A Generalized Shapelet Learning Method for Unsupervised Multivariate Time Series Clustering

Md Monibor Rahman[1], L. Vidyaratne[2], A. Glandon[1], A. Carpenter[2], C. Tennant[2], A. Shabalina[3], and K. Iftekharuddin[1]
[1]Vision Lab, Department of Electrical and Computer Engineering, Old Dominion University, Norfolk, VA, USA
[2]Jefferson Laboratory, Newport News, VA, USA

***Abstract*— Unsupervised multivariate time series clustering is important in many application areas. Among many unsupervised methods Shapelet learning has shown promise for univariate time series signal processing. Discovering suitable Shapelets from a large number of candidate Shapelets has been widely studied for classification of univariate time series signals. However, there is no generalized Shapelet based unsupervised clustering of multivariate time series data. Consequently, this work proposes a generalized Shapelet learning framework for unsupervised multivariate time series clustering. The proposed method utilizes spectral clustering and Shapelet similarity minimization with least square regularization to obtain the optimal multivariate Shapelets for unsupervised clustering. The proposed method is evaluated using an in-house multivariate time series dataset on detection of radio frequency (RF) faults in the Jefferson Labs Continuous Beam Accelerator Facility (CEBAF). The dataset constitutes of three-dimensional time series recordings of three RF fault types. The proposed method shows successful clustering performance with a precision of 0.732 with a standard deviation of 2.3%, recall of 0.7172 with a standard deviation of 1.7%, F-score of 0.732 with a standard deviation of 0.9%, rand index (RI) score of 0.812 with standard deviation of 0.9%, and an average normalize mutual information (NMI) of 0.56 with a standard deviation of 3.6%, in a five-fold cross validation evaluation.**

Keywords—Multivariate time series, Shapelet learning, Rand Index, Normalize mutual information.

I. INTRODUCTION

Clustering is widely used in unsupervised machine learning to partition a given set of data into non-overlapping groups. Clustering is also used to reveal the underlying structure of the data. Shapelets are subsequences of a given time series that contain salient features used to perform clustering [11]. Shapelet learning is a process of discovering those Shapelets which contains most informative features of the time series data. In [11], Shapelet based clustering is shown to consistently outperform several other methods for univariate time series clustering. However, many real-world applications require processing more complex multivariate time series data characterized by more than one dependent variable.

Most studies in literature address the problem of univariate time series classification and clustering. Generally, a univariate time series signal is considered as a point in multidimensional space. Consequently, Euclidian distance is often employed to search for similarity in multidimensional space [2]. However, due to the complexity introduced by high dimensionality, most methods tend to use a dimensionality reduction technique such as, Principal Components Analysis (PCA)[2], Discrete Fourier Transform (DFT) [19] , Discrete Wavelet Transform (DWT)[20], Singular value Decomposition (SVD)[21]. Typically, these techniques allow only a few coefficients to be chosen to represent the original signal. This tends to oversimplify the representation of the signals, and therefore reduces the performance of clustering.

One of the primary challenges of time series data analysis is to find the most informative features. Classification of time series are based on global properties of a time series which can potentially be improved with local patterns [4]. One such local pattern representations is known as Shapelets. Consequently, the discovery of suitable Shapelet is a research focus that may lead to improved classification as well as clustering of time series signals. Accordingly, Ye et al. [5] proposes a Shapelet based classification method by searching within a group of segmented patterns to determine the class of a time series signal. They rank each Shapelet based on the distance computations and entropy pruning of the information gain. The top ranked shepelets are then used for classification of the time series signal. Jesin et al. [7] propose a clustering algorithm called “u-Shapelets” to cluster univariate unlabeled data. The method selects a set of unsupervised Shapelets to separate the original dataset by searching and removing a subset (outliers) to maximize the gap between different groups divided by the unsupervised Shapelets. John et al. [8] have proposed another clustering algorithm called “k-Shapelet”, which shows good clustering results on several univariate datasets across many disciplines. Although the method [8] offers good performance, it requires a costly search over many Shapelet candidates. Instead of searching for a large number of candidate Shapelets, Grabocka et al. [6] proposes that suitable Shapelets can be discovered using a regression technique known as Shapelet learning. In this method, an intial Shapelet is extracted from the original time series and subsequently updated using a regression based learning method. This technique reduces time of the typically cumbersome Shapelet search process. Utilizing this regression based Shapelet learning, Qin et al. [9] have proposed a Shapelet learning method to cluster univariate unlabeled time series data. However, these Shapelet learning algorithms are tailored for univariate time series inputs, and hence not directly applicable for complex multivariate inputs.

