Mining Interesting Temporal Patterns: An Information Theoretic Approach

Avleen Singh Bijral\textsuperscript{1} and Girish K. Palshikar\textsuperscript{2}

\textsuperscript{1} Department of Computer Science, University of Colorado, Boulder 80309, USA.
avleen.bijral@colorado.edu,
\textsuperscript{2} Tata Research Development & Design Centre, 54 B Hadapsar Industrial Estate,
Pune 411 013, India.
gk.palshikar@tcs.com

Abstract. Interesting patterns in temporal databases have equivalently been treated in literature as novelties or surprises. In this paper we propose a Information Theoretic framework for detecting interestingness that can handle most subjective notions of what might be interesting in temporal data. Based on this framework, a mutual information measure is derived using kernel based nonparametric probability density estimation techniques to assess interestingness of a temporal sub-sequence given another. This measure is then utilized in a sliding windows algorithm to detect all possible interesting sub-sequences with respect to a reference sequence. The algorithm's effectiveness is shown to be independent of the parameter choice for the density estimates and its advantages over other prominent techniques are discussed.

1 Introduction

The problem of finding interesting patterns in time series data has been formulated in literature as one of finding novelties, anomalies, faults etc [2], [4], [8]. The word interesting in our discussion includes all that can be novel, surprising or anomalous. Our notion of interestingness is different from the one in [3] where interestingness is related to the novelty of a discovered data mining rule. It remains a challenging problem because of a lack of a concrete definition as to what constitutes as interesting. Each of the aforementioned papers defines a notion of interestingness and go on to describe algorithms based on these. Such approaches while automated are only partially effective and the results cannot be universally valid as they have been devised keeping in mind the author’s notion of what might be novel. Moreover they cannot be expected to include all possibilities, some of which may be domain specific or completely subjective.

The issue, we feel, is then to be understood and tackled at various sub-levels. A likewise attempt [15] has been made to understand the emergence of numerous clustering algorithms. This is related to the many definitions(or inductive principles) of what constitutes a cluster. This diversity resulted in numerous clustering algorithms that correspond to one of the inductive principles, for e.g. the total squared error or the entropy of a cluster. Similarly there is a need to establish formally the various inductive principles that underlie the problem of interestingness detection. It is within the context of a fixed inductive principle that the benchmarks then can be established and various techniques can be compared. An algorithm can be then deemed superior if it produces
better results in one particular criterion and handles more than one of these principles. It seems unfair to test an algorithm with another’s notion of interestingness, as seems to be the case with many of the papers concerned with the problem.

Instead of introducing a new definition of interestingness in a time-series we define various inductive basis and propose to tackle some of these through a controlled and parameterized probabilistic approach.

The various inductive principles defining the notion of interestingness can be:

1. **Pattern Frequency**: In this case both rare and frequently occurring patterns can be considered interesting.

2. **Global Deviation**: If a pattern is a deviation from a certain global trend or structure, it could be considered.

3. **Outliers**: Any set of points which are indicative of an extreme deviation from a regularly occurring trend is an obvious candidate.

In this paper we tackle inductive principle (2) as we feel it is the one which is the most representative form of the interestingness problem. In this form it has been dealt with in most earlier work especially in [2], [4], [10]. We develop an information theoretic approach for the mentioned inductive principles using the non-parametric probability density estimate of points across the time series.

We then test our approach on time series datasets obtained from the UCR time series archive [17] and also on real time physiological datasets [18] available for public download and display the results for the datasets in the appendix.

## 2 Related Work

The task of detecting interesting sub-sequences in time series data has been actively researched in the areas of nonlinear dynamics, biology and astronomy etc. Researchers in these areas are interested in any aberrations or deviations from regularities of form or structure, which naturally leads to the idea of some form of probabilistic occurrence of such events. This makes necessary the inclusion of Shannon’s information theoretic measures to mathematically formalize the concept of occurrence of an event or an abnormality in the analyzed data.

