MTH 517 COURSE PROJECT

Anomaly Detection in Time Series

Aditya Degala - 12051 Shreesh Laddha - 12679

1 Introduction

In data mining, anomaly detection (or outlier detection) is the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset. Typically the anomalous items will translate to some kind of problem such as bank fraud, a structural defect, medical problems or finding errors in text. This technique is especially useful in the context of abuse and network intrusion detection, where the interesting objects are often not rare objects, but unexpected bursts in activity. This pattern does not adhere to the common statistical definition of an outlier as a rare object, and many outlier detection methods (in particular unsupervised methods) will fail on such data, unless it has been aggregated appropriately. Instead, a cluster analysis algorithm may be able to detect the micro clusters formed by these patterns.

In this dosument we investigate the problem of anomaly detection for univariate time series. Although there has been extensive work on anomaly detection , most of the techniques look for individual objects that are different from normal objects but do not consider the sequence aspect of the data into consideration. Such anomalies are also referred to as point anomalies. If the data is treated as a collection of amplitude values, while ignoring their temporal aspect, the anomaly cannot be detected.

Time series data are common in many application areas including medical and environmental sciences, aerospace and industrial engineering, and finance and agriculture. Detecting emergent anomalies in time series data provides significant information for each application. For example, anomaly detection in electrocardiogram (ECG) signals or in aircraft health management saves lives, detecting anomalies in disease incident provides useful information about the possibility of occurring outbreaks, and in industrial process it may help to diagnose the incident faults. In time series, anomaly can be considered as the occurrence of any unexpected changes in a subsequence of data. The term "unexpected change" makes sense when we compare the available pattern in a subsequence with the existing patterns in the entire time series. As the result, one common approach to anomaly detection in time series is the use of a fixed length sliding window and generating a set of subsequences of time series. In the next step, one may use different techniques to detect and characterize anomalies i.e. assigning an anomaly score to each subsequence.
Anomalies occurring in time series can be a result of a change in the amplitude of data (e.g. a heavy rainfall in a week of a year), or it may be a change in the shape (e.g. occurring an arrhythmia within a set of normal heartbeats in ECG signals). Therefore, in this document we have reported anomalies of two types: anomalies in shape and anomalies in amplitude.
In this study, we propose a unified framework to detect both types of anomalies. For this purpose, after generating a set of subsequences of time series using a sliding window, a fuzzy C-Means (FCM) clustering has been employed to reveal the available structure within data. Then, a reconstruction criterion is considered to reconstruct the original subsequences from the determined cluster centers (prototypes) and partition matrix. For each subsequence, an anomaly score has been assigned based on the difference between the original subsequence and the reconstructed one. In the case of anomalies in amplitude, the original representation of time series along with the Euclidean distance

function is used in the clustering process, while for shape anomalies, first a representation of subsequences are considered to capture the shape information and then, the Euclidean distance in the new feature space has been employed.

2 Anomaly detection using a Fuzzy C-Means Clustering

Let us consider a time series $\mathbf{x} = x_1, x_2, \ldots, x_p$ of length p. We aim at finding a set of subsequences of $\mathbf{x}$ with length q, having highest amount of unexpected changes (in shape or amplitude) in terms of anomaly score. For this purpose, a sliding window with length q, moves thorough the time series and generates a set of subsequences. Consequently, there are n subsequences coming in the form

$$\begin{aligned} \mathbf{x}_1 &= x_{11}, x_{12}, \ldots, x_{1q} \\ \mathbf{x}_2 &= x_{21}, x_{22}, \ldots, x_{2q} \\ &\vdots \\ \mathbf{x}_n &= x_{n1}, x_{n2}, \ldots, x_{nq} \end{aligned} \tag{1}$$

Note that in each movement, the sliding window moves r time steps. As the result, the number of subsequences, n is

$$n = \frac{p-q}{r} + 1 \tag{2}$$

Considering a low value for r (e.g., $r = 1$) guarantees that no anomalous subsequences are missed, but processing a high amount of subsequences is time consuming. On the other hand, considering a high value for r (e.g., $r = q$) generates lower number of subsequences and processing time will be lower, but there is a risk of losing some anomalous subsequences. A trade off between accuracy and processing time can be considered. Selecting the value of r being proportional to the length of subsequences is a reasonable choice, i.e. selecting a higher value of r for longer subsequences and lower value of r for shorter subsequences. The length of sliding window, q, is another important parameter that can be selected based on the application purpose. However, one may consider different values for this parameter to find some appropriate results.
Fuzzy C-Means clustering proposed by Dunn and Bezdek is one of the commonly used clustering techniques. It describes n subsequences (or their representation) $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_3$ using c ($c \leq n$) prototypes $\mathbf{v}_1$, $\mathbf{v}_2$, ..., $\mathbf{v}_c$ and a fuzzy partition matrix U through the minimization of the following objective function:

