

以統計特性進行心跳序列之叢集處理

黃于珍

元智大學資管系碩士生
s956209@mail.yzu.edu.tw

林翰俊

元智大學資管系碩士生
s946204@mail.yzu.edu.tw

林志麟

元智大學資管系助理教授
jun@saturn.yzu.edu.tw

摘要

本研究提出使用心率變異度(Heart Rate Variability, HRV)的八個統計值(平均值、SDNN、SDNN Index、SDANN、RMSSD、CV、pNN50 與 pNN20)來計算兩 RR Interval 時間序列間的距離,進而根據此種距離計算的結果進行 RR Interval 時間序列的叢集處理。透過多組不同類型的 RR Interval 時間序列資料集的實驗結果顯示,此種方式所產生的叢集結果能有效地區隔出資料集內不同類型的 RR Interval 時間序列。

關鍵詞: 心率變異度、叢集分析。

Abstract

This study proposes a method to cluster RR Interval time series. First, the distance of any two RR Interval time series is defined based on eight statistic features of Heart Rate Variability (including Mean, SDNN, SDNN Index, SDANN, RMSSD, CV, pNN50 and pNN20). Then, RR Interval time series are clustered based on this distance. Experimental results with various datasets indicate that this method can effectively distinguish different types of RR Interval time series.

Keywords: Heart Rate Variability, Cluster Analysis.

1. 前言

老化與疾病常會導致副交感神經活性的下降。以冠狀動脈疾病與急性心肌梗塞的患者為例,副交感神經的活性多會下降[4,8,9,11,12]。心率變異度(Heart Rate Variability, HRV)是因自律神經系統(autonomic nervous system)活動而產生的波動,可以反映交感神經系統(sympathetic)與副交感神經系統(parasympathetic)間的平衡狀態[13]。近年來,心率變異度的臨床價值逐漸被顯現出來。歐洲心臟病學會(The European Society of

Cardiology)與北美節律暨電生醫學會(The North American Society of Pacing and Electrophysiology)在1996年共同提出心率變異度可有效地預測心律不齊併發症(arrhythmic complications),例如,心室心搏過速(ventricular tachycardia, VT)、心肌梗塞(myocardial infarction, MI)、糖尿病(Diabetes Mellitus, DM)與高血壓(Hypertension)等疾病[13]。

心臟跳動的頻率就是心律(Heart Rate),而心率變異度就是每次心跳時間間隔上的變化程度。人體心臟的跳動速度並非以一個恆久不變的速度持續跳動著,即使人處於穩定的狀態下,心臟跳動的速度也有某一程度的變化。此外,心理方面的壓力也會使心跳速度產生變化。因此如何有效分析心律便成為一個重要的課題。

心律量測最常用的方式就是計算兩個相鄰 QRS 波中第一個 R 波與第二個 R 波的時間間隔,這樣的時間間隔稱之為 RRI(RR Interval),也稱為 NN interval。之後,將所有 RRI 排成一時間序列以表現心律隨時間的變化。目前常見的時間序列分析方式可分為時域分析(Time domain analysis)與頻域分析(Frequency domain analysis)兩大類。頻域分析的方法多利用 Fourier Transform 擷取時間序列的若干個主要組成頻率與強度來代表此序列的主要特性。時域分析的方法則多利用一些統計方法來擷取此時間序列的主要特性。對 RRI 時間序列而言,時域分析方法常擷取的主要特性包括平均值(mean of the NN interval)、SDNN(standard deviation of the NN interval)、SDANN(standard deviation of the average NN interval)、SDNN Index(standard deviation index of the NN interval)、RMSSD(square root of the mean squared differences of successive NN intervals)、CV(coefficient of variation of the NN interval)、pNN50(the ratio of the number of interval differences of successive NN intervals greater than 50 ms and the total number of NN intervals)與文獻[7]中提及的 pNN20 等等。

