Designing a framework for implementing CUSUM algorithms for Climate change detection for relative diseases

Ms. Abhiruchi Mishra[1]
[1]Student, Department of Computer Science and Engineering,
[1]Shri Sant Gajanan Maharaj College Of Engineering,

Abstract— Big data have long predicted widespread irresistible infections as large-scale reactions to seasonal climate change, polarising discussion, and a particularly dreadful human virus for which financial motives and control techniques might restrict the disclosure of atmospheric intervention changes. The majority of the time, seasonal climate change illnesses are no longer common because of modifications to rural land practises, isolated incidents, changes in human conduct, and management of vectors. Most vector-borne illnesses only seldom show symptoms, and they are climate sensitive. The Cumulative Sum (CUSUM) algorithms are used in Big Data Analytics to identify climate change. The recommended CUSUM algorithm produces better results for the climate change algorithm.
Key terms--Climate change, Big data, Cusum, Data analytics.

I Introduction

To predict climatic changes and disease from the environment, large-scale climate data have been used. It integrates several data processing prototypes, including graph-based technique, closest neighbour algorithm, principal component analysis as an unsupervised model, binary segmentation models, and lastly a spatial autocorrelation model, to boost scalability and feasibility. The aforementioned strategies are thoroughly analysed in the proposed system to offer a novel paradigm for handling complex climatic data. The analysis of the solution to managing complex data has taken into account a number of aspects, such as creating a trajectory for data flow, projecting the load balancing model for data management, cooperating to handle conjunctions, and cooperating to anticipate uncertainty. It is very difficult to predict climate change and its potential effects. [1]

Numerous studies have been conducted on big data analytics-based climate modelling. Data on the climate and seasonal changes are gathered, shared, and monitored by the US Environmental Security Agency. These have been applied to evaluate the results of ongoing healthcare reforms. Complex big data solutions must be used to process enormous data sources in the healthcare sector. Recent state-of-the-art study concentrates on how big data impacts disease diagnosis. The significance of the massive data stored in big data and cloud computing has also increased noticeably. Advanced big data solutions will help in the diagnosis and prognosis of diseases, as well as in the reduction of treatment costs and improvement of healthcare quality. One of the algorithms used for one of the approaches to predict climate change is the CUSUM Algorithm. Big data analytics is becoming more and more crucial in a variety of industries, including healthcare, social networking, and weather forecasting. The state-of-the-art sensor can be used to collect climate data, which depicts seasonal changes. Weather information can be used to research seasonal diseases and follow seasonal fluctuations, while meteorological data are utilised to forecast the weather.

Analysing whether or not the average daily climate compares favourably to the previous climatic time period is a crucial step in comprehending seasonal changes. Weather information acquired from many sources has been used to identify seasonal climatic variations. Unstructured data patterns were mostly used to evolve climate testing. The key complementary situations that raise mathematical problems in a multidisciplinary context include epidemiology, health, biodiversity, natural resources, and seasonal climate change. In order to execute complex workflow pipelines and a strategy for managing heterogeneous and huge datasets, some stated scenarios demand access to the provided framework.[2]

Brownian motion drift changes are the main subject of the first exact optimality result in sequential change detection. Using a Bayesian approach, the change time is treated as an independent random variable with an exponential distribution. The alternative detection approach exhibits significant similarities to the greatest Bayesian test yet developed, while being developed using a minimax strategy. Unfortunately, it has been demonstrated that the Shiryaev-Roberts-Pollak (SRP) test, a sequential detector, is wrong while having a very strong asymptotic optimality property. Also worth noting is how easy and quick it is to compute the associated test statistic for the CUSUM and SRP tests in the classical version of the problem that we have so far looked at.

