Hans Journal of Biomedicine 生物医学, 2019, 9(1), 1-9
Published Online January 2019 in Hans. http://www.hanspub.org/journal/hjbm
https://doi.org/10.12677/hjbm.2019.91001

Feature Extraction and Recognition of Heart Sound Signal Based on CEEMDAN Sample Entropy

Miao Xiao[1], Jun Chang[1], Jiahua Pan[2], Hongbo Yang[2], Weilian Wang[1]

[1]School of Information, Yunnan University, Kunming Yunnan
[2]Yunnan Fuwai Cardiovascular Disease Hospital, Kunming Yunnan
Email: 1292179592@qq.com, wlwang_47@126.com

Received: Dec. 15th, 2018; accepted: Dec. 28th, 2018; published: Jan. 4th, 2019

Abstract

Due to the nonstationary characteristics of heart sound signal which was often disturbed by noise, a feature extraction method based on complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) with IMF sample entropy was proposed in this work. The heart sound signals were adaptively decomposed into several IMF components by using CEEDDAN, and the sample entropy of each order IMF component was calculated as the feature vector. A recommendation model based on Factorization Machines (FM) was proposed, which can deal with the disadvantages of sparse big data and solve the sparsity of sample entropy better. In order to verify the pros and cons of the model, AUC curve analysis was performed. 600 heart sounds of congenital heart disease and 600 normal heart sounds were analyzed. It is proved that the method can improve the signal feature extraction and show a higher recognition rate for the heart sound of congenital heart disease.

Keywords

CEEMDAN (Complete Ensemble Empirical Mode Decomposition with Adaptive Noise), Sample Entropy, Heart Sound, CHD (Congenital Heart Disease), Factorization Machines (FM)

基于CEEMDAN样本熵的心音信号特征提取及识别研究

肖　苗[1]，常　俊[1]，潘家华[2]，杨宏波[2]，王威廉[1]

[1]云南大学信息学院，云南　昆明
[2]云南省阜外心血管病医院，云南　昆明
Email: 1292179592@qq.com, wlwang_47@126.com

文章引用: 肖苗, 常俊, 潘家华, 杨宏波, 王威廉. 基于 CEEMDAN 样本熵的心音信号特征提取及识别研究[J]. 生物医学, 2019, 9(1): 1-9. DOI: 10.12677/hjbm.2019.91001

收稿日期：2018年12月15日；录用日期：2018年12月28日；发布日期：2019年1月4日

摘　要

针对心音信号的非平稳特性和易被噪声干扰的特点，本文提出一种基于自适应噪声的完备经验模态分解(CEEMDAN)与IMF样本熵结合的特征提取方法。将信号进行CEEMDAN自适应分解为若干个IMF分量，并计算各阶IMF分量的样本熵作为特征向量。在此基础上提出一种基于因子分解机(Factorization Machines, FM)的推荐模型，能更好的处理稀疏大数据的缺点，较好的解决了样本熵的稀疏性。为了验证该模型的优劣，进行了AUC曲线分析。通过对600例先心病病例心音和600例正常心音实验数据分析，证明该方法能够改善信号特征提取的效果，对先心病心音类型上的判断表现出较高的识别率。

关键词

自适应噪声的完备经验模态分解，样本熵，心音，先心病，因子分解机

1. 引言

心脏听诊是临床诊断的重要手段，尤其在先心病初诊和筛查中更为重要。对心音信号进行分类识别研究有重要的科学和临床意义。对信号进行分类识别，包括特征提取和分类识别两个过程。由于心音信号采集的不稳定和信号的非平稳特性，且微弱的心音信号经常被背景噪声所淹没，直接进行特征提取和分类识别有困难。经验模式分解(Empirical Mode Decomposition, EMD) [1]是由 N E Huang 等人在 1998 年提出的一种处理非线性、非平稳信号的时频分析方法，在心音分析领域内得到广泛应用。为了解决 EMD 的模态混叠现象，Huang [2]提出的完备经验模态分解(Ensemble Empirical Mode Decomposition, EEMD)，EEMD 是在 EMD 算法的基础上加入高斯白噪声，利用其均匀尺度的性质，改善心音信号的极值点特性和模态混叠现象，但 EEMD 的分解完备性较差，如果参数选择不当，分解会产生较多的虚假分量。Torres [3]等提出了一种自适应噪声完备经验模态分解(Complete EEMD with Adaptive Noise, CEEMDAN)，不仅提高了分解效果，也改善了分解的完备性。