Multivariate time series analysis for unsupervised clustering has not been extensively explored. Multivariate time series data are widely available in various fields such as multimedia, medicine, and finance [2]. In a human and computer interface, CyberGlove uses 22 sensors each generating 90 records per second, producing large scale multivariate time series [17,18]. Typically, in the clustering literature, multivariate time series are first transformed into univariate time series, using various dimensionality reduction techniques [1]. However, each variable of the multivariate time series may be significantly correlated to other variables. This naïve process of multivariate to univariate conversion [1] applied in typical clustering algorithms may result in a loss of some valuable information. Consequently, Shen et al [23] introduce a multivariate clustering method characterized by statistical features extracted on temporal patterns that exist between the multivariate signals. Many of the discovered patterns are less useful as features for classification and prediction. [24].

A. Shabalina[3] is with STFC Daresbury Laboratory, Warrington, UK

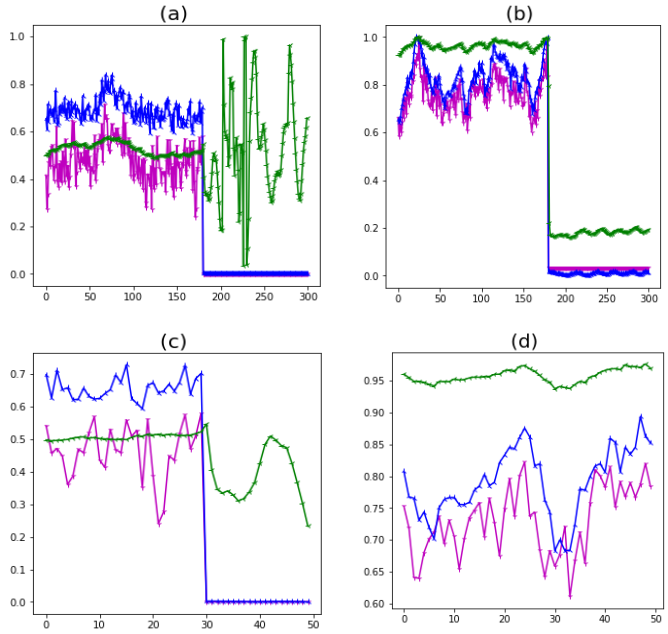

Fig. 1: (a) and (b) represent two original multivariate time-series signals of length 300. (c) and (d) presents the multivariate Shapelet of length 50 which taken from the original time series (a) and (b) respectively.

In the classification literature, there is example where the multivariate data is considered without dimensionality reduction. Bostrom et al. [3] perform multivariate time series classification using a Shapelet transform method, which inspires our work in developing a novel multivariate Shapelet learning scheme for unsupervised multivariate Shapelet based time series clustering.

This work proposes a generalization of the Shapelet learning method to multivariate time series clustering. This requires a reformulation of the cost function and the learning algorithm. In this method, we consider multivariate dependent Shapelets which are designed to maintain the phase across the channels. First, the multivariate Shapelet is converted into a Shapelet space matrix using a distance measure between the time series signal and each Shapelet. Then, the Shapelets are updated using spectral clustering, Shapelet similarity minimization and least square minimization techniques. To validate the proposed generalized Shapelet algorithm, we test our clustering on a multivariate time series dataset collected from RF cavity faults generated at Jefferson Lab CEBAF, which we describe in the motivation section. The performance is analyzed on a five-fold cross validation scheme using multiple performance indicators such as: Precision, Recall, F-score, Rand Index and Normalize Mutual Information.

Section II of this paper describes the motivation of this work. Section III describes the methodology of this work. Section IV discuss the experimental setup and results. Finally, section V concludes the work.

II. Motivation

The Continuous Electron Beam Accelerator Facility (CEBAF) in Jefferson Lab (JLab) has two linear accelerators that consist of 25 cryomodules each containing eight cavities. Operational disruptions in the whole CEBAF system can be traced back to cavity faults. Manual identification, discovery, and labeling of fault types requires substantial effort and manpower with subject matter expertise. There is a very strong coupling between cavity to cavity in cryomodules [13], which motivates an analysis of time series data from fault event in a multivariate paradigm. The proposed generalized multivariate clustering approach is expected to lead the development of a tool for automated fault identification and unsupervised data analytics for fault discovery.

