

ESTIMATION OF VARIANCE IN CLIMATE PATTERN USING BAYESIAN ALGORITHM

D.k.APARNA,K.KRITHIKA,MRS.EBENAZER ROSELIN
UG Scholar,Prince DR.Vasudevan College Of Engineering And Technology
Email: aparna.kalidas@gmail.com,kkvl1199@gmail.com,ebenazerroselin89@gmail.com

How to cite this paper:
D.k.APARNA,K.KRITHIKA and
MRS.EBENAZOR ROSELIN (2020) Paper
Title.ESTIMATION OF VARIANCE IN
CLIMATE PATTERN
https://dx.doi.org/10.4236/***.2020.*****

Received: **** *, **
Accepted: **** *, **
Published: **** *, **

Copyright © 2020 by author(s) and
Scientific Research Publishing Inc.
This work is licensed under the Creative
Commons Attribution International
License (CC BY 4.0).
<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Increasing evidence of climate change worldwide is becoming the reason to understand a lot more about weather. To forecast weather, we need to analyse a large set of data. For efficient analysis of this big data, the HDFS (Hadoop Distributed File Systems) and MapReduce concepts are used. The proposed system designs principles for estimating variability in climate simulations using a probabilistic classification algorithm called Bayesian Classifier and an efficient inference mechanism. The goal is a “Seamless Prediction” on varying temporal scales. The system aims at forecasting variations in temperature and precipitation. These factors have large impact on societal systems such as droughts, floods etc., and provides valuable insights for climatologists and business personnel. In order to know how these applications could impact normal operations this project defines various climate predictions and challenges.

Keywords

Naïve Bayes, Climate Prediction

1. Introduction

Big data is a term that describes the large volume of data both structured and unstructured, it refers to the huge data sets obtained from various sources such as social media, sensor data, public data. Big data analytics is the process of examining large data set to uncover hidden patterns, unknown correlations. The objective of this model development is to minimize the cost and to provide more accuracy in weather prediction system. The goal is to provide "seamless prediction" on a varying temporal scale. Weather forecasting system is a complicated task because large amount of elements are taken into consideration. Until now the weather forecasting result is not accurate it is based on calculation and prediction. The traditional weather prediction system is based on numerical weather prediction model (NWP). [1] Data mining knowledge are also been used in the weather forecasting problem probabilistic graphical models (Bayesian networks) in meteorology as data mining technique. Data mining involves the use of complicated data analysis tools to discover previously unknown, interesting patterns and relationships in large data set. [2][4] This Southeast Asia region is characterized by complex terrain and land-water contrasts, tropical forests, and many islands. The climate of SE Asia and especially the maritime SE Asia is mainly tropical-hot and humid all year round with plenty of rainfall. Majority of the Southeast Asia region is influenced by monsoon and much of the region is affected by extreme weather events, particularly tropical cyclones, droughts and floods. [7]

2.Existing System

These days the numerical weather prediction model is used to predict weather along with the various permutations of different models. NWP is a computerized model of the atmosphere, derived from observations can generate forecasts of how the weather will evolve into the future. Since various models using big data for forecasting came into existence the accuracy of weather forecasts has increased making society becoming more sensitive to weather. [1] previously used algorithms for climate prediction system are 1) Decision tree 2) The sliding window algorithm 3) K-Nearest Neighbor algorithm

2.1. DECISION TREE

The decision tree algorithm was previously used in climate prediction system. In the decision tree algorithm the information are been obtained and combined in a specific manner and the information are been represented in the form of tree structure

2.2.THE SLIDING WINDOW ALGORITHM

The major disadvantage of the decision tree algorithm is that it is suitable for smaller value. The historical data are not found to be accurate they are found to be changing often the process of predicting weather using historical data leads to a large amount of error.

That problem is solved by a sliding window algorithm. Weekly weather trends may not align perfectly with historical data. The sliding window algorithm works in the following manner in which the climatic data report is been obtained and the information and been splitted and grouped up in order to find the best fit.

