Weather Forecasting by using Modified K-Means Intra and Inter Clustering Algorithm

Veera Ankalu Vuyyuru, G.Appa Rao

Abstract— Urban air pollution causes biggest to the human being. Monitoring and controlling of air pollution becomes an essential thing. Deals with large dataset when forecasting the weather. Hadoop is popular for storing and processing. K-means clustering finds resemblance in small dataset. In proposed system k-means hadoop mapreduce (KM-HMR) deals with implementation of mapreduce with standard k-means clustering. And KM-I2C k-means inter cluster, it maximizes the distance between the cluster and intra cluster minimizes the distance between the clusters. This approaches increases the quality of cluster it becomes efficient and effective.

Keywords: Hadoop , Mapreduce , Clusters ,Forecast

1. INTRODUCTION

The sensor network useful in sensing ,processing and making communication with the nodes. Wireless sensor network has n number of nodes these are collecting and processing the information and send it to the centre location.

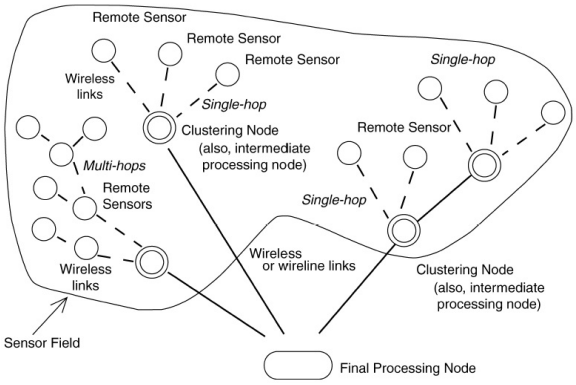


Figure 1: Arrangement of Sensor

Wireless sensor network has only the low duty cycle, limited battery life and power constrains. Capability of the sensor provides the detailed information about the gas, electric and water. And also provides the data about temperature, cooling and heating effect. Automatic notification is provided the sensor if any unusual events occurs in atmosphere.

1.1 Clustering

The sensor nodes are partitioned into cluster [3]. Each cluster has the cluster head. The data from the nodes are received by the cluster head and pass it to the base station. The energy consumption and number of active nodes communication is minimized [4]

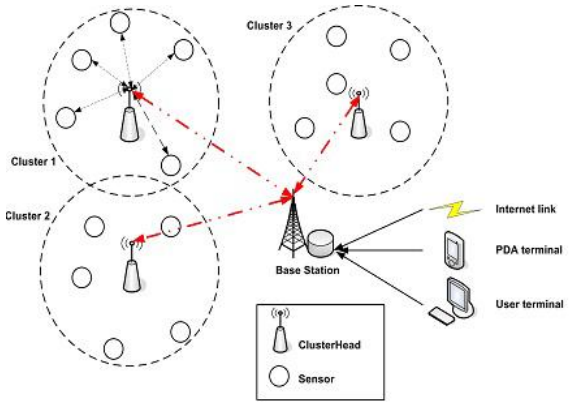


Figure 2: clustering

2. WEATHER FORECASTING APPROACHES

2.1 Persistence

The short weather condition is forecasted by persistence in a easy way. In this dependency is mainly on the weather pattern in idleness behaviour. This approach can be of use during summer or winter season. And mainly used in the region of steady state weather conditions. So it is imitated to only some regions. And cannot be taken in to consideration for real time applications and for long term forecasting.

2.2 Usage of barometer

Long range weather condition can be predicted by using the barometer. The millimetres of mercury or hetopascal pressure are taken in to consideration to measure pressure level. If the pressure level is more than 3.5 hpa or 2.6 mmHg it indicates large changes in the weather condition. During raining the pressure level get dropped suddenly becomes prediction of very low pressure. Weather prediction can be done in the clearest sky. It is considering for long predictions and not taken in to consideration when the weather changes suddenly.

2.3 Sky surveillance

The current weather is forecasted by the people by sky observation. Sky covered with the clouds during raining time. This approach is used for forecasting the instant weather state.

Revised Manuscript Received on July 10, 2019.

Veera Ankalu Vuyyuru, Research Scholar, Department of CSE, Gitam University, Andhra Pradesh, India. Email: veeraankalu14@gmail.com

Dr.G.Appa Rao, Professor, Department of CSE, Gitam University, Andhra Pradesh, India , Email : apparao.giduturi@gitam.edu

Prediction of weather condition by using sky surveillance predict cyclone and floods but its very difficult to save people, cannot gives correct observation result in all situations.

