A novel approach to extract the attractor feature of RR-Lorenz plot Based on SNN Density Clustering and Its Application in ECG Analysis

Yanling Liu[1,a],Xin'an Wang[1,b,*] and Ran Li[1,c]

[1]The Key Laboratory of Integrated Micro-systems Science and Engineering Applications, Peking University Shenzhen Graduate School, NanShan,Shenzhen, China

[a]1501213881@pkusz.edu.cn [b] anxinwang@pkusz.edu.cn, [c]1501213868@pkusz.edu.cn

*Corresponding author

Keywords: SNN, Lorenz-RR, feature extraction, ECG analysis.

Abstract. Computer-aided diagnosis has received intensive study in recent years, especially in an aging society which unforunately has limited medical resources but surging medical demands. It's a challenging work to determine health conditions with knowledge of dynamic successive long-term ECG.RR-Lorenz plot is an essential tool for analysing ECG, it's noise-immune, and time-domain HRVs (heart rate variability) are transformed to plots, which arrhythmia could be visually identified by expericed doctors. This paper proposes a novel approach for the problem based on DBSCAN(density-based spatial clustering of applications with noise) using SNN(shared nearest neighbor). It generates attractor features of RR-Lorenz with neither prior labels nor human interventions which would be later used to measure heart conditions. We employ this approach on datasets of PhysioNet CHF database and PhysioNet normal sinus rhythm database and comes with promising results(97.35% accuracy. And also, there is significant difference on the attractor features from NSR(normal sinus rhythm) and CHF(congestive heart failure) cases, Thus we believe the proposed approach are practical and clinically useful.

基于SNN密度聚类提取RR-Lorenz散点图吸引子特征及其在心电分析中的应用

刘彦伶[1,a]，王新安[1,b,*]，李冉[1,c]

[1]北京大学深圳研究生院集成微系统科学工程与应用重点实验室，南山，深圳，广东，中国

[a]1501213881@pku.edu.cn, [b] anxinwang@pku.edu.cn, [c]1501213868@pku.edu.cn

[*]通讯作者

关键词: SNN; Lorenz-RR; 特征提取; 心电分析

中文摘要. 随着计算机技术的发展，利用计算机技术补充日益紧缺的医疗资源成为目前热门的研究方向，从动态连续的长时心电信号中挖掘健康信息并利用这些信息快速识别健康状况的恶化，是一项具有挑战性的工作。RR-Lorenz散点图是分析长时间连续动态心电信号的重要工具，它对噪声不敏感，将复杂的心电信号转换为二维的图形信号，能通过图形特征快速识别心律失常。目前对于RR-Lorenz散点图的自动诊断尚停留在初步研究阶段。本文提出一种基于SNN的密度聚类方法，在无监督无人工干预的情况下自动准确提取RR-Lorenz的吸引子特征，通过使用PhysioNet心电数据库中的RR间期数据进行实验验证，使用本方法提取RR-Lorenz散点图的吸引子特征准确率可达到97.35%，通过对提取自正常窦性心律RR间期数据和充血性

心力衰竭患者的RR间期数据的两组吸引子特征作显著性检验，证明本方法提取的特征具有心电诊断意义。

1. 引言

1.1 RR-Lorenz散点图与心电分析

RR-Lorenz散点图（又称RR-Poincare散点图）采用非线性混沌学原理对心电图中的RR间期序列进行分析。RR-Lorenz散点图是在二维坐标系中，利用一段按时间排列的心电信号RR间期序列中的RR_i作为x轴坐标，RR_{i+1}作为y轴坐标形成的散点图（图1）。它是分析长时间连续动态心电信号的重要工具[1]。它对噪声不敏感，噪声与有效数据相比数据量较小，不会形成可观测的图形轮廓，干扰对心电信号的判断。

“吸引子”是混沌学中的概念。对于混沌系统来说，吸引子表征着系统的稳定定态，系统从任一初始状态出发，最终都会演化到相空间的某一局域上。连续RR间期序列的Lorenz散点图表现出混沌的很多特征[3]。吸引子体现在Lorenz散点图中就是众多散点聚成的簇，当足够数量的连续RR间期序列参与作图，图形特征不再因为数据量的增加而改变时，就能稳定地表现吸引子特征。同一性质的心律在Lorenz散点图中聚集成同一个吸引子。吸引子的数目与心律起源的变化有关[2]，当心律起源点发生变化时，新的吸引子将会产生，相应改变Lorenz散点图的图形特征，正常窦性心律的散点图主要呈现单分布的“彗星状”、“棒球拍状”（图1a）等。而大多数心律失常的散点图则呈现“三分布”（图1b）、“四分布”（图1c）或“多分布格子状”（图1d）图形[4]。因此，吸引子特征的提取能为心律失常的诊断提供依据[1-2]。