1.1 Objective

- To implement the CUSUM algorithm and Logistic Regression.
- To analyze the input data for Handling Missing data, Label Encoding and dropping unnecessary columns

1.2 Problem statement

Online change point detection is a fundamental problem in statistics and signal processing which finds applications in a plethora of practical problems in diverse fields. The most common version of the problem consists of a sequence of observations sampled independently. There is also a change point such that the underlying distribution changes from one distribution to an alternative. This problem is of major importance in many applications, such as seismic signal processing, industrial quality control, dynamical systems monitoring, structural health control, event detection, anomaly detection, detection of attacks, etc. The goal of online change point detection is to detect the occurrence of the change in statistical behavior with a minimal delay while controlling the false alarm rate. The suitable tradeoff between detection delay and false alarm rate, as in all detection problems, is of essential importance for the proper mathematical formulation of the problem. Online change detection involves monitoring a stream of data for changes in the statistical properties of incoming observations. A good change detector will detect any changes shortly after they occur, while raising few false alarms. Although there are algorithms with confirmed optimality properties for this task, they rely on the exact specifications of the relevant probability distributions and this limits their practicality.

II Literature Survey

Nandhini V et. al. states that Big data analytics has been applied to predict climate changes and climate borne disease from large scale climate data. It incorporates various data processing prototype such as linear and non-linear data processing algorithm, binary segmentation models, nearest neighborhood algorithm, principal component analysis as an unsupervised model, Graph-based technique and finally spatial autocorrelation model to increase the scalability and feasibility. In this paper, a detailed study is carried out on the above-said methods to devise a novel paradigm to handle complex climate data. The solution to managing complex data has been analyzed on a various aspect such as constructing trajectory for data flowing, projecting the load balancing model for data managing, collaborative model for conjunctions handling and joint model for uncertainty prediction. Big data analytics is improving a lot of importance in healthcare, social networking, climate modeling and so on. Climate data reflect seasonal changes, and it could be collected using the advanced sensor. Weather prediction is made using metrological data, and weather data also useful to analyze seasonal diseases, and reflects seasonal changes. This helps in improving public health. Usually, the climate data is raw, and it is unstructured format from this data meaningful information extracted using analytical techniques. Recently, robotics algorithms are also found useful for climate data analyses. The collection of a vast amount of weather data from overall climate system. Climate "normal" are used to compare the climate condition of present and past. The average climate parameter ("maximum or minimum temperature") during a period is used to calculate normal climate. [1]

Liyan Xie et. al. developed an alternative approach for solving the problem of interest which we call Window-Limited CUSUM (WLCUSUM). It consists in adopting a window-based estimate of the unknown post-change parameters and, unlike the existing window-limited GLR, we use the estimate in the updating formula of the classical CUSUM statistic. This updating mechanism is far more efficient than its window-limited GLR counterpart and by proper selection of the window size we can also guarantee asymptotic optimality. We would like to emphasize that the problem we consider is not joint detection and estimation, where the two tasks are regarded as equally important. Here, we are primarily interested in detection, with estimation being an auxiliary action that contributes towards our detection goal. For this reason we only require the estimator to be consistent without insisting on any explicit form. Compared with existing CUSUM-like procedures employing estimates of post-change parameters, the proposed WLCUSUM method applies to a far wider range of parametric distributions and not only the exponential family which is mostly the case with the available approaches.

Thomas et. al. introduced the Kernel Cumulative Sum Algorithm (KCUSUM), a new approach for online change detection. Unlike the CUSUM algorithm, this approach does not require knowledge of the probability density ratio for its implementation. Instead, it uses incoming observations and samples from the pre-change distribution. The result is that the same algorithm works for detecting many types of changes. Our theoretical analysis establishes the algorithm's ability to detect changes, and shows a relation between the delay and the MMD distance of the two distributions. These bounds should also be useful in the analysis of other non-parametric change detectors. Finally, we would like to suggest two avenues for future work. First, there are likely variants of KCUSUM that leverage more complex that may lead to improved detection performance. Secondly, the CUSUM has been investigated for detecting changes in scenarios with more complex dependencies among observations and it may also be possible to extend the kernel methods developed in this paper to detect changes in these cases.

