K-mean Clustering of Big Data Over Map Reduce

Harpreet Kaur¹ & Dr. Maninder Singh²
¹M.tech student, Department of Computer Science, Punjabi University, Patiala
²Assistant professor, Department of Computer Science, Punjabi University, Patiala

Abstract: In today’s world Data is increasing at rapid speed in every field may it be corporate, scientific, banking or social networking sites. It is not easy to process and analyze this big data. Big data analytics and Data mining are the techniques in to analyze data and to extract hidden information. But the traditional approaches do not fit well for big data as this data is of very high volume and quite complex. A major mining technique is data clustering which cluster the massive data into clusters which make it easy to extract any information from these clusters. Data clustering is used in social networking sites, weather forecasting etc. Map Reduce is a parallel programming paradigm which can be applied to fasten the processing of data by parallel mapper and reducer functions. Hadoop is an open source framework that uses the concept of map reduce and deals with big data effectively. Dataset used is NCDC weather file. K-mean clustering, a major clustering approach is applied on this dataset and maximum temperature is calculated from weather file of a particular year by making clusters through this approach. K-mean algorithm is very fast as it takes very less time to execute. F-measure is calculated based on precision and recall to check the accuracy of result and quality of clusters formed.

1 Introduction

Today big data has become buzz in the market. Big data is increasing at rapid speed these days. It is so large that applications like DBMS, RDBMS etc which were used to store and process datasets failed to handle such a large amount of data within required time. Massively parallel software is required to handle this work which runs on many servers. Big data requires such techniques which reveal insights from datasets that are diverse, complex, and of a massive scale. One such technique to handle big data is Hadoop. It uses the Map Reduce framework to analyze big data. It deals with big data and parallel processing paradigm. It has two main parts HDFS and Map Reduce.

Big data analytics is the examining and analysis of big data to find the patterns which were hidden till now and unknown correlations. Big data analytics also recognize and find useful information from large data which is used to make better decisions. A major Big data analytics technique is clustering which deals with challenges faced by big data. It provides search, recommendation and organization of data. It is a type of unsupervised learning unlike classification. In this method, similar objects of the dataset are grouped into clusters, and thus formed different clusters such that objects in the same cluster groups are very different from each other and the objects in the same group or cluster are very similar to each other.

2 K-means Clustering

K-means clustering is an important partitioning method in which Data is divided or partitioned into various clusters. It is efficient in clustering large data sets. Number of clusters is defined initially. Almost every object belongs to one of K-groups. Partitioning method leads to k number of clusters as a result where all data points belong to any of these k clusters. Every cluster has a centroid or cluster representative from which distance of all data points is measured and data points with minimum distance from centroid are kept under one cluster. This algorithm explores the structure of a data set. To apply the k-means algorithm the number of clusters needs to be predefined and randomly chosen k points may be initial centroids.

2.1 Advantages of k-mean clustering:
- K-mean clustering is simple to implement compared to other algorithms.
- If variables are huge k-means clustering is most of the times computationally faster.
- K-means produces tight clusters.
- K-mean clustering is more efficient during run time which means its speed is fast.

2.2 Dataset Used

In this research work here we test all three clustering algorithms on the dataset of NCDC ISH (Integrated Surface hourly data) weather files constructed in standard ASCII format. The dataset is globally available to everyone. This global database integrates
the surface hourly data from 20,000 stations all over the world and it includes weather files from:

- NCDC and Navy surface hourly data
- NCDC hourly precipitation data, and
- Air Force Datav3 surface hourly data,

Dataset has weather files of different years starting from year 1901. Weather is recorded on every day of a year.

The dataset consists of mostly 3 sections:

- **Control Section**: Every Record starts with fixed length control section of 60 characters. Control section usually provides information about the report. The observation date, time, station location information etc. The brief introduction of every attribute in control section is provided in table

- **Mandatory Data Section**: This section is followed by control section. Mandatory section is also of fixed length which is 45 characters long. This section contains meteorological information about the temperature, pressure, winds etc. The brief introduction of every attribute in mandatory section is provided in table

- **Additional Data Section**: This section is followed after mandatory section. It contains variable number of characters; its length is not fixed. it is not mandatory which means there can only be two sections (Control and mandatory) of a record. Also, Sometimes after additional data remarks or element quality section could also be included.

2.3 K-mean Algorithm Steps

- Initially Set number of clusters here we set \((k)=2\)
- Set two initial centeroids randomly.
- Calculate the distance from each centroid to every data point then assign nearer distant data points to the corresponding centroid and make it one cluster. Now choose mid value from data points of formed cluster and make it new centroid, again calculate distance of all points from newly formed cluster.
- Repeat this step till data points do not leave the cluster and similar data points are collected in each cluster.
- After that we partition the regions of similar data points in both clusters.
- Then we find the maximum temperature points from each region, these major data points of each region are compared with other regions in that cluster. Now we are left with few major data points from each cluster.