一般而言，隨著疾病的不同，所量測的 RRI 序列長短會有所不同。例如，鬱血性心臟衰竭(Congestive Heart Failure, CHF)患者通常需要較長時間的 RRI 時間序列資料來做進一步分析，而心房纖維顫動(Atrial Fibrillation, AF)患者則通常只量測其發病前後數分鐘的 RRI 資料。為了能一併分析不同長度的 RRI 序列資料，Yang 等人[1]於 2003 年提出一套語言學的方法（此方法又於 2007 年由 Peng 等人[2]提出改善的方式），用來計算兩 RRI 序列間的距離。此作法先根據相連 RRI 是變長還是變短，將整個 RRI 序列轉換成 0-1 的序列。之後，以將 0-1 序列的每 8 個連續的值視為一個字元；例如，第一個字元是由第一個 0 或 1 到第八個 0 或 1 組成，而第二個字元是由第二個 0 或 1 到第九個 0 或 1 組成，依此類推。最後，將各字元在各序列內的出現次數進行排名，然後依兩序列間字元排名的差異來計算出兩序列間的距離。根據此種距離的計算方式，Peng 等人將 RRI 序列作分群處理，希望能達到同類型疾病的患者落入同一叢集的效果。

本研究的主要目的在於改善 RRI 序列的分群結果。利用心率變異度的相關統計欄位作為代表一 RRI 序列的主要屬性，而序列間的距離計算方式採用曼哈頓距離 (Manhattan distance)。透過多組不同的 RRI 序列資料集的實驗，結果顯示，無論對短時間的 RRI 序列資料集、長時間的 RRI 序列資料集或混合長短時間的 RRI 序列資料集，本作法的叢集結果均優於文獻 [2] 的結果。

2. 研究方法

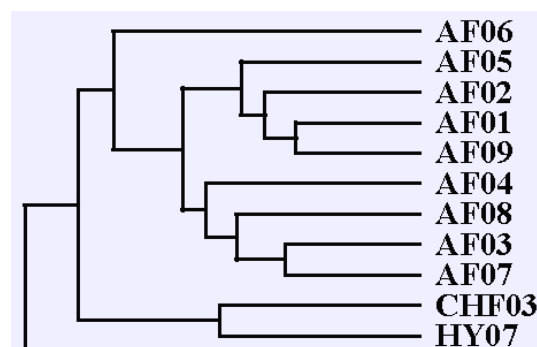
本節提出本研究所採取的 RRI 序列處理方式。以下分別就序列間距離計算方式、叢集產生方式、以及叢集品質評估方式加以說明。

假設 $\{x_0, x_1, x_2, \dots, x_n\}$ 為一 RRI 時間序列，其中 x_i 代表第 i 個 RRI。首先，吾人先計算出該 RRI 時間序列的八種統計值(mean、SDNN、SDNN Index、SDANN、RMSSD、CV、pNN50、pNN20)作為代表該 RRI 序列的屬性。有關此八種統計值之計算方式請參考文獻[13]。之後，兩 RRI 時間序列的距離則定義為這兩序列間此八個屬性的曼哈頓距離。註：當 RRI 時間序列過短以致無法計算 SDANN 與 SDNN Index 時，我們只根據剩下的六個屬性來計算距離。

要叢集 n 筆 RRI 時間序列，吾人先採上述方式計算出任兩序列間的距離，產生一個 $n \times n$ 的距離矩陣 $D=[d_{ij}]$ ，其中 d_{ij} 為第 i 筆與第 j 筆序列

間的距離，而 $i, j \in [1, n]$ 。之後，將距離矩陣 D 用 Phylip[6] 當中的 Distance Matrix Methods 之 neighbor 程式進行計算並繪製演化樹 (phylogenetic Tree)[17]；此一步驟與文獻[1,2]的作法相同，但是文獻[1,2]所輸入的距離矩陣之兩兩序列距離計算方式與本作法不同。所產生的演化樹可視為階層式叢集的結果。

