IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 19, 2018, accepted May 13, 2018, date of publication May 17, 2018, date of current version June 26, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2837660

Using Empirical Recurrence Rates Ratio for Time Series Data Similarity

MOINAK BHADURI[1] AND JUSTIN ZHAN[2]
[1]Department of Mathematical Sciences, University of Nevada at Las Vegas, Las Vegas, NV 89154, USA
[2]Department of Computer Science, University of Nevada at Las Vegas, Las Vegas, NV 89154, USA

Corresponding author: Justin Zhan (justin.zhan@unlv.edu)

This work was supported in part by the United States Department of Defense under Grant W91 1 NF-13-1-0130 and Grant 72140-NS-RIP, in part by the National Science Foundation under Grant 1625677, Grant 1560625, and Grant 1710716, and in part by the United Healthcare Foundation under Grant 1592.

ABSTRACT Several methods exist in classification literature to quantify the similarity between two time series data sets. Applications of these methods range from the traditional Euclidean-type metric to the more advanced Dynamic Time Warping metric. Most of these adequately address structural similarity but fail in meeting goals outside it. For example, a tool that could be excellent to identify the seasonal similarity between two time series vectors might prove inadequate in the presence of outliers. In this paper, we have proposed a unifying measure for binary classification that performed well while embracing several aspects of dissimilarity. This statistic is gaining prominence in various fields, such as geology and finance, and is crucial in time series database formation and clustering studies.

INDEX TERMS Time series, classification, database clustering, similarity measures, empirical recurrence rates, empirical recurrence rates ratios, bootstrapping.

I. INTRODUCTION

In the modern age, the necessity to construct efficient tools to classify and categorize time series instances is undeniable. Their applications are numerous. One may check whether global temperature variations during the present decade are similar to the ones in the last decade, or whether a volcano's eruption pattern influences that of a neighboring one. These longitudinal data use temporal dynamism and the data collection machinery, which has contributed to treating each time series as an instance. The fundamental decision is which of these instances (i.e. time series vectors) are ''similar'' to each other, and as a result, can be clustered together. As Liao [1] narrates, several application domains have witnessed such clustering exercises over the years.

As Fulcher and Jones [2] and Wang *et al.* [3] point out that the principal obstacles one needs to overcome are choosing an adequate representation of the instances and deciding on a proper measure of discrepancy or separation between the time series. Despite an extensive literature on both obstacles [1], [3], [6], unanimity on the existence of a ''best'' or ''ideal'' distance measure in a classification framework is elusive [4]. For instance, Euclidean type distances, aimed at unearthing the level of closeness between two time series, suffer in the almost constant presence of noise and misalignments in the series [3]. A large set of distance measures have been proposed [5] to circumvent problems like these. One of the goals of this paper is to offer a useful addition to the list.

The most prevalent choice is the time domain form, where the distance between two time series relies on the aggregated distance between the specific measurements. A new, unclassified time series can often be categorized by finding its similarity to another, with a known classification label. This is typically termed the ''instance-based'' approach [3], [6]. The observation that a time series of any length can be condensed into a short, summary vector of essential features (such as the mean, variance, skewness etc), enables another way of classification, [7] a ''feature-based'' type. Fulcher and Jones [2] proposed an automated method for generating such feature-based representations and noted that each classification tool could be perfectly categorized as either instance-based only or feature-based only. For instance, shapelets-based classifications [16], [17] exploit minimum distance of particular time series subsequences. Another purpose of the present work is to propose such a hybrid approach embracing good properties from both instance and feature-based categorizations. Some work [34] has been done along those lines.

The paper's layout is as follows: the next section introduces the kernel fraction. Following that, the section offers along with observations on its appealing structure. Section 3 describes ways to exploit this fraction on simulated data sets through the introduction of several metric-like measures. Section 4 describes the use of bootstrapping and its relevance in the current instance. Section 5 analyzes three real data sets to unify the proposed ideas. The final section concludes with thoughts on future directions.

II. EMPIRICAL RECURRENCE RATES RATIO

In statistical literature, random variables that model unbounded counts (such as the number of lightning strikes per month, the number of accidents per day) are often assumed to follow a Poisson distribution. The probability of having x such events is given by:

$$f_X(x) = P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad \text{for } x = 0, 1, 2, \ldots \tag{1}$$

where λ is the rate of their occurrence. Notationally, we write $X \sim Pois(\lambda)$. This variable requires the following definitions, although later on in our work, we will see how the proposal still remains relevant in more general settings.

A. DEFINITIONS AND NOTATIONS

Let $X_1 \sim Pois(\lambda_1)$, $X_2 \sim Pois(\lambda_2)$ and X_1 be independent of X_2. The conditional test (C test) developed by Przyborowski and Wilenski [9] is often used to test the difference between the two Poisson means. It is based on the fact that the sum $S = X_1 + X_2$, follows a Poisson distribution with rate parameter, $\lambda_1 + \lambda_2$, and the conditional distribution of X_1 given $S = s$ is distributed as $Binomial(s, p_{12})$, where $p_{12} = \lambda_1/(\lambda_1 + \lambda_2) = w_{12}/(1 + w_{12})$ with $w_{12} = \lambda_1/\lambda_2$. Thus, for the C-test, testing $H_0 : \lambda_1 = \lambda_2 \text{vs.} \lambda_1 \neq \lambda_2$ is equivalent to testing both $H_0 : w_{12} = 1 \text{vs.} w_{12} \neq 1$ and $H_0 : p_{12} = 0.5 \text{vs.} p_{12} \neq 0.5$.

Building upon the C test, Ho [10], [11] introduced the Empirical Recurrence Rates (ERR) statistic which was later applied to a variety of fields by Tan *et al.* [12] and Ho and Bhaduri [13]. For a given input sequence X, and discretized equidistant points $h, 2h, \ldots lh, \ldots, Nh(= T)$ for a fixed choice of the unit time h, the ERR statistic $E_{X,l}$ is defined as

$$E_{X,l} = \frac{n_{Xl}}{lh} \tag{2}$$

where n_{Xl} = total number of occurrences for X in $(0, lh)$. The ERR statistic can be extended to an Empirical Recurrence Rates Ratio (ERRR) time series to measure (through a ratio of the two empirically observed rates) the amount of dependence between two time series X_1 and X_2. The time period should first be discretized into equidistant points $h, 2h, \ldots lh, \ldots, Nh(= T)$ for a fixed choice of the unit time h. The ERRR,$R_{X_1,X_2,l}$, at these time points can then be sequentially generated as

$$R_{X_1,X_2,l} = \frac{n_{Xl}}{n_{Xl} + n_{Yl}}, \quad \text{for } n_{Xl} + n_{Yl} > 0 \tag{3}$$

where n_{Xl} = Total number of occurrences for X_1 in$(0, lh)$; n_{Yl} = Total number of occurrences for X_2 in$(0, lh)$; and $l = 1, 2, \ldots, N$.