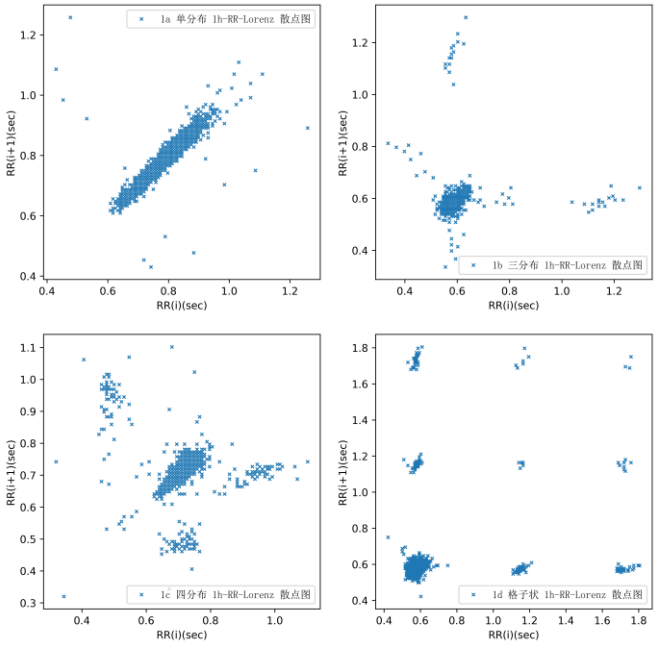

图1　使用1h连续RR间期序列所作RR-Lorenz散点图

1.2 RR-Lorenz吸引子特征定义

通过长期临床观察，李方洁[1]指出吸引子的数量，吸引子所在的位置，吸引子聚集成的几何图形形状，以及线形子图的斜率是临床使用Lorenz-RR散点图进行心律失常诊断的“诊断四要素”。本文基于“诊断四要素”的定义，将散点图聚簇形成的物理上分离的子图数指定为吸引子的数目，将吸引子所在的位置划分为等速线区、快加速区、慢加速区、快减速区、慢减速区五个区域(图2)。基于共享最近邻的密度聚类思想，设计并实现了一种能自动提取RR-Lorenz散点图吸引子的数目及所在位置的方法。

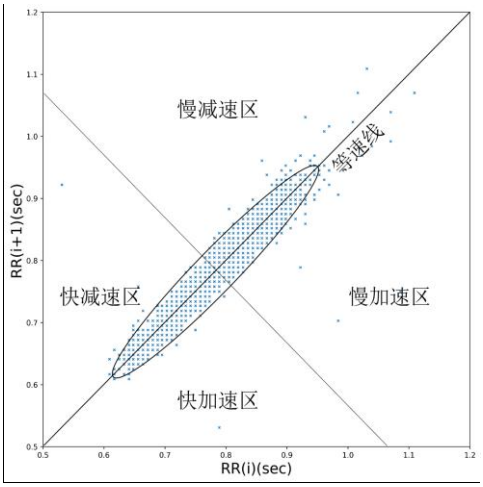

图2　本文定义的吸引子位置划分

2．特征提取方法

2.1　SNN密度聚类算法介绍

聚类算法试图在具有不同形状、大小和密度的簇中发现数据簇，或者在数据有大量噪声和离群值的情况下进行聚类[9]。实际应用中，需要根据使用场景选择合适的聚类算法，本文中，我们的目的是将所有散点划分到不同的簇中，因此考虑使用划分聚类算法。常用的划分聚类算法包括：K均值法、SOM、DBSCAN、SNN密度聚类法等。其中，K均值[7]很难处理非球形的簇和不同大小的簇；K均值和SOM算法中，簇的个数需要作为参数指定，对于无法事先确定簇个数的场景不适用。DBSCAN对于密度不同的簇之间的划分表现较差[8]。

SNN密度聚类算法擅于在存在噪声和异常值，具有不同形状、大小和密度的聚类的数据中找到聚类[10]。利用SNN密度聚类算法对二维平面中的点集*D*进行聚类，主要步骤如下：