本研究所使用的每一筆 RRI 時間序列資料均有一個額外的類別型欄位，用以存放該序列所屬之類別值。為評估階層式叢集結果是否能有效地區別出當該類別型欄位值不同時的 RRI 時間序列，吾人先將階層式叢集的結果用 TreeView[10] 這套繪樹軟體繪製出一棵有根的演化樹 (rooted phylogenetic Tree)。然後，根據此演化樹，分別計算出當這些序列分為不同叢集個數時，對此一類別欄位的預測錯誤率。預測錯誤率計算方式說明如下。一個叢集對該類別欄位的預測值為屬於該叢集的 RRI 時間序列中，最常出現的類別欄位值；而一個叢集對此一類別欄位的預測錯誤次數即為屬於該叢集的 RRI 時間序列中，類別欄位值與其預測值不同的序列個數。例如：若將圖一中的 11 筆資料歸屬到同一個叢集，因為其中有九筆的類別為 AF，一筆為 CHF，一筆為 HY，所以我們將此叢集對該類別欄位的預測值視為 AF，且有兩個錯誤（一筆 CHF 與一筆 HY）。針對某一叢集個數之叢集結果的錯誤率為各叢集之錯誤個數總和除以時間序列總個數。



圖一 單一 cluster 範例

為比較本研究與文獻[2]作法的性能差異，我們也建構了 Peng 等人[2]所採用的距離矩陣及階層式叢集結果。然後，針對我們的作法與文獻[2]作法所產生的兩種階層式叢集結果，比較在不同叢集個數下的預測錯誤率。

3. 實驗結果

由於時間序列的長短可能會影響到兩方

法的叢集效果，因此本研究採用多組長度不同的資料集進行實驗。以下分別就不同實驗所使用的資料集與實驗結果做說明。

3.1 短 RRI 序列集實驗結果

在短 RRI 序列集實驗部份，我們分別採用 AFTDB 與 AFPDB 兩組資料集做實驗。AFTDB 資料集係擷取自 The Atrial Fibrillation Termination Challenge Database [14]的訓練資料。AFTDB 資料集主要在紀錄心房顫動(Atrial Fibrillation, AF)疾病患者的心跳資料，其中的每一筆時間序列長度為一分鐘，共有 30 筆時間序列。AFTDB 資料集內的時間序列分為 3 個類別，每個類別各有 10 筆時間序列。這三個類別分別為：

- 類別N：在此時間序列之後的至少一個小時以上都沒有發生過任何AF終止現象。
- 組別S：在此時間序列之後一分鐘即發生AF終止現象。
- 組別T：在此時間序列之後一秒鐘內就馬上發生AF終止現象。

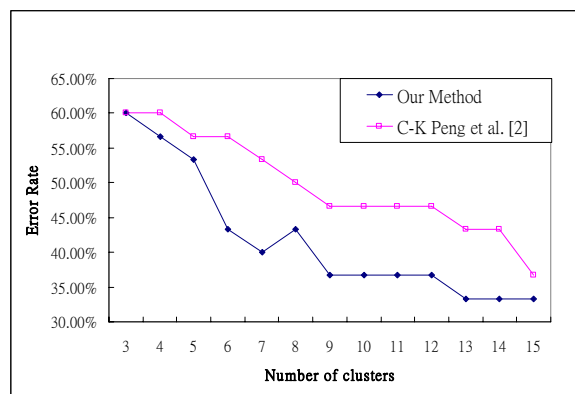
AFPDB 資料集係擷取自 The Paroxysmal Atrial Fibrillation Prediction Challenge Database[16]的訓練資料。AFPDB 資料集主要在紀錄陣發性心房顫動(Paroxysmal Atrial Fibrillation, PAF)疾病患者的心跳資料，其中的每一筆時間序列長度為 30 分鐘，共有 100 筆時間序列。AFPDB 資料集內的時間序列分為下面三個類別：

- 類別O：在此時間序列前後 45 分鐘內都沒有發生過任何PAF現象。此類序列有 25 筆。
- 組別E：在此時間序列之後馬上就發生PAF現象。此類序列有 25 筆。
- 組別N：無PAF疾病的健康受測者資料。此類序列有 50 筆。

