CHANGE DETECTION IN ONLINE MULTIVARIATE SENSOR DATA

Ph.D Registration Report

By

BIMAL KUMAR SAHOO

17IM30032

Under the supervision of

MAMATA JENAMANI

Department of Industrial & Systems Engineering
Indian Institute of Technology Kharagpur
Kharagpur-721302, India
2020

ABSTRACT

This report discusses a non-parametric method to detect changes in live multivariate data from sensor nodes. The method used is a semi parametric log likelihood method with different clustering algorithms. The SPLL method is compared using different clustering techniques. The SPLL method is modified to use for online change detection by using a sliding window method or a depth model using Mahalanobis distance. This report also compares some parametric methods (like Hotelling's t2 test) and some non-parametric methods (like KL divergence, SPLL method) using proper metrics and proposes the method working best for our dataset satisfying all the device restrictions. Some protective mechanisms are also used to check for false alarms.

INTRODUCTION

Change detection algorithms are used to detect changes in the probability distribution of stochastic (or time series) data. Change detection is widely used in fields where we have a continuous stream of time series data or an offline data store to record for changes occuring in previous data.

There are two categories of change detection based on source of data:
Offline data: The entire dataset is available before applying the model.
Online data: The change detection algorithm runs continuously as the data received is continuous.

Categories based on presumptions:
Parametric: Dataset is assumed to follow some predefined distribution like Gaussian distribution or Beta distribution etc. eg. Hotelling's t2 test assumes the data to follow a Gaussian distribution.
Non- parametric: No assumptions are made for the data.
It is usually observed that for *online change detection, non-parametric methods* are effective and produce better results on the benchmark datasets.

Metrics:
AUC-ROC : Receiver Operating Characteristic (ROC) curve is a performance measure for multiclass classification problems. The y - axis defines the True Positive Rate and the x - axis defines the False Positive Rate. The Area Under Curve (AUC) is the area under the ROC curve. If the Area under the curve is more than 0.5, the model is more likely to distinguish positive class values from the negative class values

OBJECTIVES

- The change detection algorithm is to be used in an IoT device for which our algorithm needs to be fast as well as efficient.

- The algorithm used should be a non-parametric algorithm so that it is independent of any presumptions of the data following gaussian or any other distribution.

- The algorithm used should be applicable for online multivariate data. The data expected to be received from the sensor nodes are live multivariate data. So our algorithm should not consider the availability of the complete dataset beforehand for change detection.

- The time taken for the change detection algorithm to run and be ready to take the input of the next data point should be less than the time gap between the consecutive sensory readings.

- The algorithm is to be implemented in an IoT device which might use algorithms like Arduino or Raspberry Pi, so our algorithm needs to be computationally less expensive.

- The algorithm should be accurate. Since our change detection alert is going to be sent through a blockchain. So reducing the false alerts is a necessary factor for our algorithm.

METHODOLOGY

Semi-parametric log likelihood method is a non-parametric method for change detection in offline data. It is implemented by following way:

- The dataset is first clustered into ‘K’ clusters using gaussian mixture models. But since our objective is to provide a fast change detection algorithm, we are using the K- means algorithm for clustering.

- Mean and variance of each cluster is calculated. Then the covariance matrix is calculated taking weights of each cluster proportional to the number of elements in the corresponding cluster.

- Now, the partial inverse of the covariance matrix is calculated using the pinv function from numpy library.

- This matrix is used to calculate the squared Mahalanobis distance of a datapoint from the nearest cluster center.

- The mean of the squared Mahalanobis distance of all data points is taken and a chi2 - test is conducted for n degree of freedom.

- Now the result for different confidence intervals is plotted in an AUC-ROC curve and conclusions are drawn.

WORK DONE SO FAR

- A semi parametric log likelihood method is selected among multiple methods to use as our change detection algorithm.

- A python code is written on the algorithm and the performance of which is verified using multiple datasets and AUC-ROC curves are plotted with the result.

- It is observed that the algorithm performs good with normalised as well as non-normalized dataset. The change detection algorithm performs well for small as well as big dataset (it is nearly independent from the number of clusters made). This non-parametric method with small tweaking is expected to give impressive results for online multivariate dataset.

- Some methods for advanced change detection have been searched and are yet to be applied. These methods include an ensemble method of the 3 multivariate log-likelihood methods with both nonparametric (KL divergence) and parametric (Hotelling's t2 test) methods along with the SPLL method to generate an effective model for change detection of data.

- Some methods for online change detection have also been searched which include change detection using sliding window model and energy statistics.

WORK TO BE DONE

- Noise reduction: We can use different filtering techniques to remove simple noise values which may be considered as outliers by our algorithm. Some filtering techniques may be Gaussian filtering or moving average or mean filtering etc.

- We can also have an ensemble method for change detection of some univariate methods for some subspaces and use aggregation techniques like voting methods to improvise our algorithm. We can also use an ensemble of some multivariate change detection algorithms like KL divergence, Hotelling's t2-test and SPLL method.

- We need to have this method suited for online change detection using a sliding window method or energy statistics method to effectively detect changes in the constantly recorded data.

References

[1] Yang Xiao-fei, Wu Xiao-bei, and Huang Jin-an, "*TAGPP: a tiny aggregation algorithm with preprocessing in local cluster"*, 2009 International Conference on Networks Security.

[2] Ludmila I. Kuncheva, *"Change Detection in Streaming Multivariate Data Using Likelihood Detectors",* 2013 IEEE.

[3] Eric L. Bullock , Curtis E. Woodcock, Christopher E. Holden, *"Improved change monitoring using an ensemble of time series algorithms",* 2019 Elsevier Inc.

[4] Lingzhe Guo & Reza Modarres, *"Two multivariate online change detection models",* 2020 Journal of Applied Statistics