In this study the change detection problem in a general HMM when the change parameters are unknown and the change can be slow or drastic. Drastic changes can be detected easily using the increase in tracking error or the negative log of observation likelihood (OL). But slow changes usually get missed. We have proposed in past work a statistic called ELL which works for slow change detection. Now single time estimates of any statistic can be noisy. Hence we propose a modification of the Cumulative Sum (CUSUM) algorithm which can be applied to ELL and OL and thus improves both slow and drastic change detection performance. Change detection is required in many practical problems arising in quality control, flight control, and fault detection and in surveillance problems like abnormal activity detection. In most cases, the

underlying system in its normal state can be modelled as a parametric stochastic model which is usually nonlinear. The observations are usually noisy (making the system partially observed). Such a system forms a "general HMM" (also referred to as a "partially observed nonlinear dynamical model" or a "stochastic state space model" in different contexts). The General Hidden Markov Model library (GHMM) is a freely available C library implementing efficient data structures and algorithms for basic and extended HMMs with discrete and continuous emissions. It comes with Python wrappers which provide a much nicer interface and added functionality.

Namrata et. al. Given its associated burden of disease, climate change in South Africa could be reframed as predominately a health issue, one necessitating an urgent health-sector response. The growing impact of climate change has major implications for South Africa, especially for the numerous vulnerable groups in the country. We systematically reviewed the literature by searching PubMed and Web of Science. Of the 820 papers screened, 34 were identified that assessed the impacts of climate change on health in the country. Most papers covered effects of heat on health or on infectious diseases (20/34; 59%). We found that extreme weather events are the most noticeable effects to date, especially droughts in the Western Cape, but rises in vector-borne diseases are gaining prominence. Climate aberration is also linked in myriad ways with outbreaks of food and waterborne diseases, and possibly with the recent Listeria epidemic.

III Proposed System (Methodology and Techniques)

Python Packages:

- **ocpdet**

OCPDet is a Python package for online changepoint detection, implementing state-of-the-art algorithms and a novel approach.

- **Itertools**

This module implements a number of iterator building blocks inspired by constructs from APL, Haskell, and SML. Each has been recast in a form suitable for Python. The module standardizes a core set of fast, memory efficient tools that are useful by themselves or in combination. Together, they form an "iterator algebra" making it possible to construct specialized tools succinctly and efficiently in pure Python.

- **Numpy**

NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices. Using NumPy, mathematical and logical operations on arrays can be performed. This tutorial explains the basics of NumPy such as its architecture and environment. It also discusses the various array functions, types of indexing, etc. An introduction to Matplotlib is also provided. All this is explained with the help of examples for better understanding.

- **Pandas**

Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data. Pandas allow us to analyze big data and make conclusions based on statistical theories. Pandas can clean messy data sets, and make them readable and relevant.

Methodology

- **Input data:** Here, we can collect the climate change south Africa Dataset from dataset repository. The dataset is in the format '.csv' or '.xlxs'.
- **Pre-processing**: Here, the collected input data is carried out to pre-processing step,
- Handling Missing data
- Label Encoding
- Drop unnecessary columns
- **Data Splitting**: In this step, we can split the pre-processed data into test and train for decision making,
- Test data is used for prediction (30%)
- Train data is used for training (70%)
- **Classification**: Here, we can implement the two different regression algorithms such as
- Cusum Algorithm
- Linear Regression (LR)
- **Forecasting:** Here, we can forecast or analyse the **climate change** (temperature) based on dataset attributes by using regression algorithms.
- **Performance Estimation**: Here, we can estimate some performance metrics such as
- Mean Absolute Error
- Mean Squared Error
- Root Mean Squared Error
- Forecasting Graph