2.4 Analog approach

This approach helps to decide the occurrence of next event with the signal. It is an complex approach expert only analyzes the history of analog approach. The analog signal cannot be used for predicting all the time.

2.5 Use of weather map

There are two categories of weather map surface map and upper air map. Surface map has the information of cloudiness and patterns of rain. The changes occurs in the weather gets recorded every hour. But upper makes a design every twelve hours. It analyzes the radar and satellite images contains information about temperature, wind, humidity level and pressure level of an atmosphere.

3. PROPOSED SYSTEM

3.1 Air quality

The atmosphere is affected with heavy mixing of contaminated pollution in air. Productions made by many factories and exhausts of vehicles polluting the atmosphere. Air pollution protecting is environmental saving and becomes mandatory for living of human being [1]. This contamination can controlled by measuring, monitoring and forecasting. The air quality is depend upon the ambient air . the foundation is established for managing the air quality. It monitors the air condition of particular city or region. If any emissions occur it analyzes is that an wanted. Mostly monitoring depends on online analytical processing, report editing, data mining and predictive analytics. When processing moves on it results in large volume of data. So k-means hadoop mapreduce process s used to make the clustering method applicable for an large data set [1].

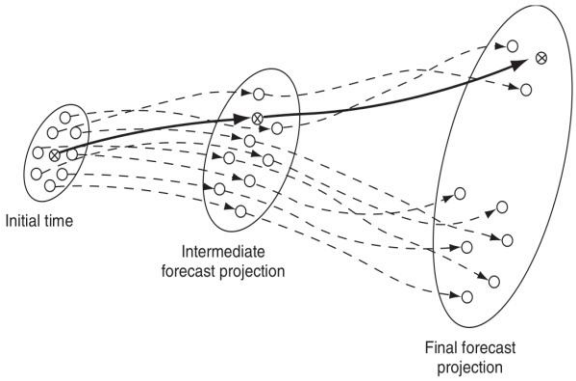
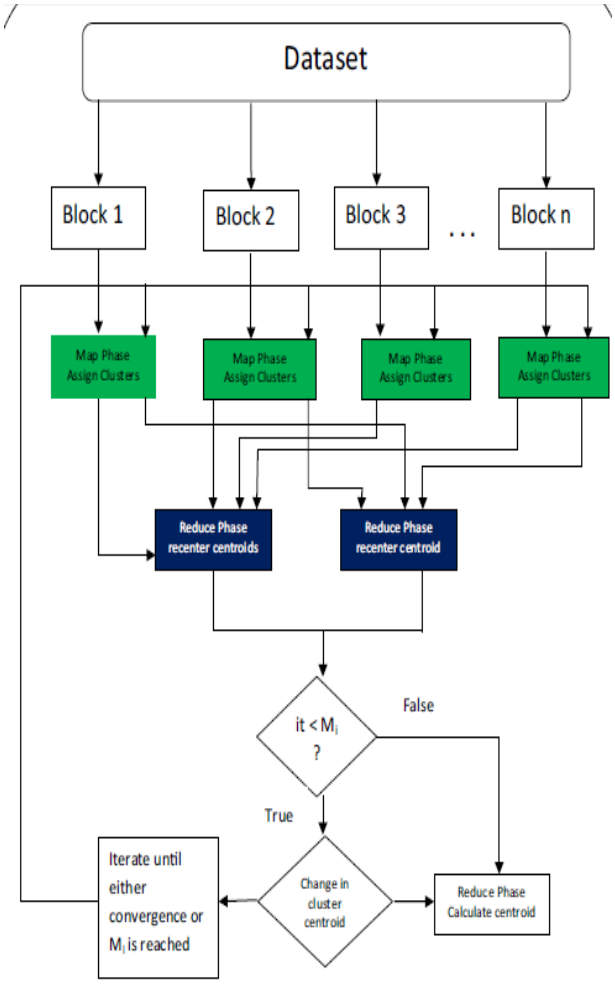


Figure 3: Projection of Forecasting

3.2 K-Means Hadoop Map Reduce

This algorithm focuses on implementation map Reduce with k-means pattern, this serves as structure for clustering. K-means hadoop mapreduce finds the clusters faster when compare to the standard k-means cluster pattern [1]. The Mapreduce splits dataset into chunks is fixed size parts it is inspired as single map.