III. Methodology

A typical example for a multivariate signal obtained at CEBAF runtime is shown in fig. 1. Figure 1 (a) and (b) shows two different signals, each with three dimensions. A small part of the multivariate time series signal is used as an initial multivariate Shapelet as presented in Fig. 1(c) and (d). In the proposed multivariate Shapelet learning method, each multivariate Shapelet is moved throughout the time series signals to determine the minimum distance between the Shapelet and the time series signal by maintaining phase across the channels. For each multivariate Shapelet, there is a minimum distance between the Shapelet and each time series input signal. For example, if a dataset contains 50 multivariate input signals and we consider 3 multivariate Shapelets randomly from the multivariate time series dataset, there will be total 50 X 3 minimum distances between the Shapelets and input signals. Accordingly, the three minimum distance measures for each input signal is considered as a three-dimensional data point in the Shapelet space. The time series data represented in Shapelet space are placed into groups called pseudo-classes. The boundary of the class label represents a pseudo-classifier. Shapelets are subsequently updated using spectral clustering, Shapelet similarity minimization, and least square minimization techniques, as discussed below. The updated Shapelet is used to determine the optimal pseudo class labels and pseudo-class boundary.

Consider a set of n multivariate time series (MTS) with length q and m number of variables: $TS = \{T_1, T_2, T_3, \dots T_n\} \in \mathbb{R}^{n \times q \times m}$. A single multivariate time series can be described as matrix $T = \{t_1, t_2\ t_3, \dots \dots \dots, t_q\} \in \mathbb{R}^{q \times m}$. The initial Shapelet is obtained as a small part of the time series signal. We consider a set of k multivariate Shapelets (MVS) of length l and m variables which can be written as $MVS = \{S_1, S_2, S_3, \dots S_k\} \in \mathbb{R}^{k \times l \times m}$. A single multivariate Shapelet can be described as a matrix $S = \{s_1, s_2\ s_3, \dots \dots \dots, s_l\} \in \mathbb{R}^{l \times m}$. A multidimensional Shapelet method is implemented in [3] for multivariate time series classification. In this method, multidimensional Shapelets are generated and the minimum distance for each multivariate series with respect to each Shapelet is calculated. By using only, a single distance for each multidimensional signal, we are effectively defining a transform that allows the algorithm to handle multivariate data.

We obtain the multivariate generalization of the Shapelet transformation introduced by Qin et al. [11] to reduce a long time series to a much shorter vector in Shapelet space. The Shapelet transform space, $X \in \mathbb{R}^{k \times n}$ is presented by calculating the minimum distance between the multivariate time series T and the multivariate Shapelet S.

The minimum distance between time series T_j and Shapelet S_i is presented as $X_{(S_i,T_j)}$ or $X_{(i,j)}$.

$$X_{(i,j)} = \min_{u=1,2,..,\bar{v}} \frac{1}{l_i}\sum_{r=1}^{l_i}\sum_{z=1}^{m}(T_{j,r+u-1,z} - S_{i,r,z})^2 \quad (1)$$

Where $\bar{v} = q_j - l_i + 1$ denotes the number of segments of length l_i in the time series signal T_j. The function in the equation (1) is non differential as the partial derivative $\frac{dX_{(i,j)}}{dS_{(i,r,z)}}$ is not defined. Therefore (1) approximated using soft minimum function [10] as follows:

$$X_{(i,j)} \approx \frac{\sum_{v=1}^{\bar{v}} d_{ijv} e^{\alpha d_{ijv}}}{\sum_{v=1}^{\bar{v}} e^{\alpha d_{ijv}}} \quad (2)$$

Where $d_{ijv} = \frac{1}{l_i}\sum_{r=1}^{l_i}\sum_{z=1}^{m}(T_{j,r+v-1,z} - S_{i,r,z})^2$.

The parameter α is the control precision. If $\alpha \to -\infty$, equation (2) become same as equation (1). In our case we have set $\alpha = 100$ following the method in [9].

Similarity matrix: A similarity measure is a real valued function that quantifies the similarity between two objects. The similarity between Shapelet transformed time series X is presented [11] by $A \in \mathbb{R}^{n\times n}$ as follows:

$$A_{i,j} = e^{-\frac{\|X_{(:,i)} - X_{(:,j)}\|^2}{2\sigma^2}} \quad (3)$$

Where σ is the radial basis function kernel.
Accordingly, the similarity between the Shapelet S_i and S_j is presented [11] by Shapelet similarity matrix $P \in \mathbb{R}^{k\times k}$. The similarity between S_i and S_j can be calculated as:

$$\mathrm{P_{i,j}} = \mathrm{e}^{-\frac{\|\mathrm{b_{i,j}}\|^2}{2\sigma^2}} \quad (4)$$

Where $b_{i,j}$ is the distance between Shapelet S_i and S_j obtained using equation (2).

Pseudo-class labels and pseudo class boundary: The main challenge of unsupervised learning is the unlabeled data. To cluster the unlabeled data, a pseudo-class label is introduced following [11]. If n numbers of unlabeled data belongs to c categories, then pseudo-class label matrix is $Z \in \mathbb{R}^{c\times n}$. Maximum value of each column of matrix Z represents the class label of the corresponding time series data. Pseudo-class labels make c categories of n time series signals with a class boundary $W \in \mathbb{R}^{k\times c}$.