1．假设$P \epsilon D$，P与D中其他所有点的欧式距离中，第K小的值为*K-Distance(P)*。计算每一个点的K最近邻集合（KNN，公式(1)）。对所有点求*KNN*，构成***KNNMatrix***；

$$KNN(P)=\{Q \in D | d(P,Q) \leq K\_Distance(P)\} \tag{1}$$

2．对于每一对散点，计算其相似度（Similarity，公式(2)），构成$N*N$相似度矩阵（***SimilarityMatrix***）；

$$Similarity(P,Q)=size(KNN(P) \cap KNN(Q)) \tag{2}$$

3．计算每一个点的共享最近邻密度（SNNDensity，公式(3)）；

$$SNNDensity(P) = size(Q | Similarity(P,Q) > 0) \tag{3}$$

4．设置共享最近邻密度的阈值*Minpts*，*Corepts*，标记核心点，删除噪声点；

$$P \in \begin{cases} \text{核心点 } if\ SNNDensity(P) \geq Corepts \\ \text{噪声点 } if\ SNNDensity(P) \leq Minpts \end{cases} \tag{4}$$

5．设置阈值*Eps*，将核心点聚类。聚类规则如下：

若核心点P，Q满足***SimilarityMatrix***$[P][Q] \geq Eps$，则P，Q是直接密度可达的；若P与Q直接密度可达，且Q与R直接密度可达，则P与R密度可达。若两个点满足直接密度可达或密度可达，则属于同一类别。

6．将未聚类的非核心点归入与之最近的核心点所在类别。

2.2 本文算法

RR-Lorenz散点图具备如下特征：1）吸引子的数量，聚集而成的簇的形状、大小无法预判；2）密度变化较大；3）心电信号采集设备输出的RR间期序列容易被噪声干扰，噪声点和异常值出现的概率较大。故而本文基于SNN密度聚类算法，设计了下述应用于RR-Lorenz散点图聚类的方法。

本文中用于聚类的点集是由原始RR间期形成的，将原始RR间期按时间序列，以相邻两个RR间期分别作为二维平面上的横纵坐标，即每个点用(RR_i,RR_{i+1})表示。本文算法第1、2步与2.1节中所述相同；第3步改进了*SNNDensity*的计算方法，详见2.2.1节；第4、5步的参数选择详见2.2.2节；由于核心点的类别直接决定了所有点最后聚集而成的类别，故而舍弃了2.1节中所述第6步，通过核心点聚类形成的类别判断吸引子数目及其位置。聚类完成之后，自动标注每个簇的位置，详见2.2.3节。

2.2.1共享最近邻密度计算方法

将相似度矩阵*SimilarityMatrix*假想为无向图，若*Similarity*(P，Q)> 0，则认为P与Q通过一条权值为*Similarity*(P，Q)的边相连。通常的SNN密度聚类算法中，使用与某个点有边连接的点数衡量它的SNN密度（公式(3)），忽略了每条边的权值。因此，考虑每条边的权值，本文使公式(5)中的计算方法。

利用提取自PhysioNet心电数据库的RR间期序列样本（详见第3章）分别使用公式(3)和公式(5)计算*SNNDensity*，得到*SNNDensity_Num*和*SNNDensity_Sum*序列，分析两个序列的趋势（图3）。其中，横坐标为点的序号，纵坐标为SNN密度。可以看到，二者总体趋势相同，但前者呈现阶梯状，而后者曲线更为平滑，这说明使用公式(5)计算SNN密度，每个点密度之间形成差异的分辨率越高。