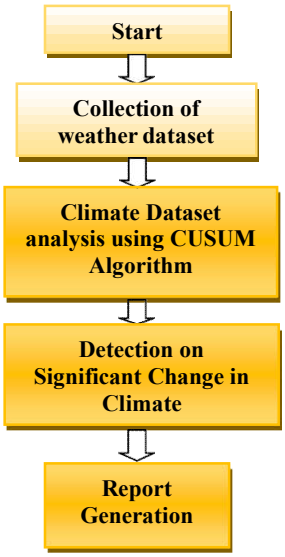

Fig. 1 Flow Chart for Detection of Climate Change

CUSUM Algorithm

A cumulative sum chart (CUSUM), a type of control chart, is used to detect the divergence of the individual values or subgroup mean from the changed target value or to monitor the variation from the target value. CUSUM charts can be used in place of Shewhart control charts. The CUSUM chart has the advantage of being more sensitive to the little shift of the process mean when compared to the Shewhart charts (Individuals I-MR or Xbar charts).The cumulative sum chart and the exponentially weighted moving average (EWMA) chart both display the process' mean, but they differ significantly from Xbar charts in that they also display the means of the previous values at each point. These charts are regarded as accurate approximations when the correct standard deviation is included. Instead of calculating the subgroup means independently, the data from the current and previous samples is displayed using a CUSUM chart. As a result, the CUSUM chart is consistently better than the Xbar plots at identifying minute changes in process mean.

Logistic Regression

Logistic regression determines the probability that an event, such as voting or not voting, will occur based on a collection of independent variables. Since the outcome is a probability, the range of the dependent variable is 0 to 1. The odds, or likelihood of success divided by probability of failure, are transformed using the logit formula in logistic regression.

In the context of artificial intelligence, the supervised machine learning model family includes logistic regression. Additionally, it is recognised as a discriminatory model, which means that it tries to differentiate across classes. It cannot, contrary to a generative algorithm like naive bayes, generate information of the class that it is trying to predict (for example, a picture of a cat). Essentially, logistic regression does supervised categorisation. In a classification problem, the goal variable (or output), y, can only take discrete values for a particular set of features (or inputs), X. Contrary to popular belief, a logistic regression is a regression model. The model builds a regression model to predict the probability that a particular data entry will fall into the group denoted by the number "1". Similar to how linear regression presumes that the data follows a linear distribution, logistic regression models the data using the sigmoid function. Logistic regression only becomes a classification strategy when a decision threshold is added. The logistic regression's key element, the threshold value, is determined by the classification problem itself. The threshold value selection is significantly influenced by the precision and recall levels. Precision and recall ought to be equal in a perfect environment, but this is rarely the case. The setup of the reference levels must therefore be carefully considered. If there is no obvious rule generated from the data itself or by prior knowledge about the variable values, it is advised to set a reference level with a minimum sample size to provide appropriate statistical power. Another approach to aid understanding is to select categories that have a similar connection to the pertinent event. If you believe that those who are older and those who are undergoing innovative treatments have lower mortality rates, use these two groups as a point of comparison. Alternately, you might use conventional medical care and younger patients. It is doable and not inaccurate to choose older individuals and receive regular care, but doing so will make it more challenging to comprehend the results.

IV Performance Analysis

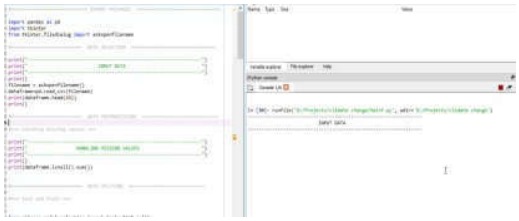

Fig. 1 Editor Window

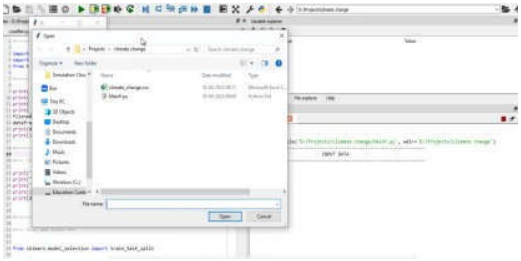

Fig 2 Selection of dataset

```
----------------------------------------------------------------
                      DATA SPLITTING
----------------------------------------------------------------

Total No. of data's in input              : 308

Total No. of data's  in training data     : 246

Total No. of  data's in testing data      : 62

----------------------------------------------------------------
                      LINEAR REGRESSION
----------------------------------------------------------------

1. Mean Absolute Error ( MAE)     : 0.07393626391432558

2. Mean Squared Error (MSE)       : 0.008097773547311104

3. Root Mean Squared Error (RMSE)  : 0.08998762996829678
```