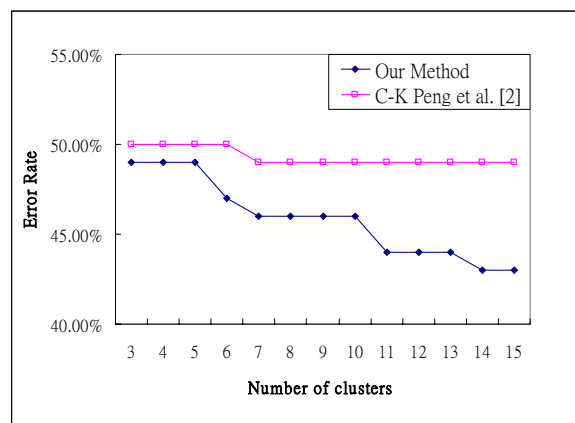
圖二為 AFTDB 資料集的實驗結果。當叢集個數為 3 時，我們的作法與文獻[2]的作法有相同的錯誤率。然而，隨著叢集個數增加，兩者間的錯誤率差距也逐漸加大。直到叢集個數為 7 時，兩者的錯誤率差距最大。此時，我們的作法之錯誤率為 40.00%，而文獻[2]的作法之錯誤率為 53.33%，兩者相差達 13.33%。雖然當 cluster 數量為 8 時，我們的作法之錯誤率有上升，但仍較文獻[2]的作法所產生之錯誤率低。

圖三為 AFPDB 資料集的實驗結果。從圖三

中可以發現我們的作法仍比文獻[2]的作法有較低的錯誤率。此外，隨著叢集個數增加，文獻[2]的作法對錯誤率的降低較不明顯，只在叢集個數由 6 到 7 時可以看出錯誤率的降低。反觀我們的作法，隨著叢集個數增加，錯誤率降低的趨勢較為明顯。因此，當叢集個數越多時，兩者的錯誤率也相差越大。



圖二 AFTDB 資料集之叢集結果的錯誤率

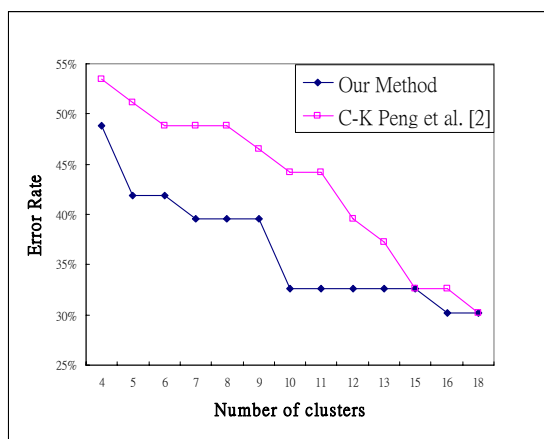


圖三 AFPDB 資料集之叢集結果的錯誤率

3.2 長 RRI 序列集實驗結果

在長 RRI 序列集實驗部份，我們採用 CHF 資料集做實驗。CHF 資料集內的時間序列分為 4 個類別，分別為 I、II、III、III-IV。CHF 資料集係擷取自 Congestive Heart Failure RR Interval Database [3]的 29 筆序列（類別 I、II、III 分別各有 4、8、17 筆）與 BIDMC Congestive Heart Failure Database[15]的 15 筆序列（均為類別 IV）。每一筆時間序列長度為 20 小時，共有 44 筆時間序列。

圖四為 CHF 資料集的實驗結果。從圖四可知我們的作法之錯誤率比文獻[2]的作法低約 4%到 12%。在叢集數為 11 時兩者差異最大；叢集數為十五時，兩者錯誤率相同。這個結果與 AFTDB 資料集的實驗結果相當類似。