Thus an ERRR can be expressed as a ratio of two ERR's:

$$R_{X_1,X_2,l} = \frac{E_{X_1,l}}{E_{X_1+X_2,l}} = \frac{E_{X_1,l}}{E_{X_1,l} + E_{X_2,l}} \tag{4}$$

We must be careful to disregard a few initial points that make the denominator vanish (the burn-in period). If the rates are independent of time, the ERRRs are essentially tracks of the maximum likelihood estimators (MLEs) of p_{ij}s. Exploiting MLE's invariance, these can be used to find the MLEs of $w_{ij} = \lambda_i/\lambda_j$. Various time series vectors in applied science are plagued with numerous zero values. Examples include sandstorms or strong hurricane counts. This zero inflation, seasonally or otherwise, creates problems in stochastic analyses through unreliable parameter estimates and volatile forecasts. ERRR analyses provide an improved methodology. Through cumulation or the summing up of past values, it reduces the number of zeros. It generates ''pseudo-observations'' over quieter time regions. Additionally, its structure makes it bounded by 0 and 1. We will demonstrate below how the nature and strength of dependence between the two time series instances are contained in this statistic as well.

Algorithm 1 ERRR Calculation Based on Two Incoming Time Series

Input: A stream of $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots..$
Parameter: $sf(x)$: sampling frequency of the X series.
$sf(y)$: sampling frequency of the Y series.
Ensure: $sf(x) = sf(y)$
if $sf(x) = sf(y)$
then
/* variable declarations */
$sum(x)$: sum of the X values from x_1 to x_i
$sum(y)$: sum of the Y values from y_1 to y_i
$ERRR(X, Y)$: ERRR value from the X and Y series.
for all new (x_i, y_i) in stream **do**
$ERRR(X, Y) \leftarrow sum(x)/(sum(x) + sum(y))$
end for
else
$ERRR(X, Y) \leftarrow NA$
end if

B. MOTIVATING EXAMPLES

To elucidate ERRR's workings, we reiterate the artificial example studied by Ho and Bhaduri (2007). Let us assume that a pair of discrete time sequences is given by:

$$X_1 = 1, 0, 0, 2, 4, 2, 2, 4, 2, 0, 0, 0, 0, 0, 0, 2, 4, 2, 2, 4, 2, 0, 0, 0, 0, 0, 0, \tag{5}$$

$$X_2 = 2, 4, 2, 1, 0, 0, 0, 0, 0, 2, 4, 2, 2, 4, 2, 0, 0, 0, 0, 0, 0, 2, 4, 2, 2, 4, 2, \tag{6}$$

A passing glance confirms their inverse dependence: active periods of one sequence usually accompanying dormant

periods of the other. At times, negative correlations like these contribute to the overall dissimilarity in this particular instance, however, it simply affirms a horizontal shift. The measures we introduce later can weed out these shifts, both horizontal and vertical). The ERRR curve constructed out of this pair resembles a clear sinusoidal pattern, depicted in the first panel of Fig (1).

Owing to the similarity of their cumulative strengths, the oscillations are about 0.5. Put differently, $R_{X_1,X_2,l} = 0.5$ at a given time $t = l$ will indicate the two sequences are fairly equally active (or equally competitive) till that time while $R_{X_1,X_2,l} < 0.5$ will imply X_2 is more active. Similarly, $R_{X_1,X_2,l} > 0.5$ will mean X_1 is more active. Thus deviation from the baseline 0.5 in either direction suggests deviation from perfect independence and the existence of an underlying bond.

This can be further confirmed by another pair, which unlike the first, is directly related:

$$X_1 = 3, 4, 5, 6, 7, \ldots \tag{7}$$

$$X_2 = 1, 2, 3, 4, 5, \ldots \tag{8}$$

Here, the first series drags up the second along with itself and the generated ERRR curve, lying on one side of the 0.5 line shows clear monotonicity (the second panel of Fig (1)). We note that the first pair, though inversely related, were equally intense (the non-zero numbers were same for both). Ho and Bhaduri [15] examine other examples of inverse dependence with varying intensities leading to a baseline different from 0.5. In the present context of unearthing similarity, we shall not pursue such generalities,

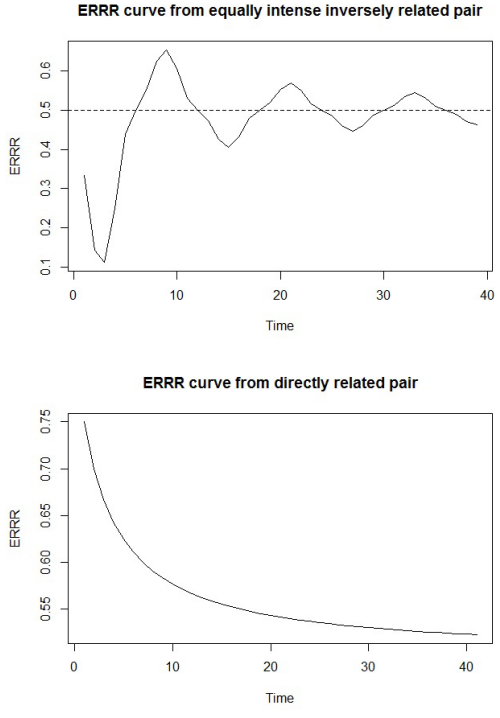

FIGURE 1. Behavior of ERRR curves under different artificial scenarios.

but instead reflect on the insights these two synthetic data sets have to offer: if two time series are negatively (inversely) dependent, the generated ERRR curve should exhibit a wavy pattern (if moreover, they are equally intense, then it should fluctuate about the 0.5 line), if they are positively (directly) dependent, the ERRR curve should show a monotonic trend. Later sections will describe how in many instances, this notion of dependence induces that of similarity.

C. EXPLOITING THE STRUCTURE OF ERRR

In section 3, we shall define distance measures that will quantify several features of the ERRR curve and aid our understanding of the mutual interplay between the participating time series. But even without those measures, the structure of ERRR is amenable to a wide array of important interpretations. Its usefulness in differentiating dependence from independence under mild parametric assumptions offers a glaring case in point.