Advantage of this algorithm is that the data that are been used for observation need not be the data that was used last year for the process of calculating the weather. Hence a small amount of memory space is enough in order to store the information and the process of predicting the climate is faster.

Disadvantage of this algorithm is that historical data values are not taken into consideration which leads to skewed prediction. That is the slight change or deviation from the actual output.

2.3 K-NEAREST NEIGHBOR ALGORITHM

The base of Nearest-neighbor classifiers is learning by resemblance, which is by comparing a given test sample with the available training samples which are similar to it [2][4].

3.PROPOSED SYSTEM

Until now, whatever we've seen is known as traditional methods of weather forecasting. Since there where a lot of disadvantage in the traditional method in order to overcome this we use the naïve bayes algorithm.

The remaining machine learning technique is a *Bayesian Network* which ultimately uses machine learning algorithms to find the most optimal Bayesian Network and parameters [temperature, humidity, outlook, etc.]. The computation cost of the Bayesian Network is very expensive because of a large number of different dependencies [2][4].

Naïve Bayes algorithm is found to be more effective than decision tree due to the following reason a lot leads to complex trees and raises probability you are over fitting. Decisions trees will pick the best features for you from tabular data. Decision trees work better with lots of data compared to Naive Bayes. Naive Bayes is used a lot in robotics and computer vision, and does quite well with those tasks. Decision trees perform very poorly in those situations.

The accuracy of naïve bayes algorithm 0.737. Meanwhile the average accuracies of Decision Tree and k-NN are 0.589 and 0.567, respectively [2].

Naïve bayes algorithm is found to be more effective than Sliding Window algorithm due to the following reason because the Sliding window algorithm uses historical data that has been sampled may not reference the current trend whatsoever. This could happen due to impactful weather conditions that had not been observed in previous years. Occurrences like this would result in a skewed prediction due to the lack of similar data. Meanwhile the naïve bayes algorithm can handle both continuous and discrete data. It is highly scalable with the number of predictors and data points. It is fast and can be used to make real-time predictions. It is not sensitive to irrelevant features

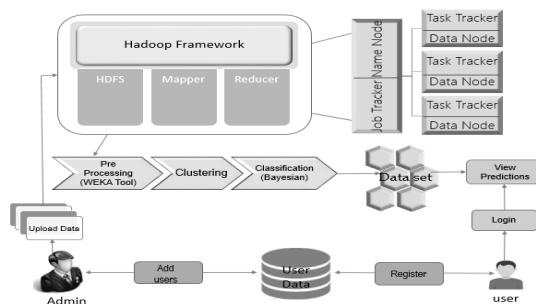


Figure 1. System Architecture

USER AUTHENTICATION:

Every last client login the page at that point makes the exchange and utilize this application. Validness is confirmation that a message, exchange, or other trade of data is from the source it cases to be from which includes verification of character. We can check validness through confirmation. Client need to Fill the all pre-requisite for security reason. Every one of the subtle elements spared in various ways and stored in table. Make new table for every client and spare points of in-

terest in like manner table

DATA UPLOADING:

The readied information are been stored in HDFS (Hadoop Distributed File System)only less memory space is been used to store the information in hdfs. HDFS consist of two nodes namely the data node and the data node .The information are been stored in the data node and the data node is been controlled by the name node. The name node acts as a master node while the data node acts a slave node.

PREPROCESSING:

Information pre-processing is an information mining method that includes changing crude information into a reasonable arrangement. True information is frequently inadequate, conflicting, as well as ailing in specific practices or drifts, and is probably going to contain numerous blunders. Information pre-processing is a demonstrated strategy for settling such issues. The transferred information recover from the hdfs. The recovered information going to the map reduces calculation and information will be composed into organized configuration. In this procedure have expelling the unusable qualities from the informational indexes. The information diminishment process is lessened portrayal of the information in an information distribution centre.

