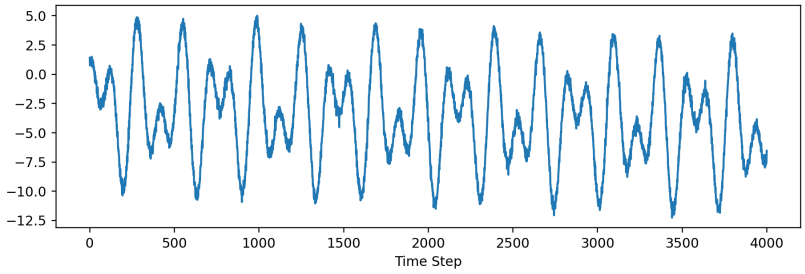
Anomaly Detection in Clinical Data

1 Introduction

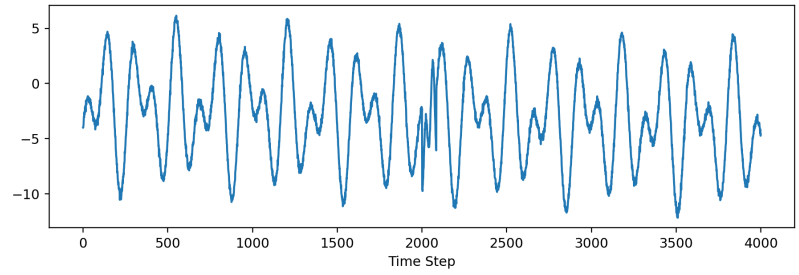
The existence of anomalies in data are sometimes an indicator of differing outcomes than normal. In this context, the existence of an anomaly are a predictor of poor after surgery outcomes for patients. An example of normal behavior is in figure 1.

Figure 1: Example of Normal Behavior



The data in this case follows a periodic structure that repeats for the time interval. In the case of an anomaly, there is a noticeable violation of the normal periodic structure.

Figure 2: Example of Anomalous Behavior

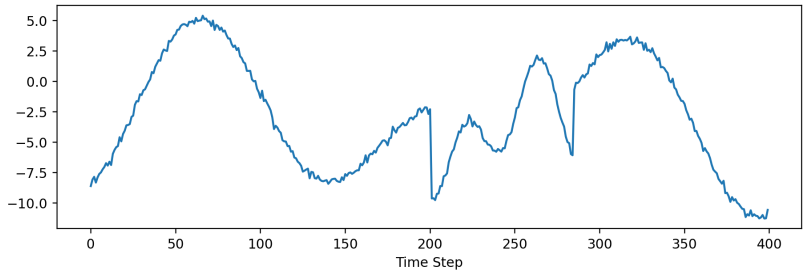


Around index 2000 in the figure above there is a sudden change in the data which goes against the normal periodic structure. The existence or lack of such anomalies can be used as predictors of patient outcomes.

2 Calculations

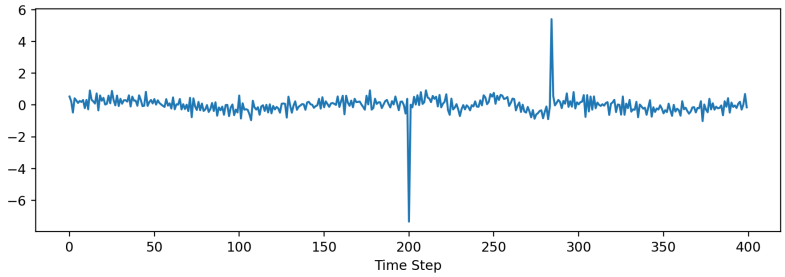
In normal behavior, the value of the time series will not change much between subsequent steps. When an anomaly occurs there is a sudden jump between one step and the step after it. The figure below shows a sudden jump at time 200.

Figure 3: Anomaly at time 200



In order to detect anomalies, we can calculate the differences between subsequent time steps of the time series. In formal terms, let f be a function from time to the real numbers where $f(t)$ is the value of the time series at time t. We can find sudden spikes in the data if the value $f(t) - f(t-1)$ is large at any time t.