样本熵[4]是在近似熵的基础上发展而来的一种新的时间序列复杂性的度量方法，与近似商相比，其抗干扰能力更强，有更好的统计稳定性。

因子分解机[5] (Factorization Machines, FM)算法是由 Steffen Rendle 等提出的一种基于矩阵分解模型的机器学习算法，它将支持向量机的优势融入到了因子分解模型中。在对特征进行建模的过程中，因子分解机模型不仅考虑了单个特征对模型的影响，而且也考虑了特征与特征之间的相互关系。因子分解机算法的出现，为特征稀疏的问题提供了一种有效的途径。

本文根据心音信号分析中存在的问题，提出了一种基于 CEEMDAN [6]分解与样本熵[7] [8]相结合的心音特征提取方法，用 CEEMDAN 方法将心音信号分解为若干个 IMF 分量，求其样本熵，并构造成特征向量，实现心音特征向量化。在分类识别算法上，本文构造了二分类因子分解机(Factorization Machines, FM)模型与 SVM 分类器对利用上述方法提取的特征进行了智能分类比较，通过对 600 例先心病病例心音

和 600 例正常心音实验数据分析，实验结果表明 CEEMDAN 与样本熵相结合的方法在心音类型分类识别中的准确性高。

2. 基本原理

2.1. CEEMDAN 算法

EMD 将信号分解为一组 IMF 和余项之和。EEMD [7]在 EMD 基础上附加频率均匀分布的高斯白噪声，使信号在不同尺度上具有连续性，以减少模态混叠的程度。CEEMDAN [9] [10]通过每个阶段附加自适应高斯白噪声来计算唯一的残余信号，设 $S(t)$ 为原始信号，令 $Z_i(t)$ 为 EMD 分解得到的第 j 个模态分量，$n_i(t)$ 为第 i 次添加的零均值，方差为 1 的白噪声序列，ε 为信噪比控制系数。则 CEEMDAN 模态分解步骤如下：

1) 向待处理信号 $S(t)$ 中分别加入多次白噪声序列 $n_i(t)$，构造出 $S_i(t)=S(t)+\varepsilon n_i(t), i=1,2,\cdots,I$，对每个 $S_i(t)$ 进行 EMD 分解，直到分解出第 1 个 IMF 分量 $Z_1(t)$，定义 CEEMDAN 的第 1 个 IMF 分量为：

$$Z_1(t)=\frac{1}{I}\sum_{i=1}^{I}d_1^i(t) \tag{1}$$

2) $j=1$ 时，计算第 1 个残余量

$$r_1(t)=S(t)-Z_1(t) \tag{2}$$

3) 进行 i 次试验，每次对 $r_1(t)+\varepsilon_1 E_1(n_i(t))$ 进行分解，直至得到第 1 个模态分量，定义第 2 个模态分量

$$Z_2(t)=\frac{1}{I}\sum_{i=1}^{I}E_1\left(r_1(t)+\varepsilon_1 E_1(n_i(t))\right) \tag{3}$$

4) 对于 $j=2,3\cdots J$，计算第 j 个残余量

$$r_j(t)=r_{j-1}(t)-d_j(t) \tag{4}$$

5) 在分别对第 j 个信号 $r_j(t)+\varepsilon_j E_j(n_i(t))$ 进行分解，直到分解出第 1 个模态分量，同时定义第 $j+1$ 个模态分量

$$Z_{j+1}(t)=\frac{1}{I}\sum_{i=1}^{I}E_1\left(r_j(t)+\varepsilon_j E_j(n_i(t))\right) \tag{5}$$

6) 令 $j=j+1$，返回步骤(4)，直到最后残余量不能再进行分解时终止分解。最后的残余量

$$R_j(t)=S(t)-\sum_{j=1}^{J}Z_j(t) \tag{6}$$

即：信号 $S(t)$ 经 CEEMDAN 分解为 J 个本征模态函数和一个剩余分量

$$S(t)=\sum_{j=1}^{J}Z_j(t)+R_j(t) \tag{7}$$

2.2. 样本熵

样本熵(SampEn) [11]是基于近似熵(ApEn) [12]的一种用于度量时间序列复杂性的改进方法，都是通过度量信号中产生新模式的概率大小来衡量时间序列复杂性，新模式产生的概率越大，序列的复杂性就越大。与近似熵相比，样本熵具有两个优势：样本熵的计算不依赖数据长度；样本熵具有更好的一致性，即参数 m 和 r 的变化对样本熵的影响程度是相同的。样本熵的定义过程如下：