The algorithm is improvised by combining k-means with an map reduce. The given dataset is segregated into blocks, those blocks are assigned into clusters. Send the mapper output to the reducer. The number of reducer can be assigned to the mapper.

The reducer group recenter the centroids and merging the centroids to global centroids. the result is put in the the output directory.The mapper phase measure the distance between the objects and clusters. The Euclidean distance is used as distance metric in k-means hadoop mapreduce. Once distance is calculated, object is assigned to the cluster which is closer to the object. Single map task is initiated for each inputsplit, and gets performed by map function of each record of inputsplit[2]. The size of the inputsplit is 128MB. Objects takes place in same clusters are sents to reduce phase.

The new cluster centroids are calculated by the reduce phase for next map reduce job.Cluster centroids are produced during the last stage of initial iteration gets, accumulated in the old cluster file. Once stored, new cluster centroids value is tested and get updated. The value of iterations is increment by one. This process executed again when there is no change in the cluster centroid value [1]. The final output clusters are in result file.

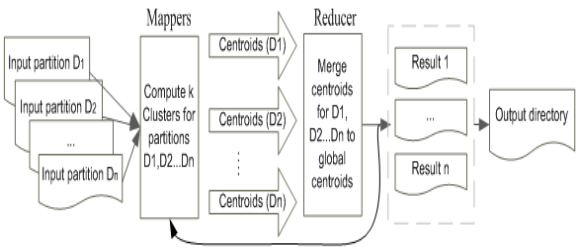


Figure 4: K-Means Mapreduce

It improves the efficiency of the clusters. Handling large dataset with HDFS. The dataset is large and unstructured type of data processed and result generated in faster manner. Large volume of data is processed parallel with mapreduce programming model.

K-means hadoop mapreduce :

```
Input:
Obj : { obj1,obj2,obj3,…..objn}- clustering the
objects,
C :  no.of. clusters
Ni: no. of iterations (maximum)
Output:
KHMapReduce(data)
X= 0
for each datapoint(dp)_ DP do
        y= select(c, dp)
        input(dp)
                write(y)
z=y
while (true)
        mapper-invite work.mapper( )
        reducer- invitework.reducer( )
n = examine()
// do again until convergence
if inform((n, z) > 0 )
        z = n
else
        inform n to result
x++
res = examine( )
```

The input partition is put in to mapper to invoking the working of an mapper. Mapper computing the clusters and reducer merging the centroids .the result is examined and put in the output directory.

3.2 K- means intra clustering and inter clustering (KM-IaeC)

The best cluster has high level of similar intra clusters and low level of similar inter clusters. This algorithm divides the distinctive clusters in to sub clusters. The dividing of clusters based on the intra clusters and inter clusters. Formation of clustering maximizes the distance of the inter clusters and minimizes the distance of the inter clusters. The clusters quality is determined with similarity and implementation measures.The distance measured in inter and intra clusters

$$Ie = \frac{1}{2}\left(\frac{\sum_{i=1}^{O1}\sum_{j=1}^{O2}(Ai - Bj)^2}{O1 * O2}\right)$$

$$Ia = \frac{1}{2}\left(\frac{\sum_{i,j=1}^{O1+O2}(Ai - Bj)^2}{(O1 + O2) * (O1 + O2 - 1)}\right)$$

Ie – inter cluster
Ia – intra clusters.
O1,O2 – cluster – data points .
Mapper protocol:

```
Input:
        N dimensional data objects (nobj1,nobj2,
nobj3,....nobjm) in each mapper
        C:  no.of clusters
        Initial cluster centroids ce1,ce2,ce3,ce4,....cek
Output:
        output  <key, value>
newlist :  list of new centroids
value = zero
newlist = zero
        for (dp = Dp)
                for (ci= N) do
ccentrodi ← Ø centroid nearer to object
ie ←infinity
ia ←infinity
for ( obji= O)
        l(obji) ← F(obji,objj) q = {1,2,3,...n}
x=zero
ccentroid= zero
do again
for (ei= E)
        minimumdistance← F (obji, objj) q ={1,2,3,...n}
if (currentcentr = zero or one  (obji) < minimumdistance)
        updating  inter cluster
else
        updating intra cluster
ccentroidi← ccentroidi+1
x =x+1
 output list is created <key, value>
```

```
Input:
        key=1(obji),
        value = mappers allocate objects  to centroids
        mo: mappers output list
Output:
newlist: list of new centroids(gc)
newlist = 0
gc Ø
for ( x = mo)
        centroid =x.key
        data object= x.value
        gc= dataobject
for  (ci= M)
gc= Ø
Sumobjects= Ø
Numobjects= Ø
for  (obji= O)
SumObjects + = Object
Numobject++
gc (Sumobjects/Numobjects)
Outputlist=gc list u gc
Return gc
```

The dataset is stored in hadoop distributed file system as key, value pairs. The dataset distributed to a number of mappers. It shares the files which contains all cluster

centroids. The mapper produces the key ,value pair that are combined by the hadoop for reducer phase. The final output is stored in gc file.