2.3.1 Stopping criteria in k-mean clustering:
The stopping criterion in k-mean clustering usually comes

- When no object or data point change clusters, or
- When within iterations very few change clusters.
- Stopping criteria can also come when number of iterations is predefined. Internal quality measure could also be used to define stopping criteria of k-mean algorithm.

3. Implementation framework

The major tools on which the research work has been implemented is hadoop and eclipse MARS IDE on ubuntu 15.10 operating system.

3.1 Hadoop

Hadoop is an open source framework which is suited for big data. The concept of big data is increasing at rapid speed due to large generation of massive data. So there is need for a software framework that can deal with big data. There are two main layers of Hadoop

![Figure 1-Layers of Hadoop](image)

3.2 Map Reduce

It supports parallel processing paradigm. There is one master node and many slave nodes. Map reduce phase has two stages mapper stage which accepts input in key value pairs and also produces output in key value pair. Second phase is reducer. The K-mean algorithm follows two major stages in Map Reduce parallel processing. The output of map phase is fed as input to reduce.
Mapper Stage: The input to the mapper phase is a weather file of a particular year that we selected as shown in figure. Here we extract two fields from every record: observation date and the air temperature in Celsius times ten. Also quality code is tested and checks if its value is not missing. As the map reduce paradigm is based on the key value scenario so here also Mapper divides the extracted data into key value pairs. Pass date as key and temperature as value to the reducer phase. Temperature is divided by 10. Here key is in Text Form and value is in intWritable form. The (key, value) pairs for this dataset is in the following form:

(19070101, 13.9)
(19070102, 11.7)
(19070103, 12.3)

Reduce Stage: Reduce phase accepts the output of mapper phase as its input. It receive the key value pairs in text and IntWritable form respectively. That is all the temperature values belong to a particular observation date at specified time and specific conditions. Temperature may be counted more than once but maximum up to three times on a particular day. So every key value is specific. Then reducer performance varies for different algorithms to make clusters and find maximum temperature from clusters formed.

3.3 Eclipse

Eclipse is an IDE (integrated develop environment). We use Eclipse MARS on ubuntu 15.04. It is usually used by java developers to write java code. Eclipse can extend its capabilities by adding plug-in for other languages. Here we add hadoop 2.6.0 plug-in to configure hadoop with eclipse.

4 Results and Discussions

On applying K-mean algorithm on NCDC weather dataset over map reduce framework of hadoop we analyzed following results:

4.1 Running of Hadoop for k-mean algorithm in Eclipse

K-mean algorithm produces map output records for the given input records. In the following diagram the working of k-mean algorithm over hadoop is shown. As here in particular weather file of year 1907 total map input records are 1095 whereas map output records are 360.

4.2 Precision, Recall and F-measure

K-mean is a static clustering algorithm and number of clusters need to be predefined. Here we set k=2. Number of data points in both clusters is given before and after processing. Performance of k-mean algorithm is measured based on above described factors. K-mean is very simple algorithm and takes very less CPU time to make clusters. To measure the quality of clusters formed F-measure is calculated.

- Precision

Precision is measured as fraction of pairs of data points correctly put in the same cluster. It is directly proportional to the quality of clusters formed and accuracy. Lesser is the value of precision poor is the quality of clusters formed. More is the precision of an algorithm more accurate is algorithm and more quality clusters are formed. It is measured as Deviation from actual value. For the clustering algorithms, precision is measured as clusters formed by particular algorithm divided by the actual clusters that can be formed in the dataset. Or we can say the percentage of relevant clusters returned by the algorithm.

Precision = Computed Clusters by Particular Algorithm ÷ Actual Clusters in the dataset
Recall is also measured on pair of data points, it is the fraction of actual pair of data points that were identified. Recall is also directly proportional to cluster quality.

F-measure
F-measure calculation requires two factors precision and recall. More is the precision and recall more is the f-measure. It is directly proportional to precision and recall. F-measure is calculated as

\[ \text{F-measure} = \frac{(2 \times \text{precision} \times \text{recall})}{(\text{precision} + \text{recall})} \]

<table>
<thead>
<tr>
<th>Year</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1902</td>
<td>0.49</td>
<td>0.33</td>
<td>0.39</td>
<td>4</td>
</tr>
<tr>
<td>1905</td>
<td>0.82</td>
<td>0.66</td>
<td>0.73</td>
<td>3</td>
</tr>
<tr>
<td>1907</td>
<td>0.41</td>
<td>0.24</td>
<td>0.35</td>
<td>5</td>
</tr>
<tr>
<td>2001</td>
<td>0.99</td>
<td>0.82</td>
<td>0.9</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 1. Performance measure of k-mean over different years

5 Conclusion
K-mean clustering is applied on large dataset of NCDC weather file to make clusters, for easy extraction of information. Maximum temperature of an year is calculated in very less time using this approach over map reduce framework of hadoop as it supports parallel paradigm.

6 Acknowledgement
Dataset of NCDC Integrated surface hourly data is accessible via NCEI’s Climate Data Online system (cdo.ncel.noaa.gov), FTP (ftp://ftp.ncel.noaa.gov/pub/data/noaa/), GIS services (gis.ncel.noaa.gov)

7 References