1) 按照序号组成的一组维数为 m 的向量序列，$X_m(1), X_m(2), \cdots, X_m(i), \cdots, X_m(N-m+1)$，其中 $X_m(i)=\{x(i), x(i+1), \cdots, x(i+m-1)\}$，$1 \le i \le N-m+1$。这些向量代表从第 i 点开始的 m 个连续的 x 的值。

2) 定义向量 $X_m(i)$ 和 $X_m(j)$ 之间的距离 $D[X_m(i), X_m(j)]$ 为两者对应元素中最大差值的绝对值。即：

$$D[X_m(i), X_m(j)] = \max(|x(i+k)-x(j+k)|) \tag{8}$$

其中 $k=0,1,\cdots,m-1$。

对于给定的 $X_m(i)$，统计 $X_m(i)$ 与 $X_m(j)$ 之间的距离小于等于 r 的 $j(1 \le j \le N-m, j \ne i)$ 的数目，并记作 B_i。对于 $1 \le i \le N-m$，其均值定义：

$$B_i^m(r) = \frac{1}{N-m-1} B_i \tag{9}$$

4) 将所有 $B_I^m(r)$ 求和后求均值得：

$$B^{(m)}(r) = \frac{1}{N-m} \sum_{i=1}^{N-m} B_I^m(r) \tag{10}$$

5) 增加维数到 $m+1$，计算 $X_{m+1}(i)$ 与 $X_{m+1}(j)$ $(1 \le j \le N-m, j \ne i)$ 距离小于等于 r 的个数，记为 T_i。

$$T_i^m(r) = \frac{1}{N-m-1} T_i \tag{11}$$

同理：

$$T^{(m)}(r) = \frac{1}{N-m} \sum_{i=1}^{N-m} T_I^m(r) \tag{12}$$

这样，$B_I^m(r)$ 是两个序列在相似容限 r 下匹配 m 个点的概率，而 $T^{(m)}(r)$ 是两 SA 个序列匹配 $m+1$ 个点的概率，则样本熵定义为；

$$SampEn(m,r) = \lim_{N \to \infty} \left\{ -\ln \left[\frac{T^{(m)}(r)}{B^{(m)}(r)} \right] \right\} = \frac{1}{N-m} \sum_{i=1}^{N-m} T_I^m(r) \tag{13}$$

当 N 为有限值时，可以用下式估计：

$$SampEn(m,r,N) = -\ln \left[\frac{T^{(m)}(r)}{B^{(m)}(r)} \right] \tag{14}$$

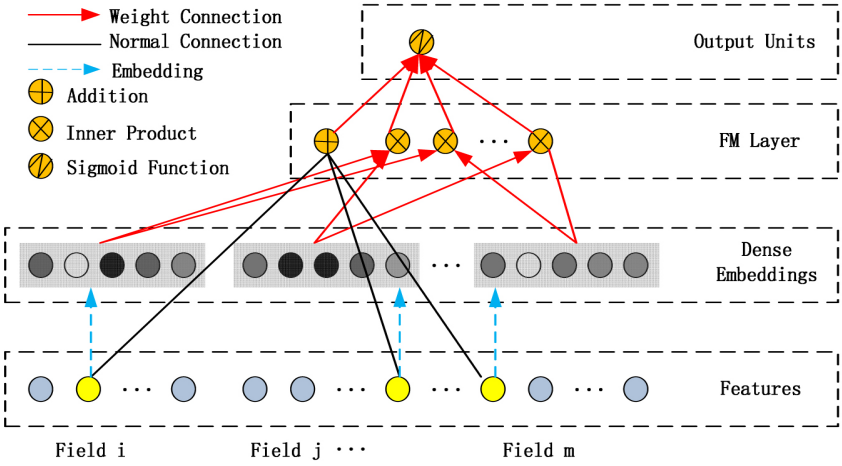
由上式子的表达式可知样本熵值的大小跟 m，r，N 的取值相关，所以确定这些参数对样本熵的计算是非常重要的。本文样本熵的参数是借鉴参考文献[7]设置的参数来选取的，$m=2$，$r=0.2$ Std (Std 表示原数据的标准差)，$N=4000$ (约一个心动周期的信号长度)。