Fig 3 Data splitting and linear regression

```
LONG SHORT TERM MEMORY ---) LSTM

Train on 277 samples, validate on 31 samples
Epoch 1/5
277/277 [==============================] - 9s 33ms/step - loss: 0.1208 - mae: 0.2571 - val_loss:
0.0890 - val_mae: 0.2821
Epoch 2/5
277/277 [==============================] - 1s 2ms/step - loss: 0.0514 - mae: 0.1811 - val_loss:
0.0106 - val_mae: 0.0812
Epoch 3/5
277/277 [==============================] - 1s 2ms/step - loss: 0.0550 - mae: 0.1882 - val_loss:
```

Fig 4 Application of LSTM

```
                    PREDICTION
------------------------------------------------

==========================================

[0] The Climate Change= -0.007342379340855132
==========================================

[1] The Climate Change= 0.42810863032528346
==========================================

[2] The Climate Change= 0.3143244543422554
==========================================

[3] The Climate Change= 0.18759464295149542
==========================================

[4] The Climate Change= 0.33909758779421627
```

Fig 5 Climate change prediction

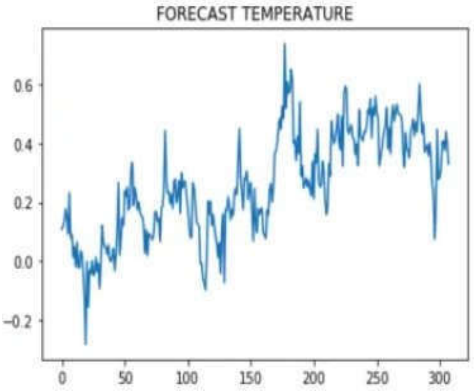

Fig 6 Climate forecasting for temperature

V Conclusion

From the perspective of practical and precise calculation, CUSUM is a costly step for detecting climate changes. The suggested approach looks for changes brought on by climate change using the CUSUM algorithm. Utilising a climatic data collection, the performance of the recommended solution is confirmed. A data cleaning technique is used to decrease the amount of data that needs to be collected, increase the accuracy of the current changes in climatic conditions, and save time and money. The moderate and radical changes in the mean assessment of a quantity of intrigue are identified using seasonal fluctuations and aggregate approaches. This method is used in a variety of activities; it shows on the screen changes in generational conditions, disease desires, fish numbers, deforestation, and investigations into wrongdoing. Finally, we would like to suggest two potential lines of inquiry for further investigation. First, it's likely that there are KCUSUM variants that use more complex non-i.i.d. statistics and could improve detection performance. Second, the CUSUM has been studied in contexts with more complex relationships between observations for change detection.

REFERENCES

[1] V.Nandhini, Dr.M.S.Geetha Devasena, "Predictive Analytics for Climate Change Detection and Disease Diagnosis", 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)

[2] Liyan Xie, George V. Moustakides And Yao Xie ," Window-Limited CUSUM for Sequential Change Detection", arXiv:2206.06777v1 [math.ST] 14 Jun 2022

[3] Thomas Flynn and Shinjae Yoo," Change Detection with the Kernel Cumulative Sum Algorithm", 2019 IEEE 58th Conference on Decision and Control (CDC) Palais des Congrès et des Expositions Nice Acropolis Nice, France, December 11-13, 2019.

[4] Namrata Vaswani," The Modified Cusum Algorithm For Slowand Drastic Change Detection In General Hmms With Unknown Change Parameters", 0-7803-8874 7/05/ ©2005 IEEE ICASSP 2005

[5] Matthew F. Chersich Caradee Y. Wright, Francois Venter, Helen Rees, Fiona Scorgie, and Barend Erasmus," Impacts of Climate Change on Health and Wellbeing in South Africa, Int. J. Environ. Res. Public Health 2018, 15, 1884; doi:10.3390/ijerph15091884 www.mdpi.com/journal/ijerph

[6] C. Yu, R. Zhang, Y. Huang, and H. Xiong, "High-dimensional knn joins with incremental updates," GeoInformatica, 2010.