$$\sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} \parallel \mathbf{v}_i - \mathbf{x}_k \parallel^2 \tag{3}$$

where m ($m > 1$) is a fuzzification coefficient, and $\parallel . \parallel$ denotes the Euclidean distance function. The defined objective function in (3) can be minimized by calculating the cluster centers using (4) and partition matrix using (5) in an iterative fashion.

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^{m} \mathbf{x}_k}{\sum_{k=1}^{n} u_{ik}^{m}} \tag{4}$$

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{\|\mathbf{v}_i - \mathbf{x}_k\|}{\|\mathbf{v}_i - \mathbf{x}_k\|} \right)^{2/m-1}} \tag{5}$$

Clustering subsequence $x_1, x_2, \ldots, x_n$ leads to the generation of a set of prototypes representing the normal structure of subsequences. Each normal subsequence in data set is similar to one or more

prototypes or it can be similar to a combination (in the form of weighted average) of prototypes. The more the subsequence is similar to the prototypes, the less anomalous it is. To evaluate how much a subsequence is similar to the revealed prototypes (or their combination) a reconstruction criterion has been considered in this paper. This criterion also was exploited for clustering spatio-temporal data in. Pedrycz and de Oliveira noted that FCM can be viewed as an encoding scheme of data and the original data points (here subsequences) can be decoded (reconstructed) using the estimated cluster centers and partition matrix. Assuming that $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_n$ are the reconstructed version of subsequences $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, by minimizing the following sum of distances:

$$F = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} \parallel \mathbf{v}_i - \hat{\mathbf{x}}_k \parallel^2 \tag{6}$$

one may arrive at:

$$\hat{\mathbf{x}}_k = \frac{\sum_{k=1}^{c} u_{ik}^{m} \mathbf{v}_i}{\sum_{i=1}^{c} u_{ik}^{m}} \tag{7}$$

After calculating the reconstructed version of each subsequence using (7), the reconstruction error in (8), that is a squared Euclidean distance between a subsequence and its reconstructed version is considered as the evaluation criterion to estimate how much a subsequence is similar to prototypes. In other words, for each subsequence the calculated reconstruction error using (8) is considered as its anomaly score.

$$E_k = \parallel \mathbf{x}_k - \hat{\mathbf{x}}_k \parallel^2 \tag{8}$$

When detecting anomalies in shape is of concern, the generated subsequences cannot be employed directly in clustering. The reason is that the generated subsequences are not synchronized and using the Euclidean distance function is not efficient as similarity measure. Although there are number of viable distance functions to measure the dissimilarity of asynchronous time series with respect to their shapes (e.g. dynamic time warping distance), one has to be aware of the challenges we may encounter for optimizing the FCM objective function in dealing with those distance functions. In this paper we confine ourselves to the Euclidean distance function. To compare subsequences based on their shape information, each subsequence is normalized to have a zero mean and a standard deviation equal to one. Then, each normalized subsequence is represented using a set of autocorrelation coefficients. Considering x_k as a subsequence with length q, its autocorrelation coefficient for lag s can be estimated using (9).

$$y_{k,s} = \frac{\sum_{t=s+1}^{q} (x_{k,t} - \bar{x}_k)(x_{k,t-s} - \bar{x}_k)}{\sum_{t=1}^{q} (x_{k,t} - \bar{x}_k)^2} \tag{9}$$

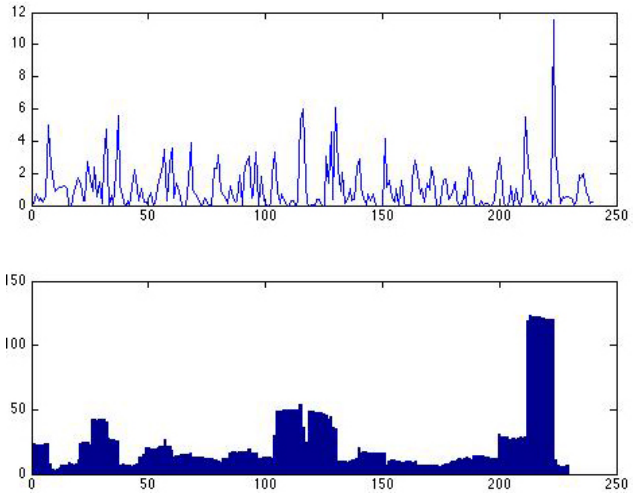

Figure 1: A monthly precipitation Time Series along with estimated Amplitude Anomaly Scores.

