

COLLABORATIVE APPROACH FOR TREND ANALYSIS USING CLUSTERING MECHANISMS AND BIG DATA TECHNOLOGIES

Shefali Arora, Ruchi Mittal and M.P.S Bhatia

Division of Computer Engineering, Netaji Subhas Institute of Technology, India

Abstract

The rapid growth in technologies and social media provides us the enormous amount of data, and it opens a wider window for researchers to work on such data. One of the critical analyses of the data is to check the changing trends in data. These days, massive volumes of data are being generated and processed using Hadoop and its ecosystem tools. These tools help in fast and efficient computing of a significant amount of data. In this paper, we collaborate few popular clustering algorithms with big data technologies to analyze the usage of mobile phones and networks in various locations. We loaded and processed this dataset in Apache Hive to examine the number of users and most prominent systems in given areas, based on their location codes. Further, we compared the time taken to build the clustered model on our framework to that on Weka tool. It was observed that Weka takes comparatively longer to process the dataset. This analysis would not only help in management and segregation of a considerable amount of data but would also help mobile service providers to understand the patterns of usage by customers and network problems, which may persist in some regions.

Keywords:

Big Data, Clustering Methods, Machine Learning, Hive

1. INTRODUCTION

Earlier, datasets contained only a limited amount of data. These days, the advent of social networking and digital transactions has made it possible to collect a vast amount of information. This data is difficult to analyze and manage using traditional processing tools. Big data is the term used to refer to this enormous amount of data obtained in various forms. This information is characterized by its volume, velocity, variety, and veracity, i.e., the 4Vs of big data. Big data gives us an opportunity to analyze and explore new insights from the available data. This data could come from many sources like social media, mobile phones, etc. Big data applications are making their way in a large number of industries: health, telecom, medical and various other sectors. Big data could be unstructured or structured, where structured data is simple and unstructured data does not have any predefined model [1]. Big data opens new avenues in the field of data mining and machine learning. As traditional techniques are not able to manage such huge amounts of data, new machine learning techniques and algorithms are being developed to generate new mechanisms and analysis techniques. Clustering is a remarkable machine learning technique required for analysis of big data. Clustering has been applied to various problems to discover trends in datasets [2]. Different clustering algorithms have been developed to divide data into useful groups and to identify new information among documents and web pages. K-means is a viral unsupervised learning technique in machine learning. It is used in various tools and technologies for analysis of big data. In this paper, we give an overview of different

clustering methods for big data analysis. Self-organizing maps and hierarchical clustering methods are some of the techniques that would be reviewed. Data is further analyzed using Apache Hive, which is famous tool of the Hadoop ecosystem, with its foundation on the concept of MapReduce [3].

The paper is divided as followed. Section 2 gives an overview of the related work and detailed review of the popular clustering techniques. Section 3 shows an excerpt of implementation based on K means clustering and analysis results in Hive. Section 4 concludes the paper.

2. RELATED WORK

2.1 CLUSTERING METHODS

Given a data set $X = \{X_1, X_2, \dots, X_m\}$ and integer k , clustering is used to map data points into k clusters. A data point belonging to one cluster is similar to other points in that cluster. Usually Euclidean distance is used as measure of similarity between clusters [4]. K means is a popular method of clustering. It is an unsupervised machine learning technique, where data is not labeled beforehand. Initially, clusters are assigned based on selection of centroids. Then new cluster centres are computed based on distance of data points from their nearest centroid [5]. The steps involved are as follows:

- N data points are separated into k disjoint subsets.
- Among these k sets, a centroid μ is randomly assigned to each set from among the data points
- Every data point is assigned a cluster which is nearest to the centroid.

The following formula is used to calculate the sum of squares and assign the appropriate cluster. This process is continued till there is no change in the process of cluster assignment.

$$\text{Arg min} \sum_{i=1}^k \sum_{x \in S} \|x - \mu_i\|^2 \quad (1)$$

Here x is the data point and μ is the centroid.