For two gamma distributed variables $X \sim Gamma(m, \lambda)$ and $Y \sim Gamma(n, \lambda)$, it is known that the ratio $R = \frac{X}{X+Y}$, under the assumption of independence, will have a $Beta(m, n)$ distribution. Noting the similarity in structure, one can devise ERRR based tests to choose between:

$$H_0 : X \text{ and } Y \text{ are independent} \tag{9}$$

$$H_a : X \text{ and } Y \text{ are dependent} \tag{10}$$

where the critical region will be

$$\{R_{X,Y} \in [0, 1] : \beta_{1-\frac{\alpha}{2};m,n} \leq R_{X,Y} \leq \beta_{\frac{\alpha}{2};m,n}\} \tag{11}$$

with $\beta_{1-\frac{\alpha}{2};m,n}$ and $\beta_{\frac{\alpha}{2};m,n}$ being the lower and upper $\frac{\alpha}{2}$ point from a $Beta(m, n)$ density. This ''body'' rejection region follows since unlike traditional normality based tests, under the assumption of dependence (one may set $Y = X$ for instance, a case of perfect dependence), the fraction will tend to cluster around 0.5.

1) SIMULATION STUDY

To elucidate the test's performance in terms of power, we performed the following simulation study: 10^4 copies of two variables U and V were generated from an exponential distribution with rate 5 − ''correlation tracker''. 10^6 copies of another variable W were generated from an exponential distribution with rate ''correlation tracker''. New variables X and Y were defined as:

$$X = min(U, W) \tag{12}$$

$$Y = min(V, W) \tag{13}$$

and theoretically it can be shown that X and Y have exponential distributions with rates 5 and 5 and are positively correlated if:

$$0 < \text{correlation tracker} < min(5, 5) = 5 \tag{14}$$

The variable ''correlation tracker'' measures the amount of dependence between X and Y: higher its value, stronger is the dependence. Using the reproductive property of exponentials,

we have $\sum_{i=1}^{10000} X_i =^d \sum_{i=1}^{10000} Y_i \sim Gamma(10000, 5)$ and thus, under the assumption of independence, the ERRR statistic:

$$R_{X,Y} = \frac{\sum_{i=1}^{10000} X_i}{\sum_{i=1}^{10000} X_i + \sum_{i=1}^{10000} Y_i} \sim Beta(10000, 10000) \tag{15}$$

Using 10^6 simulations, $\alpha = 0.1$ and a critical region of the form (11), we have evaluated the probabilities of correct dependence identifications, condensed in the power curve shown in Fig (2).

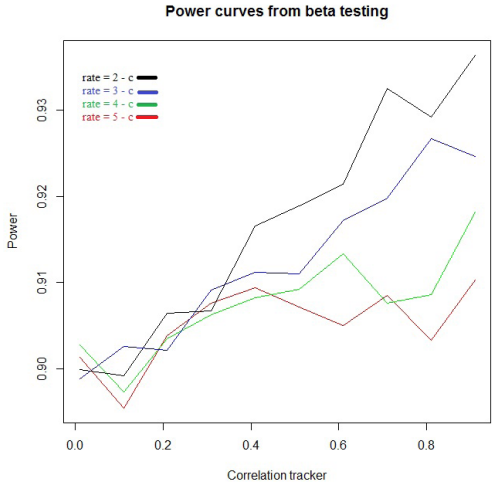

FIGURE 2. Power curve from Beta testing for independence.

To demonstrate stronger confirmation of the test, we sequentially changed the exponential rate from 5 to 4, 3, and 2. We found that the test is able to pick up even meager amounts of positive dependence with remarkable precision. As the correlation tracker increases, implying intensified positive dependence, the identification naturally becomes simpler, which justifies the overall increasing monotonicity of the power curve.

The Gamma parameterization, at first glance, might seem too restrictive. We point out that for non-standard distributions, the law of large numbers takes over both the numerator and the denominator of the ERRR statistic. Cauchy-like distributions (as the ratio of two normal-like densities) may then be employed as the null density. Hinkley [18], among others, provides excellent resources for that purpose. Noting that the present section intends to prove the utility of the form of ERRR through an easy example, we refrain from pursuing more intricate technicalities.

III. MEASURES AND INDICES FOR ERRR CURVES

A cursory glance at the ERRR curve suggests clues regarding the dependence dynamics: a wavy ERRR curve in general, implies inverse dependence while a monotonic ERRR curve suggests a direct dependence. Stronger quantification of these intuitions can, however, be had from the following indices, each of which enjoys metric-like properties. In the definitions to follow, $I(A)$ denotes the indicator of an event A, taking two values: 1 if the event A happens, 0 otherwise.

A. INDEX OF COMPETITIVENESS I_c

Given two time series X and Y of length n each, this index captures the proportion of times the generated ERRR curve lies above the 0.5 line. Formally, thus:

$$I_c^n(X, Y) = \frac{1}{n} \sum_{i=1}^{n} I\{R_i(X, Y) > 0.5\} \tag{16}$$

this is extremely useful in detecting shifts of the form identified by [4] among others. The underlying idea revolves around the notion that if one time series consistently dominates or is dominated by another, with or without maintaining a similar shape, the ERRR curve will consistently be on one side of the 0.5 line. For the pair depicted in the first panel of Fig (3), the Y series is created by adding a white noise of average magnitude 2 (and a very small variance, to preserve the shape) to an already existing time series X.

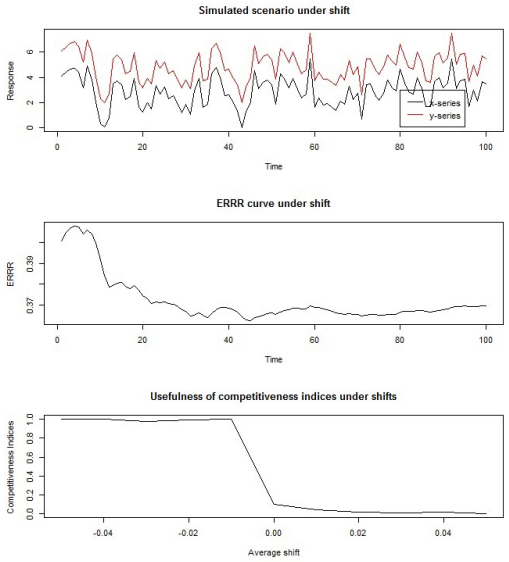

FIGURE 3. Shift detection using I_c index.

The resulting ERRR curve is shown in panel 2 and indicates a competitiveness index of 0, all the ERRR values being less than 0.5. Flipping the roles of X and Y, one could still preserve a shift translation, but could have an I_c index of 1. Extreme values of competitiveness indices close to 0 and 1 are thus indicative of the existence of possible shifts.