2.3. 因子分解机 FM 模型

模型描述

因子分解机(Factorization Machine, FM)是一种通用的因式分解模型，其借鉴矩阵分解的思想，将每个参数用一个隐向量表示，用隐向量之间的点积表征组合特征的参数，可以用于解决分类、回归以及排序问题。

如图 1 所示，FM 的输出是一个加法单元和若干内积单元的总和；对于二阶因子分解机模型，其模型方程如下：



(图中 Weight Connection 红色箭头是默认连接权重为 1；Normal Connection 黑色线条连接的是权重之间的学习；Embedding，蓝色虚线箭头表示要学习的潜在载体；Add-ition 意味着将所有输入相加在一起；Inner Product 表示该单元的输出是两个输入向量的内积。Sigmoid Function 作为输出函数)

Figure 1. FM network structure diagram
图 1. FM 网络结构图

$$\hat{y} = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left\langle V_i, V_j \right\rangle x_i x_j \tag{15}$$

其中，参数 $w_0 \in R$，$W \in R^n$，$V \in R^{n \times k}$。$\left\langle V_i, V_j \right\rangle$ 表示的是两个大小为 k 的向量 V_i 和向量 V_j 的点积，表示如下：

$$\left\langle V_i, V_j \right\rangle = \sum_{f=1}^{k} v_{i,f} \, v_{j,f} \tag{16}$$

2 阶 FM 刻画的是所有单个变量和成对变量之间的关系。其中 x_i 表示数据变量，ω_0 是 globalbias 表示，ω_i 表示的是第 i 个变量对结果的影响程度，即权重系数。在 FM 中，并不是用一个单独的系数 $\omega_{i,j}$ 来表示变量间的相互作用，而是 v_i 和 v_j 内积的结果。因为对于任何正定矩阵 W，总会存在一个矩阵 V 使得 $W = VV^T$，即可表示任何相互作用的系数矩阵 W。FM 的系数学习可以通过随机梯度下降法 SGD 得到，即：

$$\frac{\partial \hat{y}}{\partial \theta} = \begin{cases} 1 & if\ \theta = w_0 \\ x_i & if\ \theta = w_i \\ x_i \sum_{j=1}^{n} v_{j,f} x_j - v_{i,f} x_i^2 & if\ \theta = w_{i,f} \end{cases} \tag{17}$$

3. 基于 CEEMDAN 样本熵与 FM 的心音识别

基于 CEEMDAN 样本熵与 FM 模型的心音识别流程图如图 2 所示。

详细步骤如下：

1) 首先利用 CEEMDAN 算法对采集到的正常心音信号以及先心病动脉导管未闭(PDA)心音信号进行分解，得到若干个从高到低的不同频率成分的 IMF 分量；由于篇幅有限，仅列出正常和病例心音各一例分解图，如图 3、图 4 所示。

2) 为了进一步对心音信号进行量化处理，本文采用具有表征复杂信号复杂程度的样本熵对各个分量进行量化处理 $T = [SampEn1, SampEn2, \cdots, SampEnn,]$。首先对正常心音和 PDA 的心音信号经过 CEEMDAN

分解按照频率从高到低得到各阶 IMF 分量；然后分别计算其分解的各阶 IMF 分量的样本熵，图 5 为这两类部分信号的各阶 IMF 样本熵，横坐标为两类信号的 1~12 阶 IMF，纵坐标表示各阶 IMF 对应的样本熵。从图中可以看出，其样本熵是随着 IMF 的增大而减小，说明 IMF 阶数越高复杂度越低；也可以看出异常信号的样本熵是低于正常信号的样本熵。