4. MULTIDIMENSIONAL DATA SET

Each pollution air qualities daily sub-index is measured and stored depends upon the gathered data. The data which is stored are taken in to process for producing data in a particular location for particular pollutant. The datas are gets loaded into multidimensional OLAP cube giving possibility of analyzing the query in big amount of dataset [1].

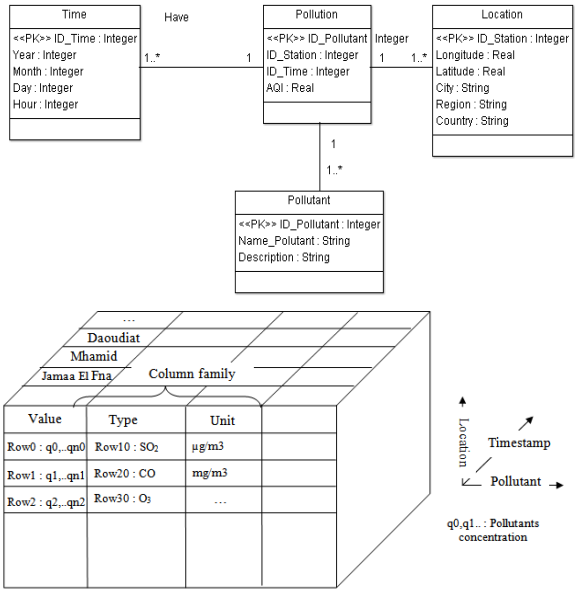


Figure 5: Multidimensional Cube Of Air Quality Monitoring

5. EXPERIMENTAL RESULTS

The air pollution data base considered the results. the nodes are clustered to the nearest nodes initially. It is capable of clustering the largrset with k-means hadoop mapreduce.

Cluster Number	*Cluster Maximum data*
Cluster0	221.376238
Cluster1	118.562500
Cluster2	41.523529

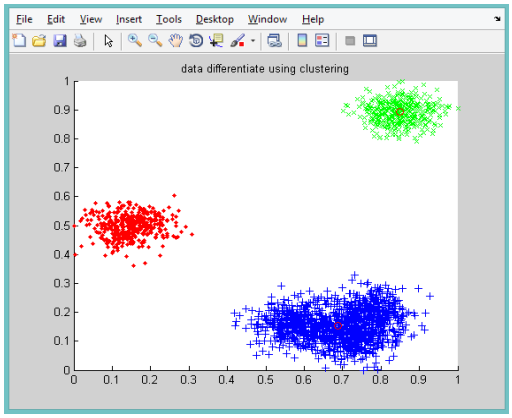


Figure 6: Data Cluster

The cluster quality and execution time is improved with k-means inter cluster and intra clusters. The distance between the inter clusters is maximized and intra clusters is minimized. The cluster quality is improved by k-means hadoop mapreduce.

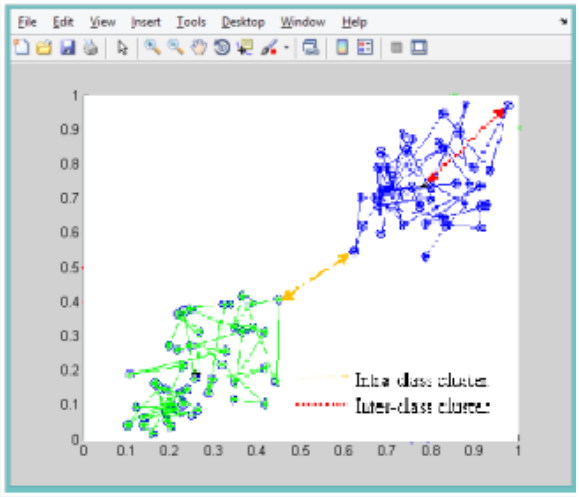


Figure 7: k means intra and inter clusters

6. CONCLUSION

The climatic change occurs frequently .so prediction of weather condition become a greatest challenge. The proposed k means hadoop mapreduce clustering algorithm is used for predicting the air pollution with large data set by implementation mapreduce with k-means. K-means inter and intra clusters increases the quality of clusters. It improves the accuracy of the cluster and clustering time.