The least square method is use in regression analysis to minimize the error. Consequently, to improve the clustering we apply the following least square minimization function [11]:

$$\min_{W} \|W^T X - Z\|_F^2 \quad (5)$$

Spectral analysis: Spectral analysis is a process by which the frequency contents of a continuous-time signal is determined in the discrete domain [22]. Spectral analysis is widely used in the clustering [12]. The similarity matrix A determines the similarity of the data set. The spectral regularization term can be formulated as [11]:

$$\begin{aligned}&\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{i,j}\left\|Z_{(:,i)} - Z_{(:,j)}\right\|_2^2 \\ &= tr(Z(D-A)Z^T) \\ &= tr(ZLZ^T) \quad (6)\end{aligned}$$

Where, $D_{i,i} = \sum_{j=1}^{n} A_{i,j}$ is the degree matrix and L is known as Laplacian matrix.

Unsupervised learning model: Based on the spectral analysis, least square minimization and Shapelet similarity minimization the unsupervised learning model presented [11] in the eqn. (7) below.

$$\min_{\mathrm{W,S,Z}} F = \min_{\mathrm{W,S,Z}} \frac{1}{2}(\mathrm{tr(ZLZ^T)} + \gamma_2\|\mathrm{W^T X - Z}\|_\mathrm{F}^2 + \gamma_1\|\mathrm{P(S)}\|_\mathrm{F}^2 + \gamma_3\|\mathrm{W}\|_\mathrm{F}^2) \quad (7)$$

Where, γ1, γ2 and γ3 are regularization parameters. The eqn. (7) is the objective function, F has three variables W, Z and S. We update each variable by keeping other two constant. Updates for pseudo class boundary or pseudo classifier W, and pseudo class label Z can be found by solving eqn. (7). Pseudo classifier W is found by derivative of eqn. (7) while Z and S is fixed. We obtain the updated value for W by setting $\frac{\partial \mathrm{F(W)}}{\partial \mathrm{W}}$ =0. We apply similar procedure to obtain updates for Z.

$$\mathrm{W} = \gamma_2(\gamma_2\mathrm{XX^T} + \gamma_3\mathrm{I})^{-1}\mathrm{XZ^T} \quad (8)$$

$$\mathrm{Z} = \gamma_2\mathrm{W^T X(L} + \gamma_2\mathrm{I})^{-1} \quad (9)$$

Updated value of S can be found by fixing Z and W from eqn. (7), as shown below:

$$\min_S F(S) = \frac{1}{2}(tr(ZL(S)Z^T) + \gamma_1\|P(S)\|_F^2 + \gamma_2\|W^T X(S) - Z\|_F^2) \quad (10)$$

The function in equation (10) is non-convex with respect to Shapelet S. To optimize the objective function, we follow an iterative algorithm. The updated Shapelet after following the iterative algorithm will be $S_{i+1} = S_i - \eta\nabla S_i$. Where, $\nabla S_i = \frac{\partial F(S_i)}{\partial S}$ and η is the learning rate.

The derivative of the function in equation (12) will be:

$$\frac{\partial \mathrm{F(S)}}{\partial \mathrm{S_{ks,ls,ms}}} = F_1 + F_2 + F_3 \quad (11)$$

With $\mathrm{F_1}$, $\mathrm{F_2}$, and $\mathrm{F_3}$ are defined as follows:

$$\mathrm{F_1} = \frac{1}{2}\left(\mathrm{Z}\frac{\partial \mathrm{L(S)}}{\partial \mathrm{S_{ks,ls,ms}}}\mathrm{Z^T}\right)$$

$$\mathrm{F_2} = \gamma_1\mathrm{P(S)}\frac{\partial \mathrm{P(S)^T}}{\partial \mathrm{S_{ks,ls,ms}}}$$

$$\mathrm{F_3} = \gamma_2\mathrm{W(W^T X - Z)}\frac{\partial \mathrm{X(S)^T}}{\partial \mathrm{S_{ks,ls,ms}}}$$

where $ks = 1,2,\ldots.k, ls = 1,2\ldots..l_i, and\ \ ms = 1,2\ldots\ldots,m.$

We know from equation (6) that, L=D-A, $D_{i,i} = \sum_{j=1}^{n} A_{i,j}$. Therefore, $\frac{\partial L(S)}{\partial S_{ks,ls,ms}}$ can be obtained by calculating $\frac{\partial A(S)}{\partial S_{ks,ls,ms}}$ as follows;