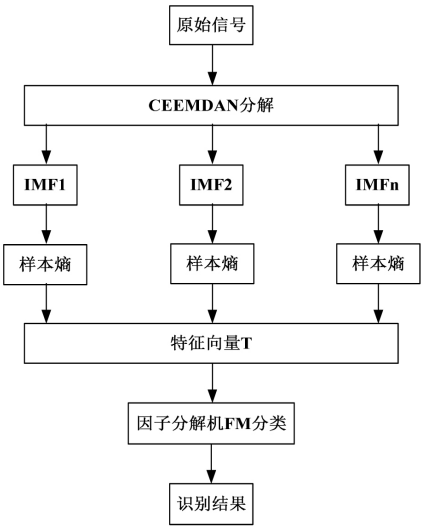


Figure 2. Heart sound recognition flow chart
图 2. 心音识别流程图

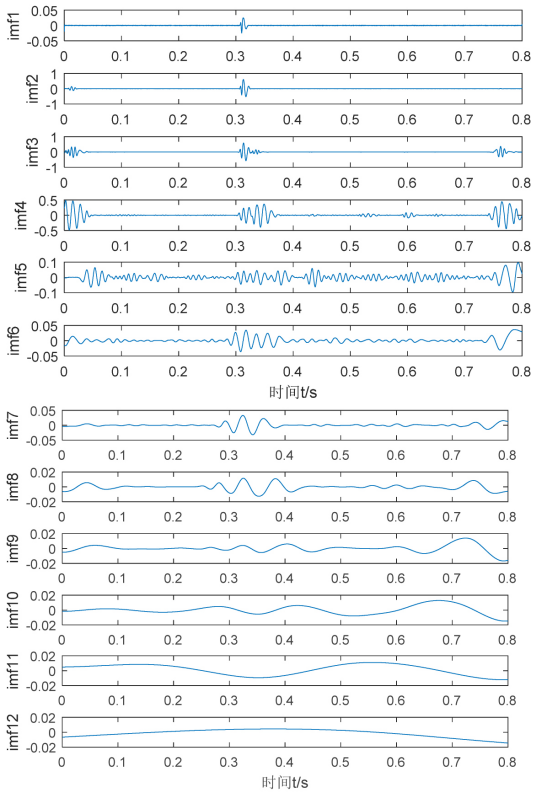


Figure 3. Normal heart sound CEEMDAN exploded view
图 3. 正常心音 CEEMDAN 分解图

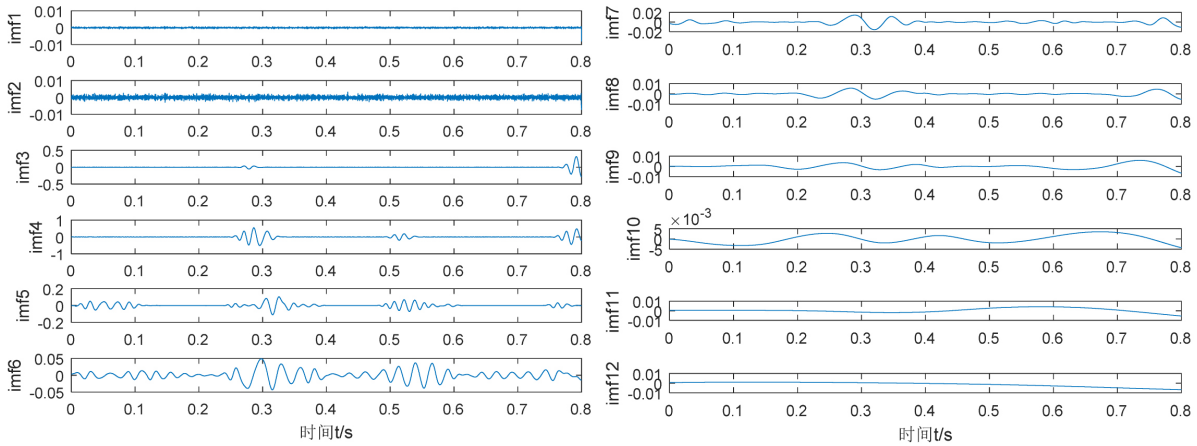


Figure 4. PDA heart sound CEEMDAN exploded view
图 4. 动脉导管未闭(PDA)心音 CEEMDAN 分解图

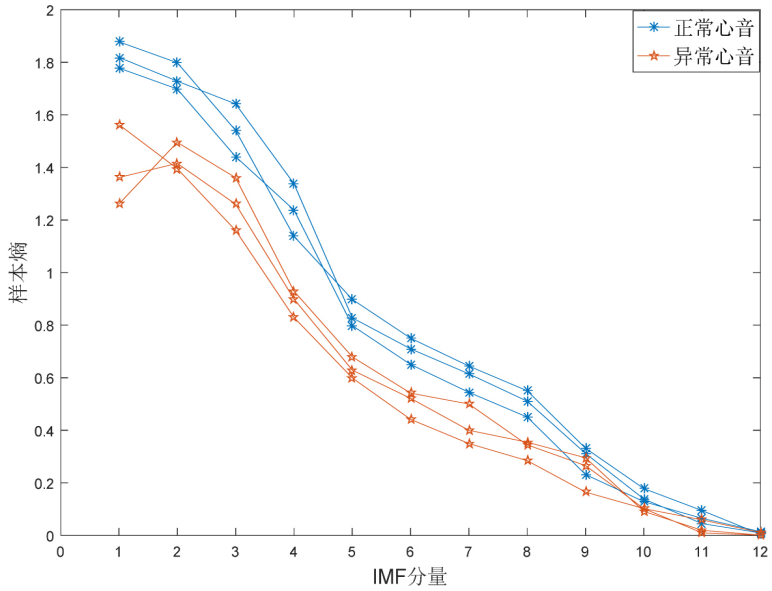


Figure 5. IMF sample entropy of each order of two types of heart sound signals
图 5. 两类心音信号的各阶 IMF 样本熵

3) 建立二分类因子分解机 FM 模型，通过网络学习和对待测样本的测试，实现心音信号的识别。

4. 实验结果与分析

研究用的心音数据源于本课题组从云南省阜外心血管病医院临床上采集的先心病心音数据库，采样频率为 5000 HZ，采集的信号以 lvm 格式保存。为了验证本文所提出方法的有效性，对采集的 600 例正常心音信号及 600 例异常心音信号分别进行 CEEMDAN 分解得到若干个 IMF 分量，再对其求熵作为特征，本文以 IMF 分量的个数作为特征维度(本文分解的个数为 12)，即网络的输入，此外，1200 例样本，其中训练样本占总样本数的 0.8，即 960 个用于训练，240 个用于测试。

采用的是机器学习中的模型性能指标 AUC (area under ROC curve)来评判其模型性能优劣。

