TEST Engineering & Management

July – August 2020
ISSN: 0193-4120 Page No. 570 - 574

Meteorological Big Data mining: A survey

Keshani Vyas[1], Swati Patel[2]

[1] *Final Year Student of ME (IT), L. D. Engineering College, Ahmedabad, Gujarat, India*

[2]*Assistant Professor, L. D. Engineering College, Ahmedabad, Gujarat, India E-mail: [1] khvyas1997@gmail.com, [2]swati.ldce@gmail.com*

Article Info
Volume 83
Page Number: 570 - 574
Publication Issue:
July - August 2020

Article History
Article Received*: 06 June 2020*
Revised*: 29 June 2020*
Accepted*: 14 July 2020*
Publication*: 25 July 2020*

Abstract

Meteorological data is generated on day to day basis through various techniques. Hadoop distributed file system is used to handle such a rapid growth of data by storing data in distributed system. There are many architecture exist which uses Hive and pig query tools to extract data from HDFS and then apply Mapreduce for analyzing data. Apache spark is also used for faster analysis. To further improve time efficiency for large datasets, data mining algorithms are also used in distributed manner. All work has been done on traditional framework with the lack of any user interface and effective visualization. We will propose a framework which is more efficient to handle meteorological data with user interface and effective visualization.

Keywords— Meteorological data, big data, data mining.

I. INTRODUCTION

Meteorological data comes under the big data as it is sharing the same characteristics of big data. Meteorological data is growing rapidly. Meteorological data is large volume of data which is generating at rapid rate. In addition the meteorological data is unstructured data with inconsistencies and uncertainty. To store, access and handle these much amount of data is a crucial task [1].

There are different tools and techniques available for handling large volumes of meteorological data. **Hadoop** is core of most technologies. Hadoop provides the best framework for storing unstructured data in a distributed manner. It also provides fault tolerance and high throughput for accessing the data. **Hive** is the data warehouse tool, built on Hadoop platform. It is query language through which user can query the data. **Apache spark** is used to perform some analysis on the data faster than mapreduce. It also provides support for multiple languages like java, Scala and python. It also supports SQL queries, machine learning, streaming and graph algorithms. **HBase** is column oriented NoSQL database which is used to store data in table format and we can use hive, pig query tools to extract data from it. **Cassandra** is also column oriented NoSQL database which provides flexible data storage. It is used where write is major operations as it performs write operations faster.

The meteorological data is unstructured data which is in raw format so we have to transform it into usable format. Data cleaning is performed to handle missing values and noisy data. Data transformation is also applied on the data by normalizing and deriving new attributes in data.

Different data mining techniques can be applied on the meteorological data. If the dataset is labelled then to classify the day as rainy, foggy and sunny supervised algorithms like Decision tree, random forest, Naive Bayes and support vector machine classification

July – August 2020
ISSN: 0193-4120 Page No. 570 - 574

algorithms can be used. To predict the rainfall in future, regression can also be applied on the data. K means is applied on unlabeled data to find the clusters based on similarity of data point values. Anomaly detection algorithms are also applied to detect abnormal events like flood, storm and many more.

II. REVIEW OF VARIOUS BIG DATA MINING APPROACHES

A. The Establishment and data mining of meteorological data warehouse.

In this paper [2], the authors have proposed warehouse system for handling increasing growth of meteorological data. The design of warehouse system is based on Hadoop distributed file system. The proposed system is basically master slave based architecture with three layers as shown in below Figure 1.