$$\frac{\partial \mathrm{A_{n1,n2}}}{\partial \mathrm{S_{ks,ls,ms}}} = \frac{\partial}{\partial \mathrm{S_{ks,ls,ms}}}\left(e^{-\frac{\left\|X_{(:,i)}-X_{(:,j)}\right\|^2}{2\sigma^2}}\right)$$

$$= \frac{-\mathrm{A_{n1,n2}}}{2\sigma^2}(\mathrm{X_{n1,ks}} - \mathrm{X_{n2,ks}})\left(\frac{\partial \mathrm{X_{n1,ks}}}{\partial \mathrm{S_{ks,ls,ms}}} - \frac{\partial \mathrm{X_{n2,ks}}}{\partial \mathrm{S_{ks,ls,ms}}}\right)$$

$$\frac{\partial \mathrm{X_{n1,ks}}}{\partial S_{ks,ls,ms}} = \frac{G_1 \frac{\partial G_2}{\partial S_{ks,ls,ms}} - G_2 \frac{\partial G_1}{\partial S_{ks,ls,ms}}}{{G_1}^2}$$

Where, $G_1 = \sum_{v=1}^{\bar{v}} e^{\alpha d_{\mathrm{n1,ks,v}}}$

$$G_2 = \sum_{v=1}^{\bar{v}} d_{\mathrm{n1,ks,v}} e^{\alpha d_{\mathrm{n1,ks,v}}}$$

$$\frac{\partial G_1}{\partial S_{ks,ls,ms}} = \sum_{v=1}^{\bar{v}} (1 + \alpha d_{\mathrm{n1,ks,v}}) e^{\alpha d_{\mathrm{n1,ks,v}}} \frac{\partial d_{\mathrm{n1,ks,v}}}{\partial S_{ks,ls,ms}}$$

$$\frac{\partial G_2}{\partial S_{ks,ls,ms}} = \sum_{v=1}^{\bar{v}} \alpha e^{\alpha d_{\mathrm{n1,ks,v}}} \frac{\partial d_{\mathrm{n1,ks,v}}}{\partial S_{ks,ls,ms}}$$

$$\frac{\partial d_{\mathrm{n1,v,ks}}}{\partial S_{ks,ls,ms}} = \frac{\partial}{\partial S_{ks,ls,ms}}\left(\frac{1}{l_i}\sum_{r=1}^{l_i}\sum_{z=1}^{m}\left(T_{ks,r+v-1,z} - S_{n1,r,z}\right)^2\right)$$

$$= -\frac{2}{l_i}\left(T_{ks,ls+v-1,ms} - S_{n1,ls,ms}\right)$$

$$\frac{\partial P_{\mathrm{i,j}}}{\partial S_{ks,ls,ms}} = -\frac{1}{\sigma^2} P_{\mathrm{i,j}} b_{\mathrm{i,j}} \frac{\partial b_{\mathrm{i,j}}}{\partial S_{ks,ls,ms}}$$

$b_{\mathrm{i,j}}$ is the distance between Shapelet i and Shapelet j. If both Shapelet is equal in length, we can write the derivative $\frac{\partial b_{\mathrm{i,j}}}{\partial S_{ks,ls,ms}}$ as follows:

$$\frac{\partial b_{\mathrm{i,j}}}{\partial S_{ks,ls,ms}} = \frac{\partial}{\partial S_{ks,ls,ms}}\left(\frac{1}{l_i}\sum_{r=1}^{l_i}\sum_{z=1}^{m}\left(S_{j,r,z} - S_{i,r,z}\right)^2\right)$$

$$= \frac{2}{l_i}\sum_{r=1}^{l_i}\sum_{z=1}^{m}(S_{j,r,z} - S_{i,r,z})\left(\frac{\partial S_{j,r,z}}{\partial S_{ks,ls,ms}} - \frac{\partial S_{i,r,z}}{\partial S_{ks,ls,ms}}\right)$$

$$= \begin{cases} \frac{2}{l_i}\left(S_{ks,ls,ms} - S_{i,ls,ms}\right) & for\ j = ks, i \neq ks \\ \frac{2}{l_i}\left(S_{ks,ls,ms} - S_{j,ls,ms}\right) & for\ i = ks, j \neq ks \\ 0 & Otherwise \end{cases}$$

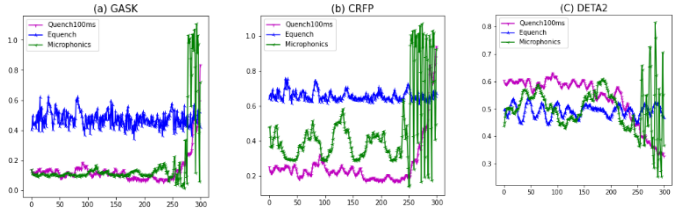

Fig. 2: Examples of dataset signals for different channels and different faults

If lengths of Shapelet i and Shapelet j are not equal then $\frac{\partial b_{\mathrm{i,j}}}{\partial S_{ks,ls,ms}}$ can be calculated using the derivative of eqn. (2).