对于二分类问题，可将样例根据真实类别和学习器的预测类别的组合划分为 TP (true positive)、FP (false positive)、TN (true negative)、FN (false negative)，真正例率(TPR, True Positive Ratio)和假正例率(FPR,

False Positive Ratio)的定义为

$$\begin{cases} TPR = \dfrac{TP}{TP + FN} \\ FPR = \dfrac{FP}{TN + FP} \end{cases} \tag{18}$$

$$ACC = \frac{TP + TN}{TP + FN + TN + FP} \tag{18}$$

受试者工作特征曲线(receiver operating characteristic curve, ROC)能够较好的来评判一个二值分类器的优劣性，ROC 的横坐标和纵坐标分别为 TPR 和 FPR。AUC 是 ROC 曲线下的面积，由于 ROC 曲线在交叉时很难通过曲线特性评判模型的好坏，所以常采用 AUC 来衡量模型的泛化性能，AUC 的曲线面积越大，其分类效果越好。采用 SVM 作为算法的参考对照，在相同的测试样本及训练样本情况下的泛化性能指标如下表 1 所示。

Table 1. Classifier generalization performance
表 1. 分类器泛化性能表

分类器	ACC	AUC
SVM	0.842	0.819
FM	0.887	0.892

从图 6 可以看出黑色曲线(FM)下的面积大于绿色曲线(SVM)下的面积，也就是说跟 SVM 算法比较，FM 在 AUC 和 ACC 上均有明显的提升，ACC 是指准确率，验证的是特征是否适合该模型，AUC 是衡量各个类别间的分类效果，能够避免类别数据不平衡带来评价误差，通过对结果的分析，FM 模型对于样本熵特征数据有较好的分类，不难解释：SVM 虽然是机器学习中最常用的分类器之一，但对于样本熵值的这种稀疏性特征，SVM 很少被使用，一是因为对于某个参数的学习，可能没有足够多的非零项，SVM 很难学习到一个可靠的分类，二是因子分解机 FM 兼具 SVM 模型的优点，能够当做一个通用的分类器来用，还能从高度稀疏的数据中学习到可靠地分类。