The measure has a metric-like property: $I_c^n(X, X) = 0$ since $R_i(X, X) = 0.5 \ \forall i = 1(1)n$, by definition. However, $I_c^n(X, Y) = 0$ does not necessarily imply $X = Y$ as the above example on shift suggests. To understand the sensitivity of this measure, we have paired several Y series, with various amounts of average shifts, to the constant X series shown in panel 1, found the ERRR curves and collected the resulting I_c indices in panel 3. As it is clear that the index picks up even negligible shifts with remarkable precision.

Algorithm 2 $I_c^n(X, Y)$ Calculation Based on Two Incoming Time Series

Input: A stream of $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots..$
Parameter: $sf(x)$: sampling frequency of the X series.
$sf(y)$: sampling frequency of the Y series.
Ensure: $sf(x) = sf(y)$
if $sf(x) = sf(y)$
then
/* variable declarations */
$sum(x)$: sum of the X values from x_1 to x_i
$sum(y)$: sum of the Y values from y_1 to y_i
$ERRR(X, Y)$: ERRR value from the X and Y series.
$IND(X, Y)$: logical, whether$ERRR > 0.5(= 1)$ or not (=0)
$I_c(X, Y)$: competitiveness index.
for all new (x_i, y_i) in stream **do**
$ERRR(X, Y) \leftarrow sum(x)/(sum(x) + sum(y))$
if $ERRR(X, Y) > 0.5$
then
$IND(X, Y) \leftarrow 1$
else
$IND(X, Y) \leftarrow 0$
end for
$I_c(X, Y) \leftarrow mean(IND(X, Y))$
else
$ERRR(X, Y) \leftarrow NA$
end if

B. INDEX OF EXTREMENESS I_e AND DESCRIPTIVE MEASURES

Visual similarity between two time series is often caused by an inherent correlation structure that makes one dependent on the other. Noting that the two terms "correlated" and "dependent" should not be confused, nor used interchangeably, we observe the Beta distribution related test to check for independence between X and Y detailed in the previous section. We assume that the X and Y series are purely random sequences in themselves, meaning that there exists zero correlation among the X or among the Y values. For practical time series, however, the existence of autocorrelation proves this is hardly the case, and a new measure, termed the *index of extremeness*, defined on the ERRR curve will prove beneficial in identifying dependence under this general case. We define it as:

$$I_e^n(X, Y) = \frac{1}{n}\sum_{i=1}^{n-2} I\{(R_{i+1} - R_i)(R_{i+2} - R_{i+1}) < 0\} \quad (17)$$

and it measures the proportion of extremes (i.e. both peaks and valleys) from the ERRR curve calculated from the pair under consideration. Like I_c, this index too follows metric-like properties:

a) $I_e^n(X, X) = 0$: This follows since $R_i(X, X) = 0.5$ $\forall i = 1(1)n$ and this ERRR curve is entirely devoid of peaks and valleys.

b) $I_e^n(X, Y) = I_e^n(Y, X)$: This follows since $R_i(X, Y) = 1 - R_i(Y, X)$ $\forall i = 1(1)n$ and under reflection about a constant, the number of peaks and valleys remain unaltered.

To conduct our simulation studies under dependence, we have generated 100 observations from each of two Autoregressive processes of order 1 (AR(1)):

$$X_t = \rho_1 X_{t-1} + U_t \quad (18)$$
$$Y_t = \rho_1 Y_{t-1} + V_t \quad (19)$$

such that $correlation(X_t, Y_t) = \rho$. If active periods of one series are usually accompanied by dormant periods of the other, then the two series are inversely dependent, and the ERRR curve is expected to be considerably wavy, thereby inflating the proportion of peaks and valleys over a period of stable flow. Care must be taken to ignore the first few oscillations (till around the 40th time point in this case) since they represent an unstable burn-in period, and their inclusion in the proportion calculation will be misleading. This scenario is depicted on the top panel of Fig (4), where the two AR(1) series share an extreme negative correlation of -0.9.

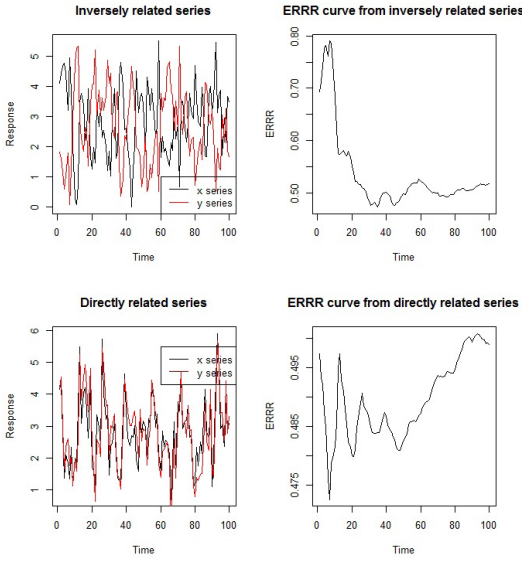

FIGURE 4. Behavior of ERRR curve under direct and inverse dependence.

On the other hand, if the two series behave similarly, then the ERRR series is expected to be a lot less wavy over the burn-in disregarded stable period and consequently, this proportion should go down. This is confirmed by the lower panel of Fig (4), where the two AR(1) series have been drawn with a high correlation of 0.9.

To confirm this pattern, we have simulated the pair using several ρ values between -1 and 1 and calculated the I_e index in each case. Figure (5) describes the findings. As the pair gets strongly positively related, the I_e index tends to fall.

Feature-based classification analyses, as detailed in the introduction, are prevalent in literature: Nanopoulos *et al.* [21] and Deng *et al.* [22] used control chart ideas on the average,

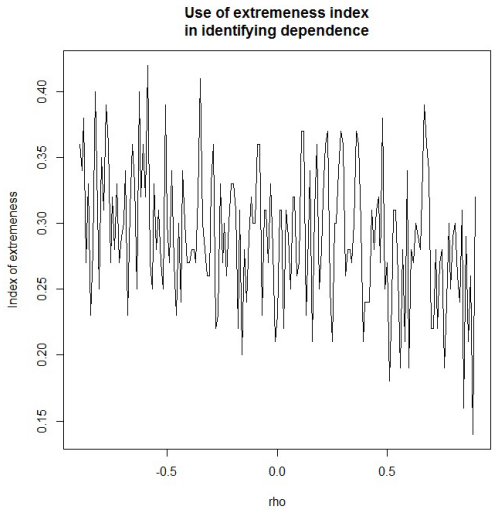

FIGURE 5. Dependence detection using I_e index using $AR(1)$ models.