The mechanism of K means clustering has been demonstrated in the Fig.1.

Authors have described the working of Bitcoin in [6]. Unlike traditional payment systems dominated by US dollar, Bitcoin has its own metric value called bitcoin. A bitcoin's value is derived from its use for making payments in the Bitcoin system. Authors have questioned the economists on whether bitcoins meet standard attributes of money or not [7].

Researchers have proposed various extensions and modifications of K means clustering algorithm. Hartigan and Wong designed a K means algorithm which takes M points in N dimensions as input to divide it into clusters, such that sum of

squares value between clusters is minimized [6]. Further authors worked on better convergence of K means clustering algorithm in [7] to make the approach more efficient.

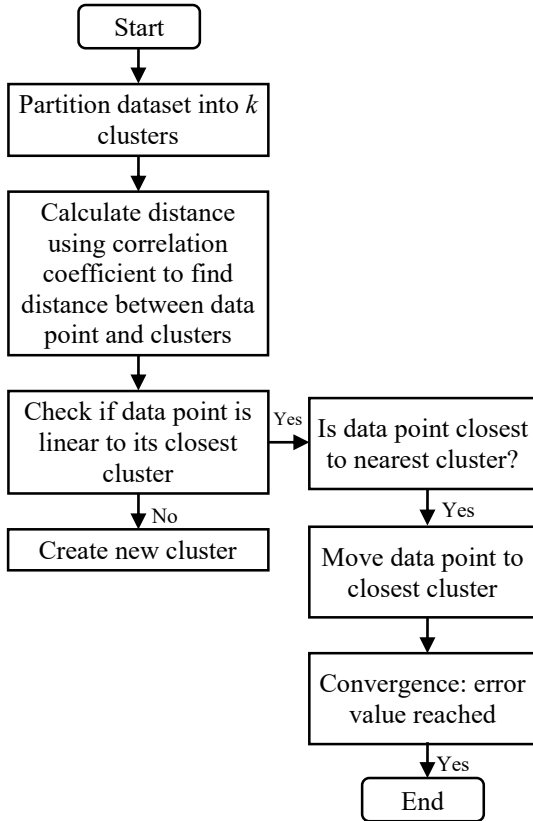


Fig.1. K means clustering mechanism

Self-organizing maps is a class of neural networks, which is another technique for unsupervised learning. SOMs are used for clustering of data and its visualization [8]. This is done by connection of nodes or neurons in a hexagonal or rectangular topology. Hence, rectangular SOMs are a two dimensional array of neurons i.e $N = \{N_{ij}, i = 1, 2, \dots, r, j = 1, 2, \dots, k\}$ where r is the number of rows and k is the number of columns. Given a set of input vector points $X = \{X_1, X_2, \dots, X_m\}$, distance between input vector N_{ij} and reference vector X_m is calculated. The closest reference vector is calculated based on the Euclidean distance between the two.

The learning rule is updated as follows:

$$N_{ij} = N_{ij}(m) + h_{ij}^m (X_m - N_{ij}^m) \quad (2)$$

Here h_{ij}^m is the neighborhood function and m is the number of iterations. This is repeated till the maximum numbers of iterations are achieved. SOM is used for solving many problems as it helps to reduce the dimensionality of the problem to determine the number of clusters. Its extensions also work with visualization tools. Therefore it is used for visualization of big data.

Hierarchical clustering is another method, which is used to build a cluster hierarchy of the dataset. These strategies could be:

- Agglomerative in which each data point is given a cluster and then pairs are merged.
- Divisive in which all data points are given a cluster and then these are recursively split.

To decide which clusters should be merged, the dissimilarity between data items is calculated. A dendrogram is used to present the results of hierarchical clustering. Authors have worked to improve hierarchical clustering in [9] which is further used to create partitions for evaluation of datasets. Expectation Minimization is another technique of clustering and used when data in question is incomplete. This