Grid search: We identify five different hyperparameters that must be obtained through a grid search. First parameter is the length of Shapelet, li. We have varied the length of the Shapelet, {0.1, 0.15,…..,0.25}×T_i, where T_i is the length of the original time series signal. The second parameter is the number of Shapelets k. We vary the number of k from 1 to 5. We have searched other three weight parameters (γ_1, γ_2, γ_3) in the range of $\{10^{-6}, 10^{-4}, \ldots\ldots, 10^{4}, 10^{6}\}$. We have used maximum internal iteration of 30 and learning rate η=0.001. For each combination of training dataset, we do grid search to find optimal pseudo class label and pseudo classifier.

The proposed multivariate Shapelet learning scheme with above steps is illustrated in the Algorithm below. The goal of this work is to learn the best Shapelets that optimizes the clustering and class boundary for multivariate time series. Initial value of Shapelet S_0 is an arbitrary segment of the input multivariate time series signal. Initial clustering for the Shapelet transform values can found by applying k-means clustering algorithm. The initial value of the pseudo classifier can be found by the center of the K-means clustering.

Algorithm: Multivariate time series clustering

Inputs:
Multivariate time series signal
$\gamma_1, \gamma_2, \gamma_2$: Regularization parameters
l_i, k: length and number of Shapelets
η, i_{max}: learning rate and maximum iteration
O: Optimization threshold value
Outputs: Z, W, S
Initialization: S_0, W_0
Repeat:
Calculate: X_t and L_t using eqn. (2) and eqn. (6)
Determine updated value of Z and W:
$Z_{t+1} = \gamma_2 \mathrm{W_t^T X_t (L_t + \gamma_2 I)^{-1} X Z^T}$
$\mathrm{W_{t+1}} = \gamma_2(\gamma_2 \mathrm{X_t X_t^T} + \gamma_3 \mathrm{I})^{-1}$
For i=1,2,3….,i_{max}
Calculate: X_i , L_i, and P_i using eqn. (2), eqn. (6), and eqn. (4)
Calculate: $\nabla S_i = \frac{\partial F(S_i)}{\partial S}$ using eqn. (11)
Update: $S_{i+1} = S_i - \eta \nabla S_i$
i=i+1
end for
$S_{t+1} = S_{imax}$
t=t+1
Calculate: F using eqn. (7)
Until: F<O
$Return$ $Z = Z_{t+1}, S = S_{t+1}, W = W_{t+1}$

IV. Experiments

A. Data acquisition and datasets description

The data acquisition system in C100 cryomodule synchronously acquires timestamps and saves waveform records of 17 different rf signals from each of the eight cavities in the cryomodule. The data acquisition system included two primary components, the LLRF and experimental physics and industrial control systems (EPICS), along with a collection of high-level applications [14]. These two components work

together to generate and save data for further analysis. Each of the recorded 17 signals are 8192 points long. The recorded data are arranged approximately 94% before the faults and 6% after the occurrence of the faults. The duration of the recorded signals is approximately 1535 ms and sample rate is 0.2ms.

In the data preprocessing stage, we normalize the data which removes the mean and scales each feature/variable to unit variance. This operation is performed feature-wise in an independent way. RF cavity faults dataset contain long time sequences with 8192 steps. One of the major issues of time series data analysis is that large signal length drastically increases the processing time. To improve the processing time, we preprocess the signals in two steps as follows: 1) we crop the signal from around 600 ms before fault to 50ms after fault to retain the maximum valuable information of the faults. 2) We down sample the signal after applying an antialiasing filter to further reduce the signal length. For this experimental analysis we consider three different faults named as Equench, Quench100ms and Microphonics. Signals from three different channel/dimension named as GASK, DETA2 and CRFP are taken from total of 17 channels based on expert opinion. The full dataset obtained for this analysis includes 358 multivariate time series examples. There are 161 examples for Microphonics fault, 100 for Equench and 97 for quench100ms fault. Figure 2 presents the examples of different faults and signal from different channels after preprocessing. Fig. 2(a) presents the Equench, Quench100ms and Microphonics faults signal collected from channel named as GASK. Similarly, Fig. 2 (b) and Fig. 2(c) presents those faults collected from channel CRFP and DETA2.