Both [6] and [9] use information theoretic entropy measures to detect dynamical changes in the time series by studying variations in the average Information content of patterns across the data. In [9], the information measure is introduced to naturally value the degree of interestingness of each occurrence of a pattern as a continuous and monotonically decreasing function of its probability of occurrence. This is directly related to the degree of interestingness of a pattern as formulated in the concept of information gain. This approach is limited in the sense, that its definition of interestingness is restricted to infrequent occurrences and only represents one aspect of what could be interesting in a complex time series.
However, we feel that our approach is most similar to [1], at least intuitively, wherein the extent of dissimilarity between two strings \(x\) (the reference string) and \(y\) (test string) is measured by the change in the compression properties of the strings before and after concatenation. This change then, is used as a measure of the degree of the anomalism. The only difference being that, we employ the change in the probability density estimates of the reference string after concatenation with the test string as a measure of its interestingness.

3 Information Theoretic Interestingness Model

The precepts of information theory intuitively conform to our definition of interestingness since they derive measures which essentially measure information content associated with event(s). Such measures could be then interpreted as evaluating the degree of interestingness when an event or as in our case, an unlikely or interesting pattern occurs. In this section we develop from the concept of mutual information (MI) a measure to assess the degree of interestingness of a region in a time series in contrast to a reference region. Much of the element of interestingness in a time series can be related to a deviation of some sort from the naturally occurring regularity or structural characteristic of the entire time series.

Consider two subsequences \(X_1 = < x_1, x_2, ..., x_{N_1} >\) and \(X_2 = < x'_1, x'_2, ..., x'_{N_2} >\) of length \(N_1\) and \(N_2\) respectively, of a given time series as shown in figure 1. We can think of \(X_1\) as an input to a stochastic system with \(X_2\) as output. This can be interpreted as measuring the amount of information about \(X_2\) already contained in \(X_1\). Equivalently it can be viewed as a measure of change brought about in the statistical properties of a subsequence \(X_1\) when presented with a subsequence \(X_2\) in its entirety. Such a notion is intuitively related to the concept of interestingness viewed as a deviation from the expected. Borrowing from the concepts of information theory, we derive a
non-parametrical MI measure, which is an indicator of the statistical deviation between two given sequences. Mutual information as discussed in [13] and [16], is a measure for global dependence and it is independent of the model of dependence assumed, thereby making it suitable for a temporal pattern-matching task.

**Definition 1.** For a given subsequence $X_1$, the Shannon entropy is given by

$$H(X_1) = -\sum_{i=1}^{N_1} p(x_i) \log(p(x_i)). \quad (1)$$

Where $p(x_i)$ is the probability density estimate of the $i$th data point belonging to the sequence $X_1$.

**Definition 2.** After introducing the subsequence $X_2$, the conditional entropy of $X_1$ given $X_2$ is given by

$$H(X_1|X_2) = -\sum_{i=1}^{N_1} p(x_i|X_2) \log(p(x_i|X_2)). \quad (2)$$

Where $p(x_i|X_2)$ is the conditional probability density estimate of $i$'th data point $x_i$ of $X_1$ given the entire subsequence $X_2$.

**Definition 3.** The mutual information measure is given by

$$MI(X_1;X_2) = H(X_1) - H(X_1|X_2). \quad (3)$$

as the change in uncertainty of $X_1$ after the subsequence $X_2$ is introduced

Substituting in equation (3) the expressions for $H(X_1|X_2)$ and $H(X_1)$ we can obtain the mutual information between the two subsequences $X_1$ and $X_2$ as

$$MI(X_1;X_2) = \sum_{i=1}^{N_1} \left( p(x_i|X_2) \log(p(x_i|X_2)) - p(x_i) \log(p(x_i)) \right). \quad (4)$$

This measure then can be used as an estimate of the relative interestingness of a given sequence of points $X_2$ given a sequence $X_1$. This notion resembles [2] in that interestingness is concept relative to existing or current knowledge. In the mentioned work, current knowledge is represented as a mathematical model of data, whereas we use probability estimates to represent existing information and then relative to this information we assess the interestingness.