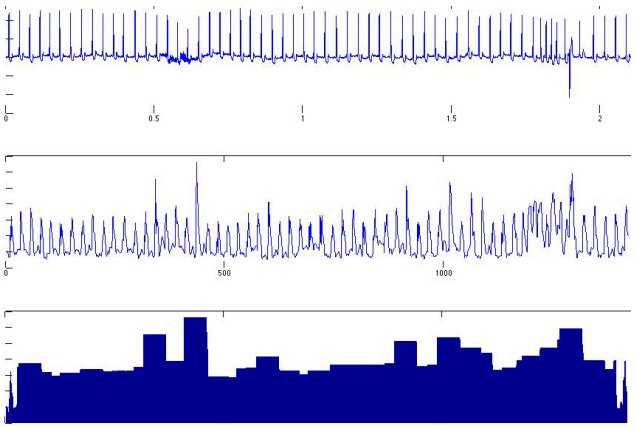

Figure 2: Detection of Sinus Arrhythmia and PVC in ECG Time Series.

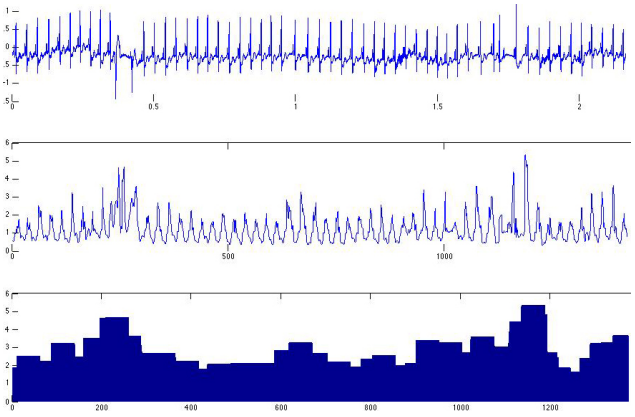

Figure 3: Application of Shape Anomaly Detection in ECG.

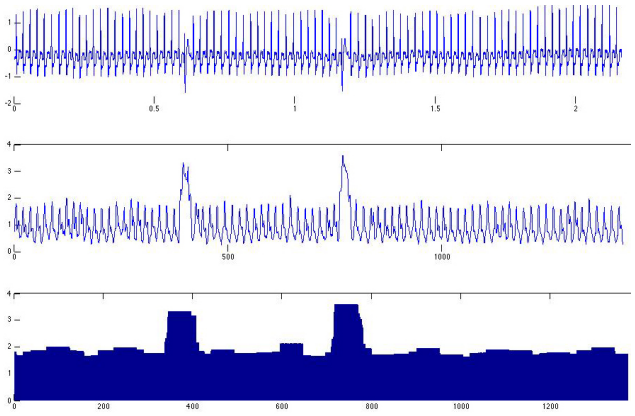

Figure 4: Another example.

References

[1] Chandola, V.; Banerjee, A.; Kumar, V. (2009), *"Anomaly detection: A survey"*

[2] Hodge, V. J.; Austin, J. (2004), *"A Survey of Outlier Detection Methodologies"*

[3] Dokas, Paul; Levent Ertoz, Vipin Kumar, Aleksandar Lazarevic, Jaideep Srivastava, Pang-Ning Tan (2002), *"Data mining for network intrusion detection"*

[4] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. *PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23):e215-e220*

[5] The United States Historical Climatology Network (USHCN), Available online at: http://cdiac.ornl.gov/epubs/ndp/ushcn/ushcn.html.

[6] V. Chandola, V. Mithal, V. Kumar, *"Comparative evaluation of anomaly detection techniques for sequence data," In 8th IEEE International Conference on Data Mining, 2008, pp. 743-748.*

[7] J.C. Dunn, "A fuzzy relative of the ISODATA process, and its use in detecting compact well-separated clusters,*J. Cyber. 3 (3), pp. 32–57, 1973.*

[8] Anomaly detection in time series data using a fuzzy c-means clustering,*Izakian H, Dept of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada, Pedrycz, W*