CHANGE DETECTION IN ONLINE MULTIVARIATE SENSOR DATA

Bimal K. Sahoo

IIT KHARAGPUR

# Abstract

Online change point detection in multivariate data includes methods monitoring changes in the distribution of data streams. The data is collected from sensors collecting information about the mangos like temperature, color changes, humidity etc. This report discusses some effective parametric and non-parametric change point detection algorithms like Semi parametric Log Likelihood (SPLL) estimation, some time series models for change detection like CUMSUM, CCDC, Chow’s Test, and energy statistic using sliding window algorithm and Mahalanobis depth algorithm to train the threshold with desired protective ability against false alarms. The algorithms are state of the art algorithms and are chosen on the basis of their efficiency and low computational power requirements/

# 

# 

# INTRODUCTION

**Data Collection:** The data received is from wireless sensor nodes (WSN). They contribute to collecting data in various environments but with limited battery power. The data are in streams collected from the sensor nodes at equal intervals. An IoT device is installed to run the proposed method and detect for any change in the data streams and notify the users through a blockchain mechanism. The data collected are multivariate, non-normalised, continuous and classification data.

**Computational Limitations:** Since the IoT devices are installed to detect any change in the data streams, the algorithm needs to be computationally less expensive and effective enough to give a good accuracy and reduce false alarms. For such conditions, some parametric as well as non-parametric algorithms are implemented along with some time series analysis on the algorithms are done so as to improve the efficiency of the algorithm.

Till now in this report, two state of the art algorithms are implemented. Yang Xiao-fei, Wu Xiao-bei, and Huang Jin-an [1] propose a method for change detection in wireless sensor nodes using fuzzy clustering and Grubbs Criteria, focussing mainly on scaling the desired data and applying the Generalized ESD test. Ludmila I. Kuncheva [2] proposed comparing three likelihood change detector models K-L Divergence, Hotelling's t2-test, and the proposed Semi-parametric Log Likelihood Estimator (SPLL) focussing on Likelihood detection algorithms.

Next, there are some papers focussing on online data and time series algorithms for change detection. Eric L. Bullock , Curtis E. Woodcock, Christopher E. Holden [3] discusses an ensemble method of time series algorithms for change detection in land covers. Lingzhe Guo and Reza Modarres [4] discusses online change detection using sliding window algorithm and Mahalanobis depth algorithm along with some improvements.

The change detection algorithm to be used should be robust and accurate. The algorithm is then implemented in an IoT device. The device should be able to run the algorithm and be ready for the next input within the time interval of 1 or 2 sec. This process includes storing the data in a ledger of a blockchain transaction database. The users then will be able to receive the change detection alert.

# ALGORITHMS

**Parametric:**

Paper 1: TAGPP: a tiny aggregation algorithm with preprocessing in local cluster

Data Alignment - It generally includes coordinates transform and time adjustment. The purpose is to align the data into a common reference point for using the algorithm

Clustering - The clustering method used is Fuzzy clustering. It returns the fuzzy weight using normal function as the membership function.

Grubbs criteria - This criteria is then used to detect the outliers. Grubbs criteria uses t statistics for detecting change in normalized data points.

Paper 2: Change Detection in Streaming Multivariate Data Using Likelihood Detectors

This paper discusses about 3 methods and includes a detailed ROC-AUC analysis. The 3 methods discussed in this paper are KL divergence, Hotelling's t2 test, the proposed method (SPLL). The paper infers the advantages of using the proposed Semi-parametric Log Likelihood method. In this proposed method, the data is clustered and the probability function of each cluster is determined, then the mahalanobis distance is calculated for each data point (here, each row). Now the p value is determined by a chi2 test upon the mahalanobis distance found and number of features (degrees of freedom) and checked with a threshold to determine if there is any change. This method is efficient for both normalised as well as non - normalized data.

**Non-parametric:**

Paper 1: Improved change monitoring using an ensemble of time series algorithms

This paper proposes an ensemble of three time series change detection models, Continuous Change Detection and Classification (CCDC) test, Cumulative Sum of

Residuals (CUSUM) algorithms for break detection and the Chow Test (Chow, 1960) for removing false positives or breaks in time series not representing land change. The ensemble methods proposes a sequential ensemble method of CCDC, CUSUM and Chow’s Test. The first algorithm, CCDC, is used to identify breaks in the time series and fit harmonic regression models for each spectral band and for every discrete time segment between model breaks. The data from the discrete time segments are

then used as inputs to the second change detection algorithm, CUSUM, which is used to identify breaks that were missed by CCDC. All break points in the time series are then tested with the Chow Test to identify false, or unnecessary breaks.

Paper 2: Two multivariate online change detection models

This paper discusses two non-parametric methods for detecting change in online data using energy statistics and Mahalanobis depth. The sliding-window model for online change detection performs homogeneity tests over two different time windows. H parameter is updated by considering the data stream of train input values and then taking the (1- alpha) quartile. Then the energy states are calculated by taking the euclidean mean of the energy states and checking for change using the f-statistics. The next method proposed is the Depth method which uses clustering technique to cluster each dataset and calculates the difference of the clusters using probability divergence algorithms then the Mahalanobis distance is calculated and checked with L statistics for crossing the threshold (raise alert) or not.

1. **CODE IMPLEMENTATION:**

Two of the above methods have been implemented in different library datasets in python language. The fuzzy clustering and Grubbs criteria for outlier detection detects outliers but is not that effective in mean changes. This criteria works fine if the change expected is any change in the variance or for any big changes. This algorithm fails to detect small changes in data mean. The accuracy of the model is not impressive. Then the method proposed by [2] was implemented on datasets like iris, satimage, wbc, etc. This method was robust, efficient in detecting changes in mean as well as variance and variance along with some trend. The algorithm proposes to use gaussian clustering models for cluster probability function but using K- prototype algorithm requires less computational expense and gives good results. The probability functions are determined using the mean of each cluster and an weighted average of the clusters based on the number of elements. Then a Semi-parametric Log Likelihood algorithm is used which considers the upper bound of the divergence. Then the mean Mahalanobis distance is calculated which is then tested with t statistics to get the p-value and if the p-value is less than the threshold, a change is raised. This algorithm works well for offline data.

# 

# RESULTS AND FUTURE WORK

The SPLL method gives impressive results for normalised as well non- normalised data. The drawback of this method remains to the limitation of this algorithm to offline multivariate data change detection. Further this method can be improved if a time analysis factor is added to the change detection model. Thus sliding window technique could be added or a protective model for training the threshold could be determined. The resulting method would be robust, accurate, easily implementable and computationally less expensive for change detection in online multivariate streaming 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

[5]

[6]

[7]

[8]