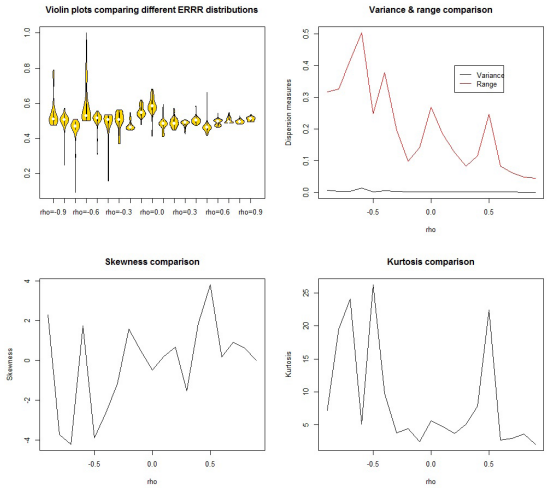

FIGURE 6. Plots of ERRR based methods using $AR(1)$ models.

standard deviation, skewness, and kurtosis of the time series. Features derived from wavelets and Fourier transforms have been examined by Morchen [23] while Wang *et al.* [24], [25] proposed thirteen summary features to characterize both univariate and multivariate time series. In a similar spirit, the distribution of ERRR values also contains useful information about the dependence structure. The violin plots graphed in the first panel of Fig (6) have been generated using the same $AR(1)$ process with different values of the correlation parameter.

These violin plots are essentially smoothed versions of histograms containing boxplots and kernel density estimates as additional measures. A change in the correlation parameter lead to several changes in the distribution of these ERRR values, the most notable one among them being a considerable drop in the dispersion among these values. These changes can be confirmed using the range and variance curves shown above.

To quantify the amount of symmetry in a distribution, one often calculates the coefficient of skewness defined as:

$$\gamma_1 = E[(\frac{X-\mu}{\sigma})^3] = \frac{\mu_3}{\sigma^3} \tag{20}$$

where μ and σ are the mean and standard deviation, respectively. For symmetric distributions such as the normal or uniform, this measure equals 0. Interestingly enough, through the skewness curve on panel 3, we observe a transition in the nature of symmetry of the ERRR distribution - as ρ increases, the distribution gradually changes from being negatively skewed (long left tail) to positively skewed (long right tail), with being roughly symmetric at independence.

Evaluating the kurtosis through

$$\gamma_2 = \frac{E(X-\mu)^4}{[E(X-\mu)^2]^2} = \frac{\mu_4}{\sigma^4} \tag{21}$$

is a useful way to garner information about the peakedness of the underlying density. Distributions such as the normal have this measure equal to 3. Panel 4 (Fig (6)) depicts how the intensity of dependence controls this feature of the ERRR density - higher the strength of dependence, sharper is the resulting ERRR distribution.

IV. SIMILARITY MEASURES FROM BOOTSTRAPPED TIME SERIES

The purpose of this section is to amplify our confidence in both the measures introduced above and the conclusions reached about the similarity of two time series vectors. For this, we resort to bootstrapping, a powerful and intuitive statistical tool introduced by Efron [28], primarily to study standard errors or parameter estimates obtained from independently and identically distributed realizations.

The time series we observed is just one instance of the infinitely many possibilities that could have stemmed from similar initial conditions. Our conclusions about similarity would be several times stronger if they can somehow withstand the test of this averaging over the numerous unobservable possibilities. Unless we have absolute knowledge about the process that governs the evolution (such as the $AR(1)$ model in the previous section), going back in time and observing another evolution of the process is not feasible. Rearranging or permuting the time series offers a structural alternative. Caution must, however, be exercised since rearranging every value indiscriminately will lead one to lose the underlying dependence structure (such as the autocorrelation) within time series. Hall [27] and Carlstein [26] thus generalized Efron's nascent idea into block bootstrapping, which was implemented recently by Ho and Bhaduri [15] in a related instance.

The process runs thus: fixing b as the block size, we will sample chunks of the original time series $Y_{I+1}, Y_{I+2}, \ldots, Y_{I+b}$ for $I \in \{0, 1, .., n-b\}$, chosen at random. Joining these blocks one after another will create a new time series with strikingly similar properties to the original one. To be precise,

assume that one can generate blocks $Y_{I_j+1}, Y_{I_j+2}, \ldots, Y_{I_j+b}$ for $j \geq 1$ continuously and create a new time series, $Y_1^*, Y_2^*, \ldots$ identical to

$$Y_{I_1+1}, Y_{I_1+2}, \ldots, Y_{I_1+b}, Y_{I_2+1}, Y_{I_2+2}, \ldots, Y_{I_2+b}, \ldots$$

The sequence of the first n values of this time series $Y_1^*, Y_2^*, .., Y_n^*$ will then constitute the block resample. For implementation purposes, sufficiently many I_js were chosen from $\{1, 2, .., n-b\}$ independently and with repetition. Technically, this is the moving blocks method.

Block bootstrapping is relevant and advantageous in the present context because by choosing blocks in their entirety, some resemblance regarding the dependence structure to the parent series can be maintained. Breakages can only happen at the joining points of the blocks. This method thus generalizes usual bootstrapping where the block size may be argued to be 1. The technique thus will be instrumental in detecting horizontal shifts.

To implement the ideas above, we shall keep the X series unaltered, while bootstrapng the Y series with the intention of answering such questions as: given X and Y are similar (or non-similar), could there be an instance when they might ever appear non-similar (or similar)? This is undeniably a stronger analysis. We have fixed the block size at $b = 10$, and for each simulation, we have found the summary measures from the ERRR curve resulting from the fixed X and bootstrapped Y, and we have done 10000 simulations. The distributions of the descriptive measures are shown in Fig 7 below:

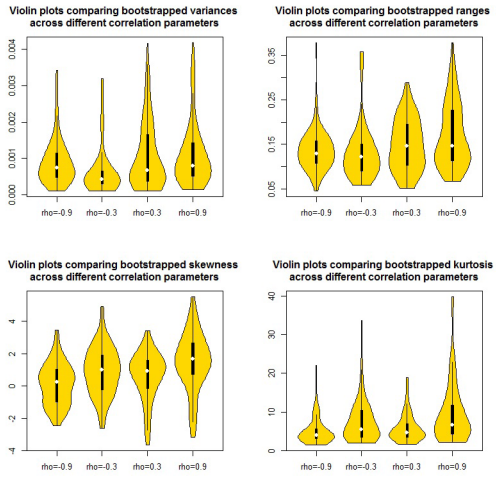

FIGURE 7. Distribution of descriptive measures from bootstrap generated ERRR curves using a block size of 10 and simulation strength of 10000.