In the next section we describe the calculation of above mentioned probabilities and information measures using non-parametric estimates obtained by kernel methods.
4 Interestingness Measure Estimation

A kernel approach to probability density transformation of a time series can be looked at as piecewise segmentation of the temporal series into shapes of the chosen Kernel function. Density estimates can be very useful for our problem as they give valuable indication of such features as skewness and multimodality in data, and such features could translate more or less directly to interesting regions in our search space. We employ both the standard kernel density and the generalized nearest neighbor estimate for density evaluation. Consider a scalar time series realization \( x_t \), where \( t = 1 \ldots N \), where \( N \) is the number of time steps considered and let \( x_i \) be the \( i \)'th data point of the time series. The generalized \( k \)th nearest neighbor estimate for the given data is defined by

\[
P(x_t) = \frac{1}{Nd_k(x_t)} \sum_{i=1}^{N} K\left( \frac{x_t - x_i}{d_k(x_t)} \right).
\]

(5)

where

\[
d_k(x_t) = |x_t - x_{t+k}|.
\]

(6)

and the standard kernel density estimate is given by

\[
P(x_t) = \frac{1}{Nh} \sum_{i=1}^{N} K\left( \frac{x_t - x_i}{h} \right).
\]

(7)

It can be seen at once that \( K(x) \) is precisely the kernel estimate evaluated at time \( t \) with window width \( d_k(x_t) \) or \( h \) if the standard kernel estimate is used, where \( d_k(x_t) \) is the absolute difference of the \( k \)th nearest neighbor(i.e. in time steps) from a particular instance \( x_t \) and \( h \) is the kernel function width. Here the overall amount of smoothing is governed by the choice of the integer \( k \) or \( h \), but the window width used at any particular point depends on the density of observations near that point, so that data points in regions where the data is sparse will have flatter kernels associated with them. This representation is suitable for our purposes considering the continuously varying density regions of a time series. The choice of the kernel function determines the shape of the "bumps" placed along the data and doesn’t significantly alter the statistical properties of the estimates obtained [14]. Since we are only interested in relative statistical deviation of two sequences, the choice of parameters for our density estimates do not significantly alter the results obtained. For our purpose we chose the Epanechnikov kernel function of the form

\[
K(x) = \begin{cases} 
\frac{3}{4}(1 - x^2) & \text{for } |x| \leq 1 \\
0 & \text{Otherwise.}
\end{cases}
\]

(8)

For the example described in the the previous section, for subsequences \( X_1 \) and \( X_2 \)
with sample space size $N_1$ and $N_2$ respectively we evaluate the kernel probability density estimates for $X_1$ as

$$P_{X_1}(x) = \frac{1}{N_1b} \sum_{i=1}^{N_1} K \left( \frac{x - x_i}{b} \right).$$ (9)

For mutual information we also need the conditional probability density

$$P_{X_1|X_2}(x) = \frac{1}{(N_1 + N_2)b} \sum_{i=1}^{N_1 + N_2} K \left( \frac{x - x_i}{b} \right).$$ (10)

Where $b = h$ or $d_k(x)$ according to the choice of the kernel estimation technique. This is nothing but the altered probability estimate for $X_1$ after appending it with subsequence $X_2$. From these estimates we can then evaluate the mutual information from equation (4) as an interestingness measure between two reference series with sample sizes $N_1$ and $N_2$.

![Fig. 2. (MI) as an Interestingness measure](image)

To understand the intuition behind our non-parametric mutual information as an interestingness measure consider an extended version of figure 1 in figure 2. After the break lines two possible sub-sequences $X_2$ and $X_3$ are shown, both leading to the final sub-sequence $X_4$. Our non-parametric estimate of mutual information is essentially an indicator of the change brought about in the density estimates of the reference subsequence when another sequence is introduced. In figure 2, subsequence $X_3$ will have little or no effect on density estimates of subsequence $X_1$, since the range of values in $X_3$ is beyond that of $X_1$ and therefore not much information about $X_1$ is contained.
FindInteresting \((\text{time-series } X_t, \text{ reference-sequence } X_{\text{ref}}, \text{ window-size } W, \text{ interesting-set } \{ I \}, \text{ step-size } S, \text{ threshold } \theta)\) 