$$SNNDensity(P) = sum(Similarity(P,Q)|Similarity(P,Q) > 0) \tag{5}$$

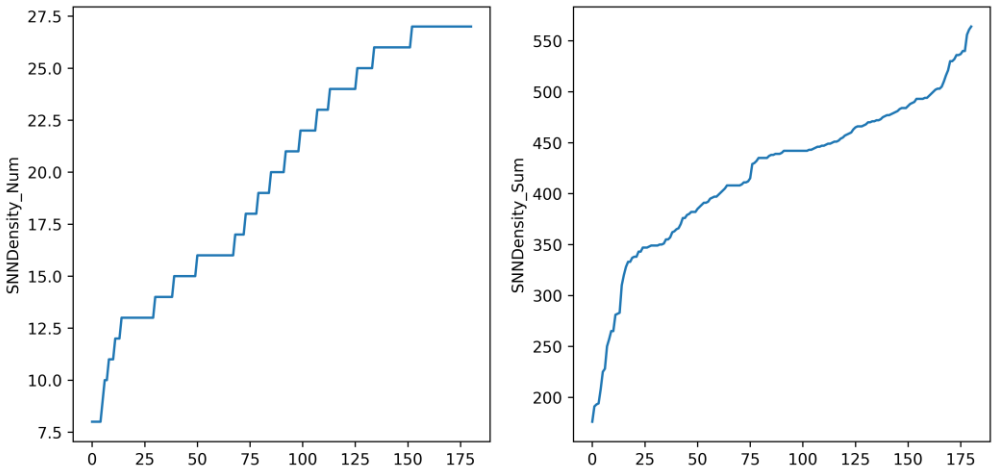

图3　两种SNNDensity计算方法所得序列趋势对比

2.2.2 *Corepts*、*Eps*选择方法

*Corepts*的选择决定了所有散点中最终参与聚类的点的数量。若选择过大，则会造成参与分类的点集过小，无法代表吸引子的分布，遗漏某些包含点集数量较少的簇（如图4中红框所标注类别）。若选择过小，则会导致参与分类的点集过大，使边界点或者噪声点参与分类，生成类别数量过多。

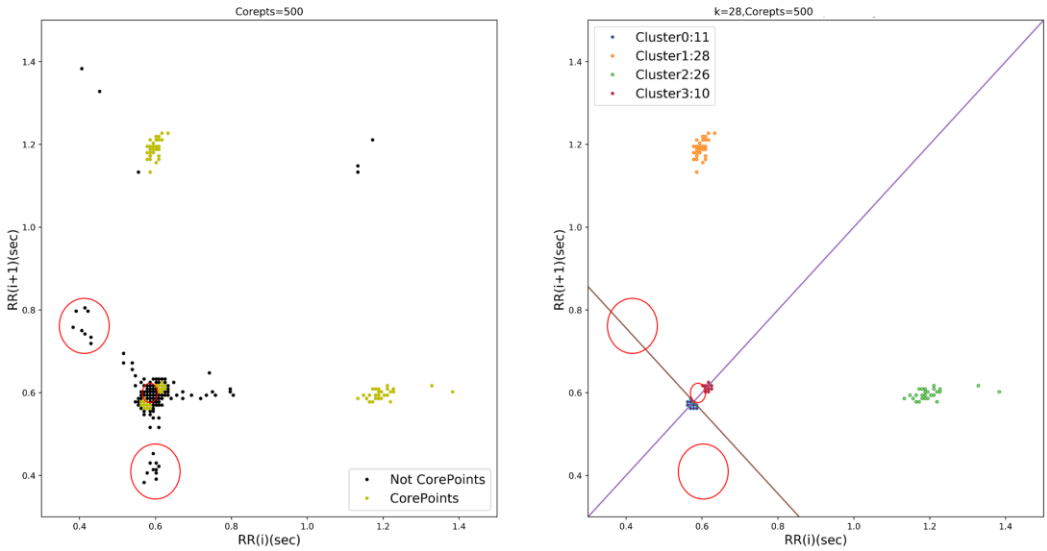

图4　Corepts选择过大导致类别遗漏

*Eps*决定核心点对是否直接密度可达。根据*SimilarityMatrix*的定义，两个核心点之间的相似度介于[*0,K*]区间。选择该参数时，可以对大量样本进行分析，观察当*Eps*在[*0,K*]之间滑动时，被判定为直接密度可达的核心点对个数情况。

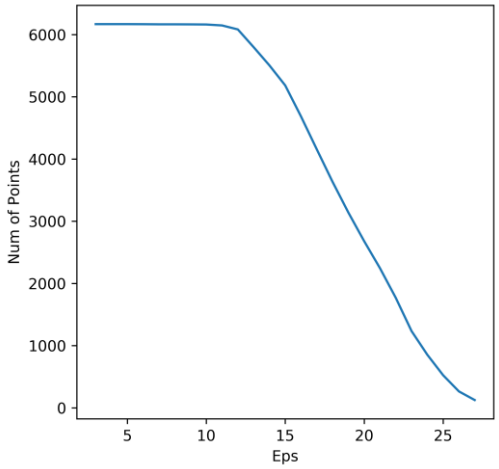

图5　直接密度可达核心点对个数随Eps变化趋势

使用RR间期序列样本作图观察直接密度可达的核心点对个数随*Eps*变化的趋势（图5）。横坐标为*Eps*，纵坐标为当前*Eps*下满足*Similarity*大于*Eps*的核心点对数。可以看到，曲线的斜率随*Eps*增大到达峰值后开始下降。这表明随着*Eps*的增大核心点群开始分裂为多个类别，分裂速度越来越快，直到达到峰值后开始变缓。可以认为，斜率的峰值为系统变化的拐点，将此拐点的横坐标作为*Eps*。