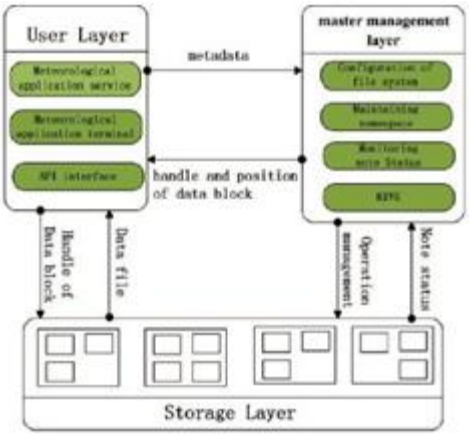

Figure 1. The framework of Meteorological data warehouse file system

In storage layer Hadoop distributed file system is used for storing large amount of meteorological data into distributed manner. Hadoop contains Hadoop Distributed File System (HDFS) which provides fault tolerance and high throughput compare to normal file system. HBase is used for storing the unstructured data into some tabular format for easily querying and analyzing data in future use. Map reduce is used for performing some computation on the data. Hive is used to generate map reduce tasks for analyzing metadata of meteorological data which is stored under HBase. In User layer data management is done through terminal and meteorological application services. In master management layer system configuration, namespace maintenance and operational management is done. The framework is tested by applying other datasets such as PokerHand, Adult, Wilt, Yeast, Balance scale and wine by considering all attributes of the datasets. Map reduce is applied on these datasets which gives output in key value form. As these datasets falls under classification category of data mining, Improved Naive Bayes algorithm is applied on the resultant data from mapreduce task and testing is performed on different dataset. Performance evaluation is done by comparing accuracy and timing of testing dataset for proposed system and traditional system. As Table 1 and Table 2 shows for all datasets, testing time is reduced and accuracy rate is increased because of storing and processing data in a distributed manner.

The table of using the improved naive bayes algorithm

Data set	Training time (s)	Testing time (s)	Accuracy rate (%)
PokerHand	28	662	92.4006
Adult	4	14	82.1560
Wilt	3	2	78.9553
Yeast	2	1	42.0866
Balance Scale	2	1	72.5642
Wine	2	1	90.3964

Table 1. Improved Naive Bayes Algorithm

The table of using the traditional naive bayes algorithm

Data set	Test time (s)	Accuracy rate (%)
PokerHand	954	91.7877
Adult	17	79.3333
Wilt	4	75.8975
Yeast	1	42.1976
Balance Scale	1	72.5642
Wine	1	91.0569

Table 2. Traditional Naive Bayes Algorithm

In future, the authors can compare the performance of other classification algorithms

July – August 2020
ISSN: 0193-4120 Page No. 570 - 574

such as Decision tree, Random Forest and Support vector machine with respect to Naïve Bayes.

B. Supervised classification of rainfall coverage data in Andhra Pradesh using support vector machine.

In this paper [3], the authors have used Meteorological data which is in the form of TIFF. Tiff format comes under the raster type of data. Raster images are collection of pixels which are represented in matrix form. The rainfall dataset is preprocessed and then data of Andhra Pradesh region is extracted from the whole dataset. For training purpose some samples are extracted from the selected data. After training phase SVM model is applied on the data. In testing phase, model is validated by giving unknown data as input to that model and find accuracy of the model. In testing phase input data should not be the part of selected training samples of Andhra Pradesh region. The performance evaluation is done using confusion matrix and error matrix which represents how many values are correctly classified or not. Regions of Andhra Pradesh map is classified as Excess if rainfall coverage is greater than 20%, Normal if rainfall coverage is -19% to +19% and deficient if rainfall coverage is -20% to -59%. The confusion matrix shows that all values are categorized correctly and model achieved 100% accuracy for classification based on district map of Andhra Pradesh.

The authors have not mentioned the data size. In future, the authors should also consider variation of rainfall coverage inside every districts.

C. SMASH: A Cloud-Based Architecture for Big Data Processing and Visualization of Traffic Data

In this paper [4], the authors have proposed one software stack named by SMASH which contains components like Hadoop distributed File System (HDFS), Spark, Apache Geomesa accumulo and Geoserver.The authors have used cloud facility provided by Australia under the NeCTAR project. To handle large amount of data the authors have created range of virtual machines with some quite good specifications like 4GB core and 30 GB storage to Highly efficient machines with 16GB cores and 480 GB storage in which storage can be extended to 100 petabyte also. The data is collected from the transport department of Sydney Coordinated adaptive Traffic system. Traffic data is spatio temporal data in nature. Geomesa is used for indexing of spatio temporal data, and Geoserver is used for visualization of data with maps. Testing of The SMASH architecture performed using Adelaide and Victoria dataset which contains 181 million records from 2008 to 2014 year. Distributed environment is created using 4 Machines with 4-core, 30GB disk and 12GB RAM specifications. SMASH is capable of processing 15GB dataset without exhausting resources. As compare to traditional Hadoop map reduce system, SMASH performs better by providing better big analytic platform. Use of Accumulo and Geomesa in platform makes it more suitable for handling spatiotemporal data efficiently. Additionally, spark is used rather than hadoop mapreduce which performs better for processing large data.

In future, the authors can give users flexibility to select areas of interest and perform analyses of it and at the end visualize it in a more effective way.

