中的每一項做偏微分來求得 w 變動時 E 的 變化量，計算過程如下式 (13) 至式 (15) 所示 ：

$$\frac{\partial E_n}{\partial y_{nk}} = -\frac{t_{nk}}{y_{nk}} \tag{13}$$

$$\frac{\partial y_{nk}}{\partial a_{nj}} = \begin{cases} y_{nk}(1-y_{nj}) & k = j \\ -y_{nk}y_{nj} & k \neq j \end{cases} \tag{14}$$

$$\frac{\partial E_n}{\partial a_{nk}} = \sum_{j}^{nclass} \frac{\partial E_n}{\partial y_{nj}} \frac{\partial y_{nj}}{\partial a_{nk}} = y_{nk} - t_{nk} \tag{15}$$

因此我們可以得到最上層(Top Layer)權重的梯度為式 (16) :

$$\frac{\partial E_n}{\partial w_{jk}} = \sum_{k} \frac{\partial E_n}{\partial a_{nk}} \frac{\partial a_{nk}}{\partial w_{jk}} = (y_{nk} - t_{nk}) x_j \tag{16}$$

而我們以 j 表示第二層的神經元，其權重的梯度計算為式 (17)

$$\frac{\partial E_n}{\partial w_{ij}} = \sum_{k} \frac{\partial E_n}{\partial a_{nj}} \frac{\partial a_{nj}}{\partial w_{ij}} = \sum_{k} (y_{nk} - t_{nk})(w_{jk})(y_{nk}(1-y_{nk}))x_i \tag{17}$$

最後將權重所產生的變化更新回舊的權重上來完成考慮標記錯誤下的類神經網路訓練。

3 實驗結果與討論

實驗中所使用的資料庫為 FAU-Aibo 情緒語料庫，分為生氣(Anger)、强調(Emphatic)、中性(Neutral)、正面(Positive)、其他(Rest)的五類情緒語料，由於五類情緒之資料不平衡問題，訓練資料與測試資料中各類別的資料分佈差異極大，因此辨識結果主要採用未加權平均辨識率(UA)作為比較基準，計算方法為式 (18) 所示，其中，A_{ij} 為類別 i 被分到類別 j 的資料數，K 為總類別數。

$$UA = \frac{1}{K}\sum_{i=1}^{K}\frac{A_{ii}}{\sum_{j=1}^{K}A_{ij}} \tag{18}$$

實驗所使用的工具為 Google 所釋出的開源軟體 Tensorflow [10] [11]，以下數據均為實驗五次後取平均後所產生的結果。實驗中使用的神經網路架構均為單層類神經網路架構：輸入層為 384 個神經元、隱藏層為 30 個神經元、輸出層為 5 個神經元，迭代的次數為 60，每一批次設定為 30，學習率固定為 0.3。

3.1 標記正確

我們先在標記正確的情況下分別對有做資料平衡與未做資料平衡的資料下做出一個基準實驗(表 2)，並與標記錯誤時的實驗進行比較。

表 2: 標記正確情況下，資料平衡與不平衡的結果

Process	平均辨識率(UA)
有做資料平衡	39.6
未做資料平衡	34.6

3.2 標記錯誤

3.2.1 非彈性標記錯誤

(1) 有做資料平衡

表 3: 資料平衡下，不同 ε 的比較

標記錯誤率	平均辨識率(UA)
$\varepsilon = 10\%$	40.6
$\varepsilon = 20\%$	41.2
$\varepsilon = 30\%$	41.3

(2) 未做資料平衡

表 4: 資料未平衡下，不同 ε 的比較

標記錯誤率	平均辨識率(UA)
$\varepsilon = 10\%$	35.0
$\varepsilon = 20\%$	34.5
$\varepsilon = 30\%$	34.7

我們針對表 3、表 4 先進行初步觀察，可以發現到非彈性標記錯誤的情況下去考慮標記錯誤時，並不會對平均辨識率有顯著的影響，其原因在於我們當初想要考慮標記錯誤的目的是認為每一個語者所講出來情緒可能會因為不同聆聽者而產生不同的結果，但是當我們使用 SMOTE 來解決資料不平衡時，我們同時也正在降低不同類別的相似程度，因為訓練資料的提升有助於讓類神經網路更加完整，縱使有標記錯誤，但是對於整體的網路卻不會有太大的影響。此外當我們使用非彈性的標記錯誤來進行模擬時，資料有 ε 機率隨機分到其他類，該筆資料也可能只有 20% 機率分到正確的類別上，除此之外語者可能當下所想表達的情緒可能不只有一種，例如 : 語者生氣的時候說出來的情緒可能包含憤怒以及難過，所以基於上述兩種原因的影響，我們進一步嘗試了彈性的標記錯誤來改善此問題。