2.2.4 提取各吸引子所在位置

各类别所在位置定义详见1.2节图3。利用前文描述的算法将散点图聚成*n*个类别，则*n*为吸引子个数。提取各吸引子所在位置的方法如下：

1．选择包含点数最多的类别为主簇，求主簇的中心点α

2．等速线为经过原点，斜率为1的直线$\gamma 1 = x + b1$；快慢速区分割线为经过α，与等速线垂直的直线$\gamma 2 = x + b2$，γ1和γ2将平面划分为等速线区、快加速区、慢加速区、快减速区、慢减速区（图2）；

3．若α在γ1上，则主簇属于等速线区；否则，遍历每个点ρ，根据ρ与γ1、γ2的关系判断ρ所在位置；

4．对于每个类别，根据其大多数点所在区域判断该类别所在位置。

3．实验验证

3.1　样本来源及预处理

RR-Lorenz散点图实际上是心电信号的一种非线性分析方法，非线性方法的特点是在相互关联的大样本或超大样本海量数据中发现隐含于其中的规律。前人通过大量实验表明，时间长度为1h的RR间期序列作出的Lorenz散点图图形轮廓与24h散点图的图形一致[1]。因此，我们以每小时为单位对RR间期数据进行作图，并以每一个1hRR-Lorenz散点图作为一个样本。

从PhysioNet CHF数据库[6]和PhysioNet NSR数据库[5]选取29例充血性心力衰竭患者共270个连续1hRR间期序列作为Ⅰ组（CHF组），53例正常窦性心律个体共520个连续1hRR间期序列作为Ⅱ组（NSR组）。对每个序列作RR-Lorenz散点图，并对吸引子数目进行人工标记。

3.2　实验结果

3.2.1去重

算法的第一步是计算KNN矩阵，需要计算任意两个点之间的距离。对原始RR间期作图发现，(RR_i,RR_{i+1})构成的散点中有大量重复点。大量重复点的存在急剧增加了KNN矩阵计算的时空复杂度；同时，每一对重复点之间的相似度会随着重复的次数增多而增大，由于相似度是相对性的，这些重复点与周围环境中的非重复点的相似度则会大大降低。类别划分本应为同一个类别的点集，可能由于部分重复点的相似度高度内聚，而被识别为多个类别。

考虑到我们的目的在于提取RR-Lorenz散点图中的吸引子的数目和几何特征，重复点的个数在二维图像中不应该影响点的密度，因此在预处理阶段去除重复点。

3.2.2实验结果

*K*是本算法中的超参，设定为28；分别从Ⅰ组和Ⅱ组中选择20个样本，通过使用2.2.2节中描述的参数选择方法观察这40个样本，最终确定*Corepts*为200，*Eps*为16，得到较好的聚类效果（图6）。使用本文提出的算法提取I组剩余的250个样本以及Ⅱ组剩余的500个样本所作RR-Lorenz散点图吸引子特征，与人工标注结果对比，验证算法的有效性。如表1所示。正常窦性心律样本描画的散点图图样较为单一，多为单分布棒状或鱼雷状，而CHF样本则展现出更复杂的散点图分布，因此本算法在NSR组中的准确率相较CHF组较高，这符合预期结果。

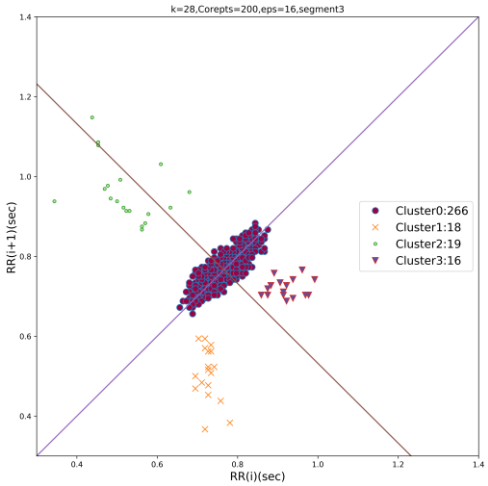

图6　算法结果图示例

4．分析与讨论

第3章验证了本算法提取特征的准确性，为了判断本算法提取的特征对于心电分析的意义，将第3章通过算法提取的来自于I组（充血性心力衰竭组）的250组吸引子特征与来自于Ⅱ组（正常窦性心律组）的500组吸引子特征进行显著性检验（表2）。除等速线上吸引子个数外，本文算法提取的其他特征在CHF组和NSR组之间均表现出显著性差异（p远小于0.001）。这表明吸引子的数目和所在位置对于正常心律和异常心律具有明显的指示意义。