If we examine the skewness panel for instance, we will find stronger credence in the fact that negative dependence between the two time series observations roughly implies a negatively skewed ERRR distribution (panel 3, Fig. 6). As the correlation parameter increases (panel 3, Fig. 7), the boxes get elevated implying it is difficult to construe an ERRR curve generated from two positively dependent series (probably thus similar in appearance) as being negatively skewed.

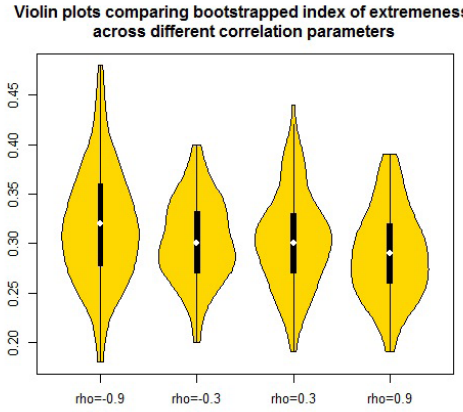

FIGURE 8. Distribution of I_e indices from bootstrapped time series.

It is also possible to depict the distribution of competitiveness or extremeness indices garnered from bootstrap generated ERRR series. Figure 8 below reveals this for the I_e index:

From the falling levels of the boxes and their shrinking spread with an increase in ρ, we prove that as the pair grows more dependent, the resulting ERRR curve grows less wave-like and situations insinuating otherwise are extremely unlikely.

V. REAL DATA ANALYSIS

The type of similarity and dissimilarity, or dependence and independence, that may abound between two time series vectors are numerous and controlling the correlation parameter between the two *AR* driven processes is only one method to generate such a type. While useful in understanding the way the proposed mechanism works, in reality, this approach would be impractical due to our lack of knowledge of the mechanism that drove the instances (fitting a time series model could be one alternative) and uncertainty due to the type(s) of internal dependencies.

Thus, to show the utility of ERRR and related measures, we have taken recourse to an extensive set of time series databases available at the UCR repository [8]. Each database houses a set of time series, with their classification labels known, undoubtedly an advantageous aspect with respect to calculating misclassification probabilities. Different databases contain different numbers of clusters and various lengths of the involved time series data sets. We choose a subset of these time series for training purposes (the same training set as has been mentioned in the repository); calculate the ERRR series and some of the desired related measures to understand how ''similarity'' and ''dissimilarity'' feel like in this particular instance (since not all types of ''dissimilarity'' can be tracked by a change in the ρ parameter); quantify the notion of distinction through some summary measures (such as percentile); and use it on the test set, to eventually calculate misclassification rates. In this work, we are concentrating on binary classification, which is why we have chosen five data sets: ''gunpoint,''

"ECG," "Yoga," "Wafer," and "Coffee," each containing two clusters.

For instance, for the data set titled "gunpoint," we have two classes and 200 time series of length 150 each. 50 of them were selected in the training set and the remaining 150 in the testing set. On the former, we have 24 time series in cluster 1, and 26 in cluster 2. We have done C_2^{24} ERRR computations on cluster 1 to get the "range" statistic, C_2^{26} computations on cluster 2, to get the same statistic, pooled the distributions together, to generate knowledge on how "similarity" is viewed through the lens of this statistic for this data set. This is graphed in the first panel of Figure 9.

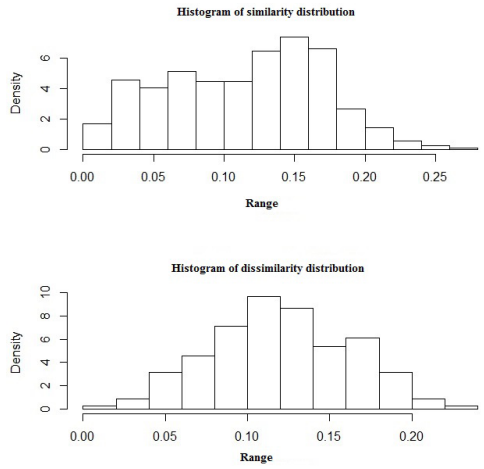

FIGURE 9. Distribution of range statistic for "gunpoint" data set.

Next, we performed 24×26 ERRR calculations, taking members from clusters 1 and 2 to quantify the dissimilarity distribution (panel 2) in this instance. We note that dissimilarity tends to inflate the range statistic, as demonstrated through simulations in the previous section: a greater proportion of observations in the dissimilarity distribution tend to cluster around the right tail, compared to that of the similarity distribution. Numerical summaries confirm this: the first quartile from the similarity distribution is 0.06935, compared to 0.09161 from the dissimilarity one. This difference is significant due to the fact that the ERRR construction bounds the range statistic within one unit.

One may use these numerical summaries to identify whether two time series in the test set belong to the same cluster. If the ERRR range for instance, from a pair, turns out to be in excess of 0.06935, we might say they belong to different categories and are dissimilar. To construct stronger cutoffs, however, we have employed bootstrapping techniques on the training set and have stuck to the range and extremeness index statistics while classifying time series in the different databases.

We have tested the method on five different datasets, using time series of varying lengths: these are "gunpoint," "ECG," "Yoga," "Wafer," and "Coffee," all containing two clusters.

According to experimental results based on these three real data sets, as shown in Figures 10 through 14, our ERRR

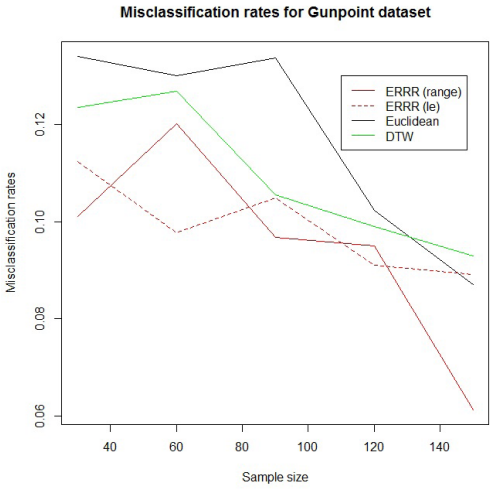

FIGURE 10. Misclassification rates for "gunpoint" data set.