B. *Performance measures*

We utilize several clustering performance measures such as: normalized mutual information (NMI)[15], Rand index (RI) [16], along with Precision, recall and F-score to measure the performance of the pseudo-classifier in the proposed algorithm. The performance measures are defined as follows:

Rand index: Rand index is computed by using the following formula:

$$RI = \frac{TP+TN}{TP+FP+TN+FN}$$

Where, TP represents the true positive, TN denotes the true negative, FP indicates false positive and FN represents the false positive.

NMI: NMI can be computed by using following formula:

$$NMI(Y,C) = \frac{2\times I(Y;C)}{[H(Y)+H(C)]}$$

TABLE I. PERFORMANCE OF MULTIVARIATE CLUSTERING

Dataset	Performance parameter				
	Precision	***Recall***	***F-score***	***RI***	***NMI***
Fold 1	0.768	0.713	0.739	0.822	0.606
Fold 2	0.736	0.724	0.724	0.814	0.593
Fold 2	0.706	0.692	0.699	0.796	0.573
Fold 4	0.716	0.739	0.728	0.81	0.549
Fold 5	0.734	0.718	0.726	0.814	0.515

Where Y= class labels, C= cluster labels, H(Y)= entropy of the class labels, H(C)= entropy of the cluster labels and I(Y; C)= mutual information between Y and C.

Precision: Precision determines the percentage of properly classified examples within the same cluster.

$$Precision = \frac{TP}{TP+FP}$$

Recall: Recall determines the percentage of elements that are properly included in the same cluster.

$$Recall = \frac{TP}{TP+TN}$$

F-measure: The F-measure combines precision and recall. F-measure determine how the clustering method is precise.

$$F\ score = \frac{2*Precision*Recall}{Precision+Recall}$$

C. *Results and discussion*

We apply our generalized Shapelet learning scheme to perform unsupervised clustering of RF cavity faults of particle accelerator data in Jefferson Lab. All folds are properly stratified to ensure representation from all three classes. We perform 5-fold cross validation to analyze the performance of the proposed algorithm. All folds are properly stratified to ensure representation from all three classes. The detailed five-fold performance figures are presented in Table I. Average precision value of the five-fold is 0.732 with a standard deviation of 2.3%. Which indicate that the multivariate Shapelet learning model can cluster on average 73.2% of relevant instances among the retrieved instances from the test dataset. Average recall value of the fivefold is 0.7172 with a standard deviation of 1.7%. The recall value indicates that 71.72% of relevant instances that are retrieved from the test examples. Average F-score value of the experiment is 0.7232 with a standard deviation of 0.9%. The model rand index is more in fold 1 which is 0.822. Mean RI is 0.8112 and standard deviation of 0.95%. Lower standard deviation indicate that model performance is stable if the training and test dataset changes. The maximum NMI were achieved in fold 1 and minimum in fold 5. The average NMI is 0.56 and standard deviation 3.6% for the cross validation.

The clustering output is shown in Fig.3. This output is generated from one of the hyperparameter combination. It is the Shapelet space transform representation using 3 Shapelets. We apply PCA to reduce the Shapelet dimension for 2D visualization purposes.

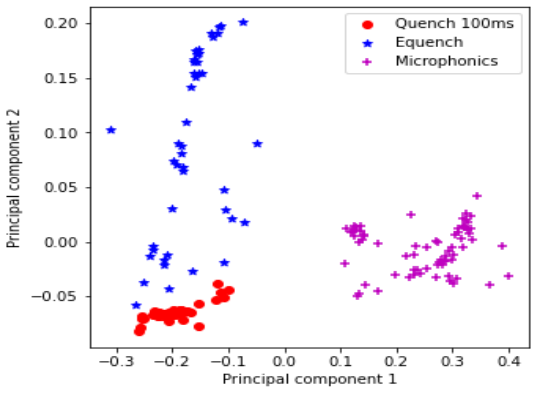

Fig. 3: Clustering output of the multivariate Shapelet learning method

V. Conclusion and Future Works

This work proposes a generalized Shapelet learning scheme for multivariate unsupervised time series clustering. The proposed method automatically learns multivariate Shapelet functions based on spectral clustering, least square minimization, and a pseudo-classification process. The performance of the proposed model is extensively evaluated using a 5-fold cross-validation scheme applied to a challenging multivariate time-series dataset obtained from the Jefferson Labs CEBAF hardware fault detection study. The results suggest that the model successfully clusters multivariate input, identifying the different RF fault types represented in the dataset.

In future we plan to benchmark the proposed method by extensive performance comparison with other works in literature. Additionally, we plan further improvements to the proposed method to efficiently perform multivariate Shapelet learning with large-scale multi-class input towards full automation of the fault discovery process in Jefferson Labs CEBAF hardware system.

Acknowledgment

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Nuclear Physics under Contract No. DE-AC05-06OR23177. This work used the available resources in the high-performance computer (HPC) facility in the Wahab and Turing cluster of Old Dominion University, Norfolk, VA.