Figure 4: Differences between sequential time steps of figure 3



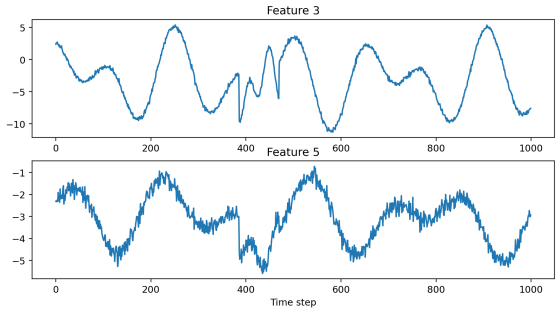
For any time series, we calculate the values $[f(t) - f(t-1)]$ for all t. In order to compare anomalies between different patients, we also need to normalize the differences. For a given time series, we calculate the standard deviations of the values $[f(t) - f(t-1)]$. The anomaly value of that time series is the greatest absolute value among the numbers $[f(t) - f(t-1)]$ divided by the standard deviation. We summarize this below:

1. Diff is the collection of numbers $\{[f(t) - f(t-1)]\}$
2. s = Standard deviation of Diff
3. Anomaly Value = max(abs(Diff)) divided by s

3 Conclusions

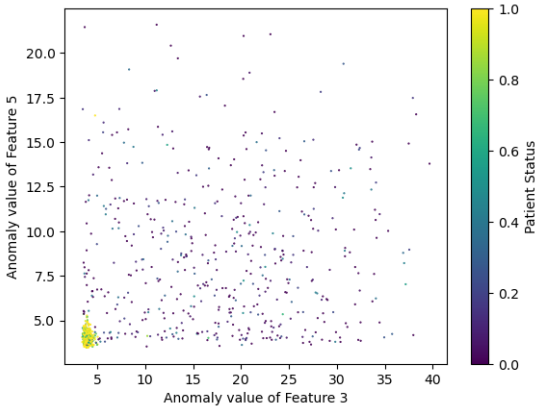
For the case of this study the anomaly values in features 3 and 5 are good predictors of patient statuses. Importantly, we observed the anomalies of features 3 and 5 tend to (but not always) occur together at the same time. An example is shown in figure 5.

Figure 5: Anomalies of Features 3 and 5



The graph below shows the relationship the anomaly values of features 3 and 5 and patient statuses. Note there are unhealthy patients without noticable anomalies.

Figure 6: Relation between Features 3 and 5 and Patient Status



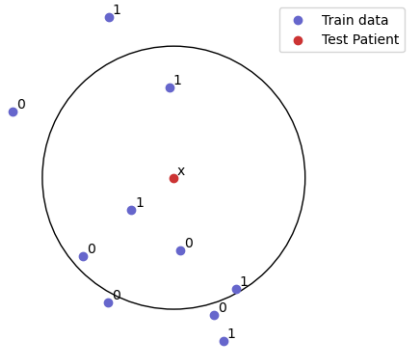
A K-nearest neighbor classifier was ran on the patients. Healthy patients are considered to have status more than 0.5 and unhealthy patients have status less than 0.5. The classifier seeks to categorize patients into these two categories. An explanation of K-nearest neighbors can be found below. The overall accuracy on a sample test set of patients was 95.0%. There are unhealthy patients which did not exhibit anomalies in features 3 and 5 which were labeled healthy by the

classifier. Of the unhealthy patients, 9.5% were predicted as healthy. Of the healthy patients, only 1.0% were predicted as unhealthy.

4 Explanation of K-Nearest Classifier

The algorithm takes a training sample of patients. Given a test patient, say patient x, it computes the k-nearest points in the training sample to x. The algorithm then computes what category the majority of the k-nearest patients are in, and predicts this as the category for x. An example is shown in figure 7.

Figure 7: K-Nearest Neighbors



In this example we have $k = 5$, so the algorithm will find the 5 closest points to x. The training points are in blue and patient x is in red. Among the 5 closest points to red, three are labeled 1 and two are labeled 0. So, the classifier will predict x is labeled 1.

Due to the simplistic nature of the classifier, unhealthy patients with anomalies can be mislabeled. In our testing sample there was one unhealthy patient with a anomalies which was labeled healthy because the closest points were all healthy.