\[ I := \emptyset \quad NI := \emptyset \]
\[ P_{X_{\text{ref}}}[1..|X_{\text{ref}}|] := \text{ProbDensityEstimate(\text{time-series } X_{\text{ref}})} \]

\[ \text{FOR } i := 1 \text{ TO } |X_t| - |X_{\text{ref}}| - 1 \]
\[ \text{IF } (\text{MutInf } (X_{\text{ref}}, X_{1+(i-1)\ast S..W+(i-1)\ast S}) \leq \theta) \]
\[ I := I \cup X_{1+(i-1)\ast S..W+(i-1)\ast S} \]
\[ \text{ELSE} \]
\[ NI := NI \cup X_{1+(i-1)\ast S..W+(i-1)\ast S} \]

\[ \] Table 1. Interesting Pattern Search

\begin{align*}
\text{MutInf } (\text{reference-sequence } X_{\text{ref}}, \text{ subsequence } X) & \{ \\
\text{MI} & := 0.0 \\
P_{X_{\text{ref}}|X}[1..|X_{\text{ref}}|\ast|X|] & := \\
\text{ProbDensityEstimate(\text{time-series } X_{\text{ref}}, X)} \] \\
\text{FOR } i := 1 \text{ TO } |X_{\text{ref}}| \\
\text{MI} & := \text{MI} + (P_{X_{\text{ref}}|X}[i] \ast \log(P_{X_{\text{ref}}|X}[i]) - P_{X_{\text{ref}}}[i] \ast \log(P_{X_{\text{ref}}}[i])) \\
\} \\
\] Table 2. Mutual Information Calculation

in \(X_3\), hence mutual information will be very low. And thus low MI estimates would indicate a higher degree of interestigness. However the subsequence \(X_2\) is contained in \(X_1\) and hence it would to a certain degree alter the density estimates, but certainly less so than the region \(X_4\) where the statistical similarity between the two subsequences is obvious. MI can then be used to assesses the degree of interestigness or confidence level as discussed in [2] of any given pattern as compared to a reference series.

The algorithm used to enumerate interesting regions in the time series is described in table 2 along with the mutual information calculation in table 3. For a given step and window size \(S\) and \(W\) respectively, we slide the window across the time series comparing each window to the reference series and assign it a degree of surprise. Then according to the threshold store the pattern window in the Interesting or the Non-interesting set.
5 Experiments

We demonstrate the performance of our algorithm on artificially generated and standard measured data. We conduct experiments on a generated dataset similar to the one used in [2], to compare performance. We also test the algorithm on the annual Sunspots data and two real public domain datasets.

5.1 Parameter Setting

We chose the following parameters for all our experiments:

1. Number of Nearest neighbors \((k)\): 4
2. Kernel function width \((h)\) : 0.1

To begin with we evaluated the kernel density estimates for the individual datasets using the above values. We also conducted experiments with the Gaussian kernel function and different values of \(k\). Instead of using the nearest neighbor function \(d_k(x)\) as the width of the kernel function we also used the standard kernel density estimate (4). The results, as we found out were more or less independent of the choice of these values. The choice of \(k\) and \(h\) affects the probability estimates, but as discussed before our measure is an indication of relative disparity between any two given sequences and hence more or less independent of the parameter choice.

Generated Data  The artificial time series was generated by the following equations

\[
Y_1(t) = \sin\left(40\frac{\pi}{N} t\right) + n(t).
\]  

\[
Y_2(t) = \sin\left(40\frac{\pi}{N} t\right) + n(t) + e(t).
\]

\[
e(t) = \begin{cases} 
n(t) & t \in [600,630], \\
0 & \text{Otherwise.}
\end{cases}
\]

As shown in figure 3 in the appendix, the first 400 points are chosen as the reference series. The added distortion represents an outlier scenario, wherein a set of points is way off the expected range of values. As shown in the figure, with a window size \(W = 30\) and step size \(S = 30\), we demarcate the most interesting region found by our algorithm, the one corresponding to lowest MI value. As we can see our algorithm successfully marks the most interesting sequence of size \(W\). We also tested our algorithm on an 800 data point noisy sine waveform with an inserted anomaly as done in [4] by halving the period in the region between two points. Our algorithm was able to detect...
Fig. 3. Results on noisy sine waveform with  a) a period anomaly  b) added distortion

this introduced anomaly as the most interesting region with the lowest MI as shown in the lower part of figure 3. The interestingness is depicted by the thickness of the line.