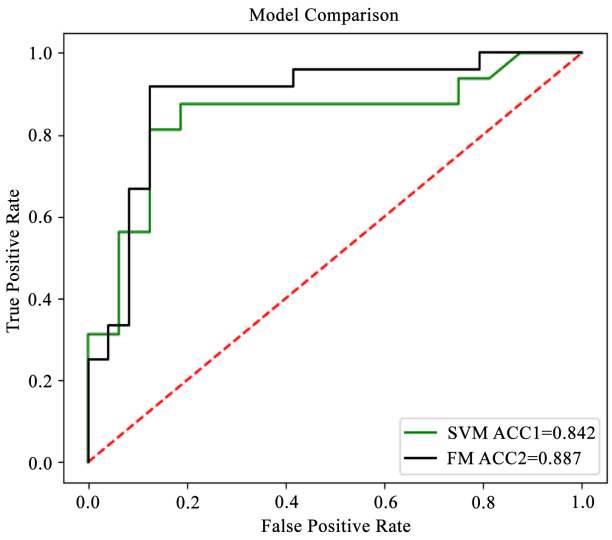


Figure 6. AUC graphs of different classifiers
图 6. 不同分类器的 AUC 曲线图

5. 结论

本文通过 CEEMDAN 分解处理非平稳复杂的心音信号，将心音信号从高频率到低频率分解成若干个 IMF 分量，结合样本熵在表征时间序列复杂程度并具有稳定的统计性特点有效地提取心音信号中的隐藏信息，为有效的提取心音特征信息，构造了二分类因子分解机模型，利用因子分解机在数据稀疏的样本情况下具有较强的二分类能力，实现了心音类型的识别，对进一步实现心脏疾病的诊断有重要的参考价值。

基金项目

国家自然科学基金项目：先心病心音库构建及特征提取算法研究，项目编号：61261008。2018 云南省重大科技专项(社会发展领域-医药健康方向)：基于人工智能与互联网+的先心病初诊辅助诊断技术研究与平台研发，项目编号：2018ZF017。

参考文献

[1] 雍希. 基于 EMD 及 SVD 的心音信号提取方法研究[D]: [硕士学位论文]. 重庆: 重庆大学, 2016.

[2] Wu, Z.H. and Huang, N.E. (2009) Ensemble Empirical Mode Decomposition: A Noise-Assisted Data Analysis Method. *Advances in Adaptive Data Analysis*, **1**, 1-41. https://doi.org/10.1142/S1793536909000047

[3] Torres, M.E., Colominas, M.A., Schlotthauer, G., *et al*. (2011) A Complete Ensemble Empiric-Al Mode Decomposition with Adaptive Noise. *IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 4144-4147.

[4] Richman, J.S. and Moorman, J.R. (2000) Physiological Time-Series Analysis Using Approximate Entropy and Sample Entropy. *AJP Heart and Circulatory Physiology*, **278**, H2039-H2049.

[5] 喻飞, 赵志勇, 魏波. 基于差分进化的因子分解机算法[J]. 计算机科学, 2016(9): 269-273.

[6] 谢志谦, 孙虎儿, 刘乐, 等. 基于 CEEMD-AN 样本熵与 SVM 的滚动轴承故障诊断[J]. 组合机床与自动化加工技术, 2017(3): 96-100.

[7] 李余兴, 李亚安, 陈晓, 等. 一种基于样本熵与 EEMD 的舰船辐射噪声特征提取方法[J]. 水下无人系统学报, 2018(1): 28-34.

[8] 丁晨莉, 马彦韬, 黄强民, 等. 利用样本熵分析针刺肌筋膜疼痛触发点的疗效[J]. 针刺研究, 2018(2): 127-132.

[9] 刘荣海, 豆龙江, 万书亭, 等. 基于 EEMD 样本熵和支持向量机的高压断路器故障诊断[J]. 华北电力大学学报(自然科学版), 2018(2): 82-88.

[10] 胡显能, 蔡改贫, 罗小燕, 等. 基于 CEEMDAN 和多尺度排列熵的球磨机负荷识别方法[J]. 噪声与振动控制, 2018(3): 146-151.

[11] 任国春, 赵永东. 基于 EEMD 样本熵与 LS-SVM 的行星齿轮箱故障诊断[J]. 山东工业技术, 2018(3): 214-215.

[12] 李振璧, 张坤, 姜媛媛, 等. 基于变分模态分解与近似熵的输电线路两相接地故障诊断[J]. 科学技术与工程, 2018(5): 70-75.

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2161-8976，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：hjbm@hanspub.org