3.2.2 彈性標記錯誤

從彈性的標記錯誤所做出來的實驗結果來看，不管是資料平衡與否，平均辨識率均有上升的趨勢，尤其是在資料平衡的時候再加入彈性標記錯誤的方法時，比起單純只做資料平衡平均辨識率上升了 2.2%。由此實驗結果 (表 5 6) 可以說明情緒標記確實存在著主觀性，也驗證了在考慮標記錯誤的情況下來實做一個類神經網路系統是可以有效

的提升整體的平均辨識率。

(1) 有做資料平衡

表 5: 資料平衡下，不同 ε 的比較

標記錯誤率	平均辨識率(UA)
$\varepsilon = 10\%$	40.8
$\varepsilon = 20\%$	41.5
$\varepsilon = 30\%$	41.8

(2) 未做資料平衡

表 6: 資料未平衡下，不同 ε 的比較

標記錯誤率	平均辨識率(UA)
$\varepsilon = 10\%$	34.9
$\varepsilon = 20\%$	35.1
$\varepsilon = 30\%$	35.7

4 結論與未來展望

在類神經系統中，訓練資料對於系統之平均辨識率應有一定之影響。因此在利用類神經系統作為辨識工具時，若能考慮訓練資料錯誤標記的因素，並適當的提供標記錯誤參數，系統之平均辨識率將可以明顯改善。根據本篇論文所實驗的結果，可以發現在類神經網路下處理特定且不確定性很高的資料集的時候，可以透過考慮錯誤標記的方式來有效的提升平均辨識率。我們的方法在考慮彈性的錯誤標記時，是將 $\frac{1}{4}\varepsilon$ 的機率平均分給其他四個類別，但是這種做法沒有去考慮語者說出來的話語中，各個情緒所佔的比例，這一點是將來研究時必須要再進一步探討的問題，是否能夠找出一個方法可以有系統的調整這些參數來使平均辨識率可以更有效的提升。另外當我們考慮錯誤標記的方法時，如果是運用在資料已經很完整且辨識度很高的資料集上就不適用於此

方法，例如：手寫辨識，未來我們希望可以在更複雜的模型上來考慮標記錯誤的情況以面對各種不同類型的資料，並將其應用在一個更完整的系統上。

參考文獻

[1] R. Van Bezooijen, S. A. Otto, and T. A. Heenan, Recognition of vocal expressions of emotion a three-nation study to identify universal characteristics, Journal of Cross Cultural Psychology, Vol. 14, no. 4, 387-406, 1983. 2011.

[2] R. W. Picard, Affective computing. MIT Press, 1997.

[3] S. Steidl, Automatic classification of emotion related user states in spontaneous children's speech. University of Erlangen-Nuremberg Erlangen, Germany, 2009.

[4] B. Schuller, S. Steidl, and A. Batliner, The INTERSPEECH 2009 Emotion Challenge, Proceedings of the Interspeech 2009, Brighton, UK, 312–315, 2009

[5] G. Tsoumakas, K. Ioannis and V. Ioannis, . A Review of Multi-Label Classification Methods, Proceedings of the 2nd ADBIS Workshop on Data Mining and Knowledge Discovery, 99-109, 2006.

[6] T. Li and O. Mitsunori. Detecting emotion in music, ISMIR. Vol. 3. 2003.

[7] M. Boutella, X. Shena, J. Luob and C. Browna1, Multi-label semantic scene classification. technical report, department of computer Science, University of Rochester, 2003.

[8] B. -C. Chiou, Cross-lingual automatic speech emotion recognition, Master's thesis, National Sun Yat-sen University, 2014.

[9] N. V. Chawla, K. Y. Bowyer, L. O. Hall and W. P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, Journal of artificial intelligence research, Vol. 16, 321-357, 2002.

[10] M. Abadi. TensorFlow: Large-scale machine learning on heterogeneous systems, Software available from tensorflow, 2015.

[11] M. Abadi . Tensorflow: Large-scale machine learning on heterogeneous distributed systems. 2016.