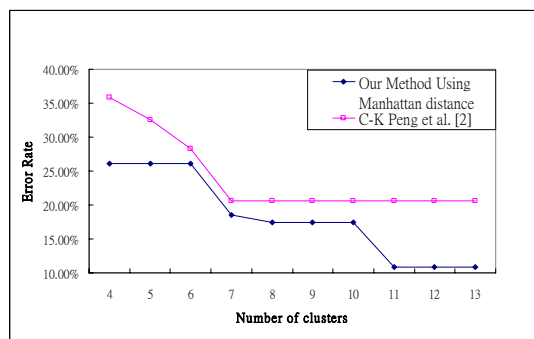


圖四 CHF 資料集之叢集結果的錯誤率

3.3 混合長短 RRI 序列集實驗結果

在混合長短 RRI 序列集實驗部份，我們採用文獻[1,2]所使用的一組資料集做實驗。此資料集內的時間序列分為 4 個類別，分別為 AF (Atrial Fibrillation)、CHF (Contestive Heart Failure)、HE (Health Elder)、HY (Health Young)。其中，CHF 類別的時間序列有 43 筆，資料來源與前述的 CHF 資料集相同；HE 與 HY 這兩個類別的時間序列來自於[5]，兩類別各有 20 筆 RRI 序列，每筆紀錄長度為兩小時；AF 類別的時間序列有 9 筆，每筆紀錄的長度為兩個小時，資料來源係由文獻[1]作者所提供。

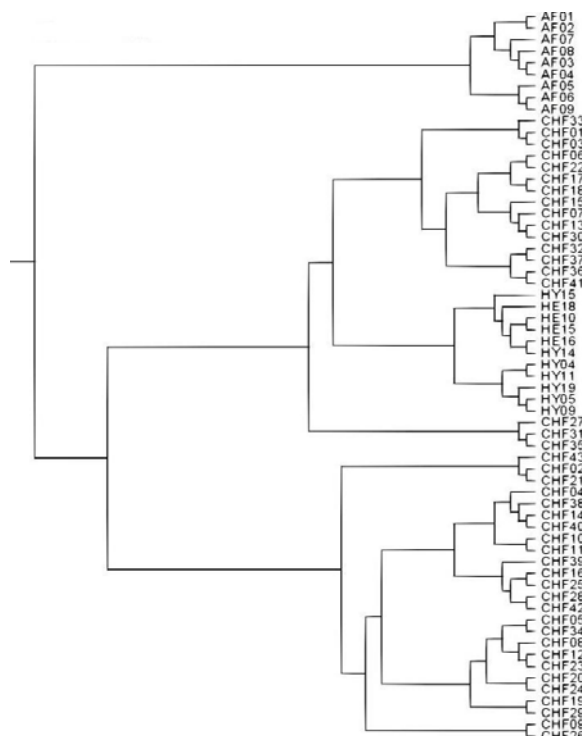
圖五為此資料集的實驗結果。從圖五可發現於在叢集個數為 4 時，我們的作法和文獻[2]的作法在錯誤率方面就已有明顯的差異。當叢集個數超過 7 以後，文獻[2]的作法之錯誤率便沒有很明顯的改善。反觀我們的作法，錯誤率隨著叢集個數的增加呈現階梯式的遞減。這個結果與 AFPDB 資料集的實驗結果相當類似。



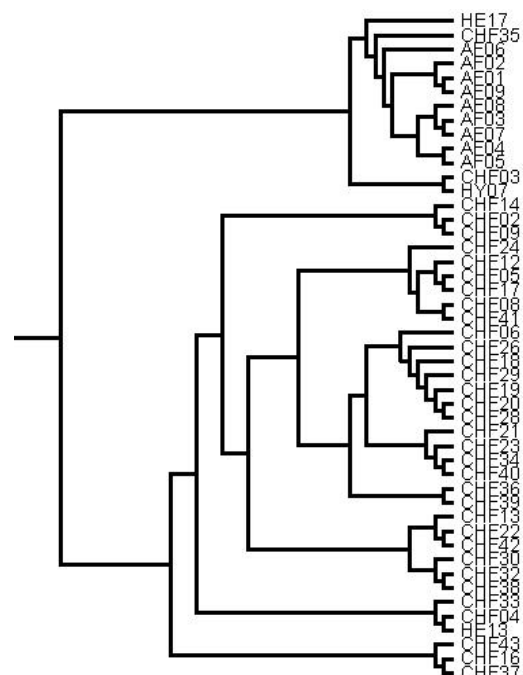
圖五 混合長短 RRI 序列之叢集結果錯誤率

圖六為我們的作法所產生之樹狀圖的部分畫面。值得注意的是，當叢集個數為 7 時，雖然我們的作法仍有 18% 左右的錯誤率，然而這

些錯誤均是將 HY 類別（或 HE 類別）誤判為 HE 類別（或 HY 類別）。換言之，當叢集個數為 7 時，我們的作法已可以完全區隔 AF、CHF、與健康人（HY 或 HE）。反觀文獻[2]的作法，便無法達到這樣的結果（請參圖七）。