D. A novel clustering of big data in hadoop ecosystem

In this paper [5]**,** the authors have proposed hybrid clustering on big data. To handle big data Hadoop has been used. The data is collected from National Climatic data center (NCDC) from 1901 year, the dataset contains hourly data from 20000 stations from full globe. The K means and Hierarchical clustering are used for hybrid approach. In the hybrid approach, K means is applied on a

July – August 2020
ISSN: 0193-4120 Page No. 570 - 574

dataset and mapper class will map the data point with the centroid of the cluster in which the data point belongs. After K means Hierarchical algorithm is applied on the remaining data points. Hash map is applied to remove redundancy from the clusters formed by both the algorithms. Input of the Mapper class is weather file for specific year. Mapper is applied on both the algorithm and output of mapper class is given as input to the reducer class. Output of Mapper class is a key value pair here key is date and value is temperature. Output of reducer is aggregated value for unique key. Reducer will merge the cluster of both algorithms. Precision, Recall and F1 score is used for performance evaluation of the algorithms. Hybrid clustering have 99% precision, 82 % Recall and 90% F1 score.

Hybrid clustering is taking more time to execute than kmeans and hierarchical clustering methods, Different parameters needed to reduce execution time of hybrid clustering.

E. A study of Rainfall over India Using Data Mining

In this paper [6], the authors have applied data mining techniques such as classification and clustering to analyze spatiotemporal data. The daily rainfall data is developed by Indian meteorological Department, Pune. The data is collected over the year 1951 to 2003 by all stations with minimum 90% data availability and other values are filled using interpolation. The authors have designed a system to handle climatic data. Algorithms are developed to extract climate data from the file which can be in the form of ASCII, NetCDF, and HDF format. All different format files needed to be converted into a single form such as ASCII format. Average rainfall over India is calculated for monsoon season and daily basis. Clustering is applied to the data and clusters are formed based on rainfall range. Cluster 1 depicts 0.1 to 3 mm rainfall, cluster 2 depicts 3 to 5 mm rainfall, cluster 3 depicts 5 to 10 mm rainfall, cluster 4 depicts 10 to 25 mm rainfall, and cluster 5 depicts rainfall greater than 25 mm. Classification is applied to the data with user defined thresholds as every region indicates different variation of rainfall. The approach can be extended by applying more data mining algorithms and more effective visualization of clusters.

III. CONCLUSION

This paper presented different approaches for handling and processing meteorological data. Data warehousing and distributed storage based frameworks are used for storing large amount of data. Data mining algorithms such as Naïve Bayes, SVM and Kmeans are used for classifying rainfall events and creating clusters based on rainfall events. However storing large amount of data and applying data mining techniques on them is not enough we need distributed and efficient processing of the meteorological data. Hadoop mapreduce will process the data in distributed manner but it takes more time compare to Spark processing. In future, the researcher should focus on developing the system that enables user to store any kind of meteorological data into distributed storage, process the data and provide visualization of results on interactive interface.

ACKNOWLEDGMENT

I owe a great many thanks to all people who helped and supported me during the completion of this paper. My deepest thanks to internal guide at L.D. Engineering College prof. Swati Patel for guiding and rectifying several actions and steps and documents with attention and care. She has presented a decent amount of attention during the course of the process and made necessary correction as and when needed.

July – August 2020
ISSN: 0193-4120 Page No. 570 - 574

REFERENCES

[1] Big data characteristics: **https://www.geeksforgeeks.org/5-vs-of-big-data/**

[2] L. Shao, J. Liu, G. Dong, Y. Mu and P. Guo, "The establishment and data mining of meteorological data warehouse," 2014 IEEE International Conference on Mechatronics and Automation, Tianjin, 2014, pp. 2049-2054.

[3] VenkataNagendra, Kolluru. (2017). Supervised Classification of Rainfall Coverage Data in Andhra Pradesh Using Support Vector Machine. International Journal for Research in Applied Science and Engineering Technology. V. 1198-1204. 10.22214/ijraset.2017.10172.

[4] Sinnott, Richard & Morandini, L. & Wu, Siqi. (2015). SMASH: A Cloud-Based Architecture for Big Data Processing and Visualization of Traffic Data. 53-60. 10.1109/DSDIS.2015.35.

[5] S. Kumar and M. Singh, "A novel clustering technique for efficient clustering of big data in Hadoop Ecosystem," in Big Data Mining and Analytics, vol. 2, no. 4, pp. 240-247, Dec. 2019.

[6] Chowdari K.K, Girisha R and K. C. Gouda, "A study of rainfall over India using data mining," 2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), Mandya, 2015, pp. 44-47, doi: 10.1109/ERECT.2015.7498985.