[7] W.Welcome,"http://www.wmo.int/pages/prog/wcp/wcdmp/ index _ en.php, "2017.

[8] Lee JG, Kang M, "Geospatial big data: challenges and opportunities," Big Data Res, pp.74–81, 2015.

[9] Native S, Mazzetti P, Santoro M, Papeschi F, Craglia M, Ochiai O, "Big data challenges in building the global earth observation system of systems," Environ Model Softw, vol.68,pp.1–26,2015.

[10] EnviroAtlas | US Environmental Protection Agency,http://enviroatlas.epa.gov/ enviroatlas, 2017.

[11] Pickard BR, Baynes J, Mehaffey M, Neale AC, "Translating big data into big climate ideas, "Solutions, vol.6,pp.64–73,2015.

[12] Lopez D, Gunasekaran M, Murugan BS, Kaur H, Abbas KM, "Spatial big data analytics of influenza epidemic in Vellore, IEEE international conference on big data (big data), "IEEE, vol.27, pp. 19–24, 2014.

[13] Lopez D, Gunasekaran M, "Assessment of vaccination strategies using fuzzy multicriteria decision making. In: Proceedings of the fifth international conference on fuzzy and neurocomputing (FANCCO-2015)," Springer International Publishing, pp. 195– 208, 2015.

[14] Lopez D, Sekaran G, "Climate change and disease dynamics-a big data perspective. Int J Infect Dis, "pp.23–4, 2016.

[15] Lopez D, Manogaran G, " Big data architecture for climate change and disease dynamics, "The human element of big data : issues, analytics, and performance. USA: CRC Press; 2016.

[16] Manogaran G, Thota C, Kumar MV, " MetaCloudDataStorage architecture for big data security in cloud computing," Procedia Comput Sci, pp.128–33, 2016.

[17] Manogaran G, Thota C, Lopez D, Vijayakumar V, Abbas KM, Sundarsekar R, "Big data knowledge system in healthcare. In: Internet of things and big data technologies for next-generation healthcare," Springer International Publishing, pp. 133–57, 2016.

[18] Manogaran G, Lopez D. Disease surveillance system for big climate data processing and dengue transmission. Int J Ambient Comput Intell, pp.88–105, 2017.

[19] Thota C, Manogaran G, Lopez D, Vijayakumar V, "Big data security framework for distributed cloud data centers.In: Cybersecurity breaches and issues surrounding online threat protection," USA: IGI Global, pp. 288–310, 2017.

[20] Manogaran G, Thota C, Lopez D, Vijayakumar V, Abbas KM, Sundarsekar R, "Big data knowledge system in healthcare. In: Internet of things and big data technologies for next-generation healthcare," Springer International Publishing, pp. 133–57, 2017.

[21] Murdoch TB, Detsky AS, "The inevitable application of big data to healthcare, "vol.309(13, pp.1351–2, 2013.

[22] Haines, A.; Kovats, R.S.; Campbell-Lendrum, D.; Corvalan, C. Climate change and human health: Impacts, vulnerability and public health. Public Health 2006, 120, 585–596. [CrossRef] [PubMed]

[23] Cohen, A.J.; Brauer, M.; Burnett, R.; Anderson, H.R.; Frostad, J.; Estep, K.; Balakrishnan, K.; Brunekreef, B.; Dandona, L.; Dandona, R.; et al. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: An analysis of data from the Global Burden of Diseases Study 2015. Lancet 2017, 389, 1907–1918. [CrossRef]

[24] MacKellar, N.; New, M.; Jack, C. Climate trends in South Africa Observed and modelled trends in rainfall and temperature for South Africa: 1960–2010. S. Afr. J. Sci. 2014, 110, 1–13. [CrossRef]

[25] Van Wilgen, N.J.; Goodall, V.; Holness, S.; Chown, S.L.; Mcgeoch, M.A. Rising temperatures and changing rainfall patterns in South Africa's national parks. Int. J. Climatol. 2016, 36, 706–721.