REFERENCES

1. Abderrahmane SADIQ 1, Abdelaziz EL FAZZIKI 1, Jamal OUARZAZI 2, Mohamed SADGAL 1, Air Quality Analysis Based On MapReduce and K-Means: A Decision Making System, Int. J. Advance Soft Compu. Appl, Vol. 9, No. 2, July 2017ISSN 2074-8523
2. Li, S., Chou, S., and Pan, S. 2000. Multi-resolution spatiotemporal datamining for the study of air pollutant regionalization, In Proceedings of the33rd Hawaiian International Conference on System Sciences, Island ofMaui, Hawaii.
3. Vikas Verma*, Shaweta Bhardwaj and Harjit Singh, A Hybrid K-Mean Clustering Algorithm for Prediction Analysis, Indian Journal of Science and Technology, Vol 9(28), DOI:

10.17485/ijst/2016/v9i28/98392, July 2016
4. Vaibhavi Mistry1, Vibha Patel2 ,Weather Condition Prediction Using Semi-Supervised Data Mining Technique, International Journal of Engineering Trends and Technology (IJETT) – Volume 20 Number 4 – Feb 2015
5. Dean J. and Ghemawat S. 2008. MapReduce: simplified data processing on large clusters, Communications of the ACM, Vol.51, No.1, 107-113.
6. McCallum A, Nigam K, Ungar LH. Efficient clustering of high-dimensional datasets with application to reference matching. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining.2000. p. 169–78.
7. Quinlan JR. Programs for machine learning. San Francisco: Morgan Kaufmann Publishers Inc; 1993. ISBN1-55860-238-0.
8. Domingos P. Linear-time rule induction. In: Proceedings of knowledge discovery and data mining (KDD-96), Portland,Oregon. 1996.
9. Fisher D. Knowledge acquisition via incremental conceptual clustering. Mach Learn J. 1987;2(2):139–72.
10. heeseman P, Stutz J. Bayesian classification (AutoClass): theory and results. In: Advances in knowledge discovery and data mining, 1996. p. 153–80. ISBN: 0-262-56097-6.
11. Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R, editors. Cambridge: AAAI/MIT Press; 1996. p. 153–80.
12. Biswas G, Weinberg J, Li C. ITERATE: a conceptual clustering method for knowledge discovery in databases. In:Braunschweig B, Day R, editors. Innovative applications of artificial intelligence in the oil and gas industry. 1995.
13. Das G, Mannila H, Ronkainen P. Similarity of attributes by external probes. In: Proceedings of the fourth international conference on knowledge discovery and data mining KDD'98. New York: AAAI Press; 1998. p. 23–9.
14. Ortega M, Rui Y, Chakrabarti K, Mehrotra S, Huang T. Supporting ranked boolean similarity queries in mars. IEEE TransKnowl Data Eng. 1998;10(6):905–25.
15. Dhillon IS, Modha DS. Concept decompositions for large sparse text data using clustering. Mach Learn.2001;42(1):143–75.
16. I.Karthika, K. P. Porkodi ,"FRAUD CLAIM DETECTION USING SPARK" International Journal Of Innovations In Engineering Research And Technology ISSN: 2394-3696 VOLUME 4, ISSUE 2, Feb.-2017
17. I.Karthika, K.P.Porkodi," AUTOMATIC MONITORING AND CONTROLLING OF WEATHER CONDITION USING BIG DATA ANALYTICS , International Journal of Advanced Research in Computer and Communication Engineering Vol. 6, Issue 1, January 2017
18. I. Karthika, P. Gokulraj, S. Saravanan" Prediction of sales using Big data analytics" Journal Of Advances In Chemistry Vol 12, No 20
19. Karthika, S. Priyadharshini " Survey on Location based sentiment analysis of Twitter data" © 2017 IJEDR | Volume 5, Issue 1 | ISSN: 2321-9939
20. S Saravanan, V Venkatachalam,"Advance Map Reduce Task Scheduling algorithm using mobile cloud multimedia services architecture" IEEE Digital Explore,pp21-25,2014.
21. S Saravanan, V Venkatachalam,"Enhanced bosa for implementing map reduce task scheduling algorithm" International Journal of Applied Engineering Research,Vol 10(85),pp60-65,2015.