The window size can be set by the user based on familiarity with the dataset and/or the nature of interesting pattern search. It is not a burdensome parameter and optimal values can be arrived at with relative ease. The step size is an optional parameter available, which can be used to arrive at exact location of the anomaly by clearly outputting the start and end points. The case where $S = W$, is the non-overlapping sliding window search which we use for all datasets.

**Standard Datasets** We also conducted interestingness search on the Sunspot data obtained from UCR Time Series archive [17].

The Sunspots data is the annual measurement of sunspot numbers from 1700-1988. It presents an example of a novelty embedded within the range of the data yet peculiar in structural characteristics and hence interesting to the astronomer. By visual inspection
of figure 4, we can clearly separate the most interesting region from the rest of the series. With points 1-90 being designated as the reference sequence, a window size of 20 and a step size \( S = W \), the output of our algorithm exactly matched the results of visual detection. The region marked with the thickest lines represents the sequence with the lowest MI value and hence the most interesting fragment of the time series.

**Real World Datasets** To include real world examples, we examined the annotated MIT-BIH Noise Stress Test database and the Santa Fe Time Series competition Chest volume (respiration force) data. We extracted 5000 points from the MIT-BIH data, which included the most interesting event. The first 600 points were taken as the reference series and with a window size of 100, we were able to detect the most interesting event as shown in figure 6.

For the Santa Fe competition data we took a 10000 point sample which included the three most interesting events. For a window size 150, we were able to detect the interesting events as shown in figure 7. The relative decrease in the line thickness is an indication of lessening degree of interestingness.

**Algorithmic Complexity** The interesting pattern detection algorithm is \( O(|X|^2) \) where \( |X| \) is the size of the time series. The bulk of the computation is in the probability density estimation part, which using elementary methods is inherently \( O(N^2) \); experiments with faster non-parametric estimation techniques like [7] and [12] are currently on. Since our work was only aimed at demonstrating new interestingness measures, we stuck to the basic methods for obtaining density estimates instead of newer faster methods, to retain focus on the problem.
Fig. 5. Results on ECG data.

Fig. 6. Results on Santa Fe Competition data
6 Comparison

We compared our technique with the Novelty detection algorithm of [2]. While it worked well with the artificial noise data, it couldn’t detect the interesting sub-sequences in the Sunspot data. This technique failed to work for the structural surprise in the Sunspot data. One major disadvantage with this approach is that careful parameter tuning is required to make it work, unlike our technique.

Other prominent approaches like [8] and [11] fail to detect the simple period anomaly added to the artificial sine wave in figure 4b) as demonstrated in [4]. These approaches have a limited notion of surprise, mostly limited to certain dramatic shifts in the signal. They wouldn’t be effective with unusual patterns within the signal range. As mentioned in [4], the wavelet tree [11] approach will not work for the anomaly in the ECG data of figure 6, since wavelet techniques are invariant to transformation of the kind.

7 Conclusions and Future Work

In this paper we present mutual information as a interestingness measure for sub-sequences in time series data, which as a part of a framework for interestingness can include most existing definitions of "surprise", "novelty" etc. We can also minimally alter this technique to retrieve similar patterns given a particular sequence by simply evaluating the MI between them since it is a measure of probabilistic similarity of two sequences. So employed this measure can also handle the pattern frequency notion of interestingness as defined in the introductory section. As an extension, faster ways of evaluating density estimates could be used and then experiments could be conducted on very large time series. We are also working on automating approximate reference series extraction; evaluation of various statistical sampling techniques is currently on. We also plan to extend the technique to multiple as well as multi-dimensional time series.

References