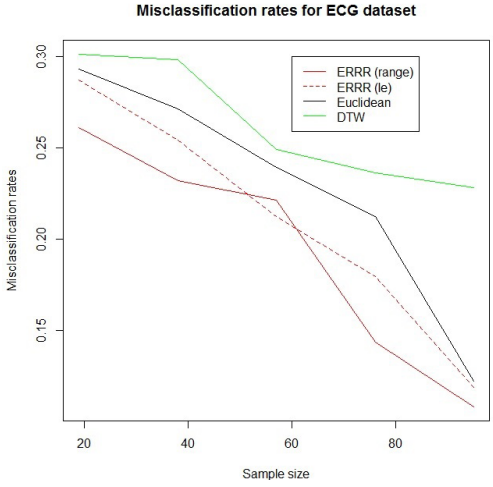

FIGURE 11. Misclassification rates for "ECG" data set.

based approach always perform better than its traditional counterparts such as the Euclidean metric and Dynamic Time Warping [19] in terms of smaller misclassification rates. In each of these figures, the misclassification rates for the ERRR based measures consistently lie below those from the Euclidean and Dynamic Time Warping technique. The Euclidean one, being the most primitive, and suffering from noise and misalignments described previously, gives the worst classification performance in most of the examples. In the "ECG" data set, where this is not the case, the problems with noise and misalignment are less severe. Dynamic Time Warping, in general, offers a marked improvement over the Euclidean metric. Between the ERRR based measures too, neither one consistently dominates the other, although we found that the I_e based extremeness index gives lower misclassification rates in most of the cases. This, arguably, is due to the range based measure's sensitivity to outliers, similar to a problem faced by the Euclidean metric. Measures capturing other descriptive properties of the ERRR distribution, are likely to perform better than the range.

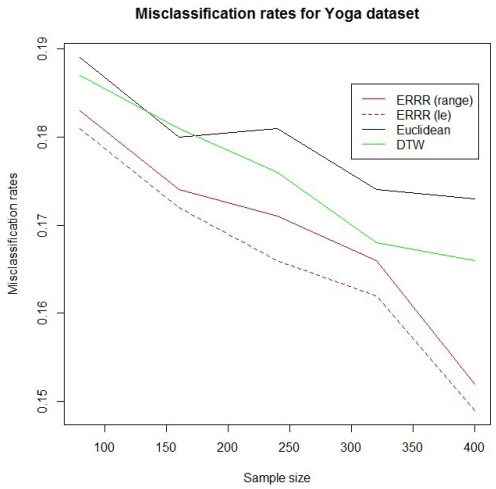

FIGURE 12. Misclassification rates for "Yoga" data set.

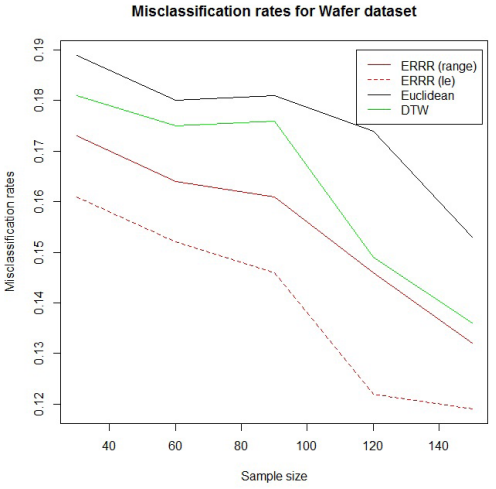

FIGURE 13. Misclassification rates for "Wafer" data set.

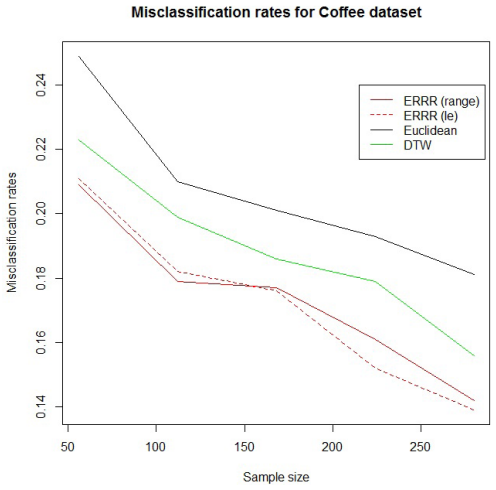

FIGURE 14. Misclassification rates for "Coffee" data set.

VI. CONCLUSION AND FUTURE WORK

This article popularized a nascent statistic termed Empirical Recurrence Rates Ratio and introduced a novel way to quantify the similarity between two time series. Three measures relying on the ERRR curve have been introduced, and many of summary measures on the values themselves have been employed in conjunction with a sophisticated method of block bootstrapping to suggest the superiority of the technique over established classification methods such as Euclidean Distance or Dynamic Time Warping. One could also combine more sophisticated feature selection [29]–[31] and classification methodologies [33]. We avoided the problem of feature correlation [32]. We have also demonstrated how, unlike others, this tool can differentiate dependence from independence between a pair of time series values, and at the same time stay robust on issues like noise and the presence of outliers. The measures defined on ERRR have several metric-like properties and are a useful addition when classifying time series databases. Currently, we are trying to extend similar measures and indices to non-binary classification problems. Serra and Arcos [20] provides multiple other similarity measures. ERRR based analyses will be comparable to them as well. The method avoids the need of decomposing the time series into seasonality components based on local polynomial regression [35], de-noising it using methods shown in [37], or weeding out outliers using Local Outlier Factors [40]. All calculations can be done using packages in R. For instance, measuring the proportion of extremes could be done using the *quantmod* package [36]. An easy and intuitive construction, computational efficiency, and effectiveness in identifying both structural similarity and underlying dependence make ERRR applicable to a wide range of classification problems.