References

[1] M. Shokoohi-Yekta, B. Hu, H. Jin, J. Wang, and E. Keogh, “Generalizing dtw to the multidimensional case requires an adaptive approach," Data Mining and Knowledge Discovery, vol. 31,no. 1, pp. 131, 2017.

[2] Kiyoung Yang and Cyrus Shahabi , “A PCA-based Similarity Measure for Multivariate Time Series”, MMDB '04: Proceedings of the 2nd ACM international workshop on Multimedia databases, November 2004 Pages 65–74

[3] Aaron Bostrom and Anthony Bagnall, “A Shapelet Transform for Multivariate Time Series Classification”, Computer Science, Mathematics,ArXiv,2017

[4] Mustafa Gokce Baydogan, George Runger, and Eugene Tuv, “A Bag-of-Features Framework to Classify Time Series”, IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 11, november 2013

[5] Lexiang Ye, Eamonn Keogh, “Time Series Shapelets: A New Primitive for Data Mining”, KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining

[6] Josif Grabocka, Nicolas Schilling, Martin Wistuba, Lars Schmidt-Thieme, “Learning Time-Series Shapelets”, KDD '14: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data miningAugust 2014 Pages 392–401

[7] Jesin Zakaria · Abdullah Mueen · Eamonn Keogh, “Accelerating the discovery of unsupervised-shapelets”, Data Min Knowl Disc, DOI 10.1007/s10618-015-0411-4

[8] John Paparrizos, Luis Gravano, “k-Shape: Efficient and Accurate Clustering of Time Series”, SIGMOD ’15, May 31–June 4, 2015, Melbourne, Victoria, Australia.

[9] J. Lines, L. M. Davis, J. Hills, and A. Bagnall, “A shapelet transform for time series classification,” in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2012, pp. 289–297.

[10] J. Grabocka, N. Schilling, M. Wistuba, and L. Schmidt-Thieme, “Learning time-series shapelets,” in Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2014, pp. 392–401.

[11] Qin Zhang , Jia Wu , Peng Zhang, Guodong Long , and Chengqi Zhang, “Salient Subsequence Learning for Time Series Clustering”, IEEE transactions on pattern analysis and machine intelligence, vol. 41, no. 9, september 2019

[12] William E Donath and Alan J Hoffman. “Lower bounds for the partitioning of graphs”. IBM Journal of Research and Development, 17(5):420– 425, 1973.

[13] C. Tennant , A. Carpenter, T. Powers, A. Shabalina , L. Vidyaratne and K. Iftekharuddin, “Superconducting radio-frequency cavity fault classification using machine learning at Jefferson Laboratory”. Physical review accelerators and beams 23, 114601 (2020)

[14] A. Solopova, A. Carpenter, T. Powers, Y. Roblin, C. Tennant, L. Vidyaratne, K. Iftekharuddin, “SRF cavity fault classification using machine learning”, 10th Int. Particle Accelerator Conf. IPAC2019, Melbourne, Australia

[15] W. M. Rand, “Objective criteria for the evaluation of clustering methods,” J. Amer. Statistical Assoc., vol. 66, no. 336, pp. 846–850, 1971

[16] Julio-Omar Palacio-Nino, Fernando Berzal,”Evaluation Metrics for Unsupervised Learning Algorithms”, https://arxiv.org/pdf/1905.05667.pdf

[17] M. W. Kadous. Temporal Classification: Extending the Classification Paradigm to Multivariate Time Series. PhD thesis, University of New South Wales, 2002.

[18] C. Shahabi. AIMS: An immersidata management system. In VLDB CIDR, 2003.

[19] R. Agrawal, C. Faloutsos, and A. N. Swami. Efficient Similarity SearchIn Sequence Databases. In FODO, 1993.

[20] K. pong Chan and A. W.-C. Fu. Efficient time series matching by wavelets. In ICDE, 1999.

[21] F. Korn, H. V. Jagadish, and C. Faloutsos. Efficiently supporting ad hoc queries in large datasets of time sequences. In SIGMOD, 1997.

[22] K. Deergha Rao, N.S. Swamy “Spectral Analysis of Signals”, Digital Signal Processing pp 721-751

[23] Pei-Yuan Zhou,and Keith C.C. Chan “A Model-Based Multivariate Time Series Clustering Algorithm”, PAKDD 2014 Workshops, LNAI 8643, pp. 805–817, 2014

[24] Alexander ShknevskyYuval ShaharRobert Moskovitch , “Consistent discovery of frequent interval-based temporal patterns in chronicpatients’data”, Journal of Biomedical Informatics 75 (2017) 83–9584