表1　SNN密度聚类提取吸引子数目与人工标注对比实验结果

组别		与人工标注结果相同比例
Ⅰ CHF组（n=250）	吸引子数	93.66%
	吸引子所在位置	92.02%
ⅡNSR组（n=500）	吸引子数	99.53%
	吸引子所在位置	98.52%
所有样本（n=750）	吸引子数	97.35%
	吸引子所在位置	96.37%

表2　CHF组与NSR组RR-Lorenz散点图吸引子特征显著性检验结果

特征	指标	CHF组（I组） (n = 250) Mean±Std	NSR组（II组） (n = 500) Mean±Std	F值	P值
各区域吸引子数	快加速区	0.43±0.497	0.09±0.282	388.447	1.327E-17
	慢加速区	0.53±0.549	0.06±0.244	592.364	6.573E-24
	快减速区	0.36±0.491	0.06±0.237	451.277	2.638E-14
	慢减速区	0.38±0.537	0.06±0.237	444.651	3.085E-14
	等速线区	1.04±0.223	1.00±0.044	78.737	0.016
吸引子总数	吸引子总数	2.73±1.356	1.27±0.789	185.249	9.535E-34

5．结论与展望

本文结合医学领域对RR-Lorenz散点图的研究，提出使用基于SNN密度聚类的算法提取自动提取RR-Lorenz散点图中的吸引子特征，并提出了使用少量样本选择SNN密度聚类参数的方法，最终准确率基本与人工提取吸引子特征相同。但由于当前算法中的超参与阈值均为固定参数，在部分散点图中的效果没有达到预期，接下来考虑在算法中使用自适应参数提升算法准确率。另外，在第3章中通过显著性检验验证了本文提取的吸引子数目和所在位置对于心电分析中鉴别正常心律与异常心律具有明显的指示意义，下一步工作将使用本文的方法，利用连续心电检测设备，建立RR-Lorenz散点图数据集，推进心电分析的自动化与智能化。

6．致谢

本论文受深圳市科技计划项目：HRV预警芯片关键技术研究（JCYJ20170306091821082）、基于足底压力监测的脑卒中步态康复训练系统关键技术研究（JCYJ20170306092000960）资助。

References

[1] Li Fang-jie, The important concepts, terms and their connotations of Lorenz plot, *Journal of Practical Electrocardiology*, vol.24(3), pp.153-157, 2015.

[2] Zhong Hangmei, Li Li, Wu Ying, Application of ECG scatterplot in quick determination of complex arrhythmias. *Journal of Practical Electrocardiology*, vol.25(1), pp.9-16, 2016.

[3] Jihong Guo, Ping Zhang, Ambulatory electrocardiography. *Beijing: PMPH*, 2003

[4] Li Fang-jie, Yang xin-chun, Bai Jing et al, The contrast analysis on diagnosis of 1153 arrhythmic patients between Lorenz plot and ambulatory electrocardiogram, *J Clin Electrocardiol*, vol.15(5), pp.330-333, 2006.

[5] Stein PK, Ehsani AA, Domitrovich PP, et al, The effect of exercise training on heart rate variability in healthy older adults. *Am Heart J*, vol.138, pp.567–76, 1999.

[6] Krum H, Bigger JT Jr, Goldsmith RL, et al. Effect of long-term digoxin therapy on autonomic function in patients with chronic heart failure. *J Am Coll Cardiol*, vol.25, pp.289–94, 1995.

[7] Hartigan J A, Wong M A. Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society*, vol.28(1), pp.100-108, 1979.

[8] Ester M, Kriegel H P, Sander J, et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. 1996.

[9] HAN Jia-Wei, KAMBER M, Data Mining Concepts and Techniques 2nd Edition, *Beijing:China Machine PRESS*, pp.251-299, 2007.

[10] Levent Ertöz, Michael Steinbach, Vipin Kumar, A new shared nearest neighbor clustering algorithm and its applications// *The Workshop on Clustering High Dimensional Data & ITS Applications at Siam International Conference on Data Mining*. 2002.