圖六、我們作法對混合長短 RRI 序列集的叢集結果之部分畫面

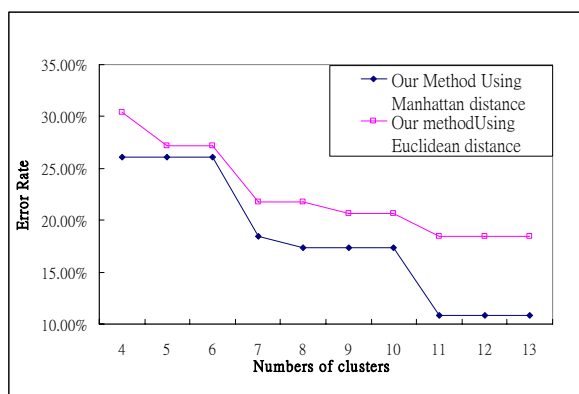


圖七、文獻[2]作法對混合長短 RRI 序列集的叢集結果之部分畫面

4. 討論

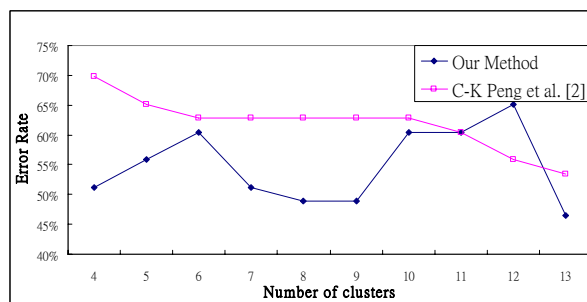
從上述實驗結果可發現，無論是對短、長、混合的 RRI 序列資料集，我們的作法都較文獻[2]的作法有較低的錯誤率。尤其是針對混合長短 RRI 序列的資料集，我們的作法可以很有效地區隔出 CHF、AF 患者與健康人。因此，我們的作法所產生的叢集品質優於文獻[2]的作法。

此外，我們也嘗試將我們作法中所使用的曼哈頓距離公式改用歐基里德距離公式。以混合長短 RRI 序列的資料集為例，從圖八中可以發現使用曼哈頓距離較歐基里德距離公式有較好的叢集品質。



圖八 我們作法使用歐基里德距離與曼哈頓距離之叢集結果錯誤率比較

對某些資料集而言，不同的預測錯誤所造成的嚴重性也有所不同。以 CHF 資料集為例，其類別欄位表示患者之 NYHA(New York Heart Association)分類等級。NYHA 等級越高，患者所允許進行的活動越受限。換言之，CHF 資料集中的類別欄位是具有順序性的。因此，將類別 1 誤判為類別 2 的錯誤應小於將類別 1 誤判為類別 3 的錯誤。針對 CHF 資料集，我們依預測值與實際值的差距做加權，將錯誤率的計算方式修改為各筆序列之預測類別與實際類別差的絕對值總和。結果如圖九所示，只有在叢集個數為 12 時，文獻[2]的做法較我們的方法有較佳之錯誤率；然而，在其他叢集個數時，我們的方法仍優於文獻[2]的做法。



圖九 加權後之叢集結果比較

5. 結論

本研究提出利用心率變異度的相關統計欄位來代表 RRI 時間序列，然後根據這些欄位來計算出任兩時間序列間的曼哈頓距離，進而根據此種距離計算結果來叢集時間序列。傳統上多認為此類統計欄位多適用於描述較長時間的量測資料，但是實驗結果顯示，無論是對長、短、或混合的 RRI 序列資料集，我們的方法所產生之叢集結果均優於 Peng 等人所提出的方法[2]。