DATA CLUSTERING:

To gather those information into those bunches whose climate information class has been as of now characterized. In this way it develops a procedure to anticipate the promoting of the up and coming days. This one procedure gathering the information and ought to be made out of focuses isolated by little separations, in respect to the separations between groups. The information will gathering in light of the value, open, high, low, shut and time. In this bunching will apply the map reduce with the help of Bayesian approach.

BAYESIAN CLASSIFICATION:

It is a statistical classifier and probabilistic prediction and it predict class membership probabilities. It is based on Bayes' theorem. Naive Bayesian classifier can compare performance with decision tree and selected network classifiers. The accuracy and speed is good when applied to large databases. It is a Class Conditional Independence. The effect of an attribute value on a given class is independent of the values of other attributes. It simplifies computations. The Bayesian Belief Networks has graphical models, and it will represent dependencies among subsets of attributes. It models dependencies between variables. It is defined by two components one is Directed Acyclic Graph and another is Conditional Probability Table (CPT) for each variable. It allows a subset of the variables to be conditionally independent.

REPORT PREDICTION:

The cluster value will be different ranges. Those values are gathered and compared to each other's. Finally we will get the low and high result based on the calculation. The performance values are been shown in the form of graph.

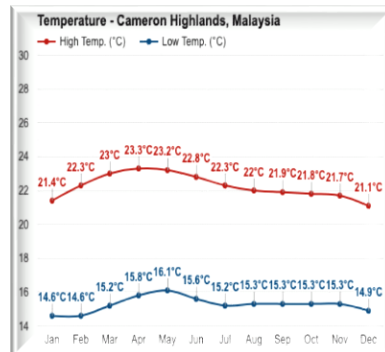


Figure 2. Climate prediction

RESULT AND DISCUSSION:

Data volume and variety are growing at very fast rate and this is becoming a great challenge in weather forecasting; as now difficulty is to mix these data to provide correct forecasts . [8] Since we have tremendous amount of complex data, so transport, storage and management of data is becoming a problem and is also increasing the overheads. Complex too large required to manage this much amount of data, which can cause millions of dollars of overhead. If in case weather forecasting goes wrong it could lead to shut down of many businesses causing severe loss to the economy and society. So if we could understand the nature of these applications and challenges, we can identify the optimal solutions to these challenges and can have more efficient and reliable applications which can save lives, improve quality of life and business, reduce risks and enhance profitability.

4.FUTURE SCOPE

In the future work, day to day predictions can be implemented with the help of APIs and accuracy should be higher. New Algorithms and features can be added for efficiency purpose

5.Acknowledgements

I would like to express great fullness to U.G..Department of Computer Science Engineering, Prince. Dr.K.Vasudevan College Of Engineering and technology, Ponmar

6.References

- [1] Sue Ellen Haupt and Branko Kosovic,National center for Atmospheric Research Boulder,CO USA,2015
- [2] Nishcala C . Barde and Mrunalinee Patole,RMD Sinhgad School Of Engineering,Savitribai Phule Pune University,Pune,India,2016
- [3] Saleem Khalid Shaab,Tharmar Community College,Thamar,Yemen,2013
- [4] Sayali D.Jadhav,H .P. Channe,Savitribai Phule Pune University,Pune,India,2016
- [5] Ugur Cayoglu,Frank Tristram,Karlsruhe Institute of Technology,Hermann-von-helmholtz-Platz,2018
- [6] Ashish Juneja,Nripendra Narayan Das,Manav Rachna Internal Institute of Research and studies ,Faridabad,India,2019
- [7] Satyaban B. Ratna,J.V. Ratnam and Toshio Yamagata,Application laboratory ,Yokohama,Japan,2016
- [8] P.Chandrasekar Reddy,St.Martin's Engineering College,Hyderabad,Telangana,India,2017
- [9] E.Sreehari,Dr.Satyajee Srivastava,School Of Computing Science And Engineering GalgatiyaUniversity,Uttar Pradesh,India,2018
- [10] Yang Yuanyuan,Zeng Tao,Yu Yongquan,Guangdong University of Technology,Guangzhou,China,2008

IJSER