REFERENCES

[1] T. W. Liao, "Clustering of time series data—A survey," *Pattern Recognit.*, vol. 38, no. 11, pp. 1857–1874, Nov. 2005.
[2] B. D. Fulcher and N. S. Jones, "Highly comparative feature-based time-series classification," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 12, pp. 3026–3037, Dec. 2014.
[3] X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, and E. Keogh, "Experimental comparison of representation methods and distance measures for time series data," *Data Mining Knowl. Discovery*, vol. 26, no. 2, pp. 275–309, 2013.
[4] U. Mori, A. Mendiburu, and J. A. Lozano, "Similarity measure selection for clustering time series databases," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 1, pp. 181–195, Jan. 2016.
[5] P. Esling and C. Agon, "Time-series data mining," *ACM Comput. Surv.*, vol. 45, no. 1, Nov. 2012, Art. no. 12.
[6] E. Keogh and S. Kasetty, "On the need for time series data mining benchmarks: A survey and empirical demonstration," *Data Mining Knowl. Discovery*, vol. 7, no. 4, pp. 349–371, 2003.
[7] L. Wang, X. Wang, C. Leckie, and K. Ramamohanarao, "Characteristic-based descriptors for motion sequence recognition," in *Proc. 12th Pacific–Asia Conf. Adv. Knowl. Discovery Data Mining*, 2009, pp. 369–380.
[8] E. Keogh, X. Xi, L. Wei, and C. A. Ratanamahatana. (2006). *The UCR Time Series Classification/Clustering Homepage*. [Online]. Available: http://www.cs.ucr.edu/~eamonn/time series data/
[9] J. Przyborowski and H. Wilenski, "Homogeneity of results in testing samples from poisson series: With an application to testing clover seed for dodder," *Biometrika*, vol. 31, nos. 3–4, pp. 313–323, 1940.
[10] C.-H. Ho, "Empirical recurrence rate time series for volcanism: Application to Avachinsky volcano, Russia," *J. Volcanol. Geothermal Res.*, vol. 173, nos. 1–2, pp. 15–25, 2008.
[11] C.-H. Ho, "Hazard area and recurrence rate time series for determining the probability of volcanic disruption of the proposed high-level radioactive waste repository at Yucca Mountain, Nevada, USA," *Bull. Volcanol.*, vol. 72, no. 2, pp. 205–219, 2010.

[12] S. Tan, M. Bhaduri, and C.-H. Ho, "A statistical model for long-term forecasts of strong sand dust storms," *J. Geosci. Environ. Protection*, vol. 2, pp. 16–26, Jun. 2014.
[13] C.-H. Ho and M. Bhaduri, "On a novel approach to forecast sparse rare events: Applications to Parkfield earthquake prediction," *Natural Hazards*, vol. 78, no. 1, pp. 669–679, 2015.
[14] C.-H. Ho, G. Zhong, F. Cui, and M. Bhaduri, "Modeling interaction between bank failure and size," *J. Finance Bank Manage.*, vol. 4, no. 1, pp. 15–33, 2016.
[15] C.-H. Ho and M. Bhaduri, "A quantitative insight into the dependence dynamics of the Kilauea and Mauna Loa Volcanoes, Hawaii," *Math. Geosci.*, vol. 49, no. 7, pp. 893–911, 2017.
[16] L. Ye and E. Keogh, "Time series shapelets: A new primitive for data mining," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 947–956.
[17] T. Rakthanmanon and E. Keogh, "Fast shapelets: A scalable algorithm for discovering time series shapelets," in *Proc. SIAM Conf. Data Mining*, 2013, pp. 668–676.
[18] D. V. Hinkley, "On the ratio of two correlated normal random variables," *Biometrika*, vol. 56, no. 3, pp. 635–639, Dec. 1969.
[19] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proc. KDD Workshop*, 1994, pp. 359–370.
[20] J. Serrà and J. L. Arcos, "An empirical evaluation of similarity measures for time series classification," *Knowl.-Based Syst.*, vol. 67, pp. 305–314, Sep. 2014.
[21] A. Nanopoulos, R. Alcock, and Y. Manolopoulos, "Feature-based classification of time-series data," in *Information Processing and Technology*. Commack, NY, USA: Nova, 2001, pp. 49–61.
[22] H. Deng, G. Runger, E. Tuv, and M. Vladimir, "A time series forest for classification and feature extraction," *Inf. Sci.*, vol. 239, pp. 142–153, Aug. 2013.
[23] F. Morchen, "Time series feature extraction for data mining using DWT and DFT," Philipps-Univ. Marburg, Marburg, Germany, Tech. Rep. 33, 2003.
[24] X. Wang, K. Smith, and R. Hyndman, "Characteristic-based clustering for time series data," *Data Mining Knowl. Discovery*, vol. 13, no. 3, pp. 335–364, 2006.
[25] X. Wang, A. Wirth, and L. Wang, "Structure-based statistical features and multivariate time series clustering," in *Proc. IEEE Int. Conf. Data Mining*, Oct. 2007, pp. 351–360.
[26] E. Carlstein, "The use of subseries values for estimating the variance of a general statistic from a stationary sequence," *Ann. Statist.*, vol. 14, no. 3, pp. 1171–1179, 1986.
[27] P. Hall, "Resampling a coverage pattern," *Stochastic Processes Appl.*, vol. 20, no. 2, pp. 231–246, 1985.
[28] B. Efron, "Bootstrap methods: Another look at the jackknife," *Ann. Statist.*, vol. 7, no. 1, pp. 1–26, 1979.
[29] A. K. Jain, R. P. W. Duin, and J. C. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.
[30] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Jan. 2003.
[31] I. Guyon, C. Aliferis, and A. Elisseeff, "Causal feature selection," in *Computational Methods of Feature Selection* (Data Mining and Knowledge Discovery Series). Boca Raton, FL, USA: CRC Press, 2007 pp. 63–85.
[32] B. D. Fulcher, M. A. Little, and N. S. Jones, "Highly comparative time-series analysis: The empirical structure of time series and their methods," *J. Roy. Soc. Interface*, vol. 10, no. 83, p. 20130048, 2013.
[33] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.
[34] G. E. Batista, X. Wang, and E. J. Keogh, "A complexity-invariant distance measure for time series," in *Proc. SIAM Int. Conf. Data Mining*, vol. 31, 2011, pp. 699–710.
[35] R. B. Cleveland, W. S. Cleveland, and I. Terpenning, "STL: A seasonal-trend decomposition procedure based on loess," *J. Off. Statist.*, vol. 6, no. 1, pp. 3–73, 1990.
[36] J. A. Ryan. (2013). *Quantmod: Quantitative Financial Modelling Framework. R Package Version 0.4-0*. [Online]. Available: http://CRAN.R-project.org/package=quantmod
[37] T. Kohler and D. Lorenz, "A comparison of denoising methods for one dimensional time series," Tech. Rep., 2005. [Online]. Available: http://www.math.unibremen.de/zetem/DFGSchwerpunkt/preprints/orig/lorenz20051dreport.pdf
[38] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *J. Mach. Learn.*, vol. 85, no. 3, pp. 333–359, Dec. 2011.
[39] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.
[40] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2000, pp. 93–104.

MOINAK BHADURI is currently pursuing the Ph.D. degree in statistics with the Department of Mathematical Sciences, University of Nevada at Las Vegas, Las Vegas.

His research interests include point processes, repairable systems, anomaly detection, and clustering.

JUSTIN ZHAN is currently the Director of Big Data Hub and a Faculty Member with the Department of Computer Science, Howard R. Hughes College of Engineering, University of Nevada at Las Vegas.

His research interests include big data analytics, information assurance, and biomedical computing.