雖然本研究只使用到 8 個統計欄位來代表 RRI 時間序列。事實上，除了統計方法所計算出的特徵值外，還有幾何方法與頻譜分析的方法[13]可以產生具有代表性的特徵值。此外，部分統計欄位值（如標準差）常會受到 RRI 時間序列長度的影響。因此，新特徵值的選擇對本方法的效能影響值得進一步的評估。不同叢集的方法也有待進一步的實驗評估。

雖然文獻[2]的方法在 RRI 時間序列的叢集表現不如我們的方法好，但是該方法的應用範圍並不侷限於 RRI 時間序列。反觀我們的方法，由於所用到的部份統計欄位（如 pNN50、pNN20 等）是在 RRI 時間序列的分析上才有意義，因此若要直接應用到 RRI 以外的時間序列，則仍需要對特徵值的選取做相當的改變。

參考文獻

- [1] Albert C.-C. Yang, Shu-Shya Hseu, Huey-Wen Yien, Ary L. Goldberger, and C.-K. Peng, "Linguistic Analysis of the Human Heartbeat Using Frequency and Rank Order Statistics," *Phys. Rev. Lett*, 90, 108103, 2003.
- [2] C.-K. Peng, Albert C.-C. Yang, Ary L. Goldberger, "Statistical physics approach to categorize biologic signals: From heart rate dynamics to DNA sequences," *CHAOS* 17, 015115, 2007.
- [3] Congestive Heart Failure RR Interval

- Database.
www.physionet.org/physiobank/database/c_hf2db
- [4] F. E. Tristani, D. G. Kamper, D. J. McDermott, B. J. Peters and J. J. Smith, "Alterations of postural and Valsalva responses in coronary heart disease," *AJP - Heart and Circulatory Physiology*, Vol 233, Issue 6 694-H699, 1977.
- [5] Fantasia Database.
www.physionet.org/physiobank/database/fantasia
- [6] Felsenstein, J. 1989, "Phylip - Phylogeny inference package," *Cladistics*. 5:164-166.
- [7] J E Mietus, C-K Peng, I Henry, R L Goldsmith, A L Goldberger, "The pNNx files: re-examining a widely used heart rate variability measure," *Heart* 88:370-380,2002
- [8] J Hayano, Y Sakakibara, M Yamada, N Ohte, T Fujinami, K Yokoyama, Y Watanabe and K Takata, "Decreased magnitude of heart rate spectral components in coronary artery disease. Its relation to angiographic severity," *Circulation*, Vol 81, 1217-1224.
- [9] K E Airaksinen, M J Ikäheimo, M K Linnaluoto, M Niemelä, and J T Takkunen, "Impaired vagal heart rate control in coronary artery disease," *Br Heart J*, December; 58(6): 592-597, 1987.
- [10] R. D. Page, "TreeView: an application to display phylogenetic trees on personal computers," *Comput Appl Biosci*. Aug;12(4):357-8,1996.
- [11] Ryan C, Hollenberg M, Harvey DB, Gwynn R,"Impaired parasympathetic responses in patients after myocardial infarction.,"*Am J Cardiol*. 37:1013-1018,1976
- [12] T Bennett, R G Wilcox, and J R Hampton, "Cardiovascular reflexes in patients after myocardial infarction. Effect of long-term treatment with beta-adrenoceptor antagonists," *Br Heart J*. September; 44(3): 265-270,1980.
- [13] Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology. "Heart rate variability-Standards of measurement, physiological interpretation, and clinical use," *Circulation*. 93:1043-1065,1996.
- [14] The Atrial Fibrillation Termination Challenge Database.
www.physionet.org/physiobank/database/afddb/
- [15] The BIDMC Congestive Heart Failure Database.
www.physionet.org/physiobank/database/c_hfdb
- [16] The Paroxysmal Atrial Fibrillation Prediction Challenge Database.
www.physionet.org/physiobank/database/afpdb
- [17] W.M. Fitch and E. Margoliash, "Construction of phylogenetic trees," *Science*, Vol.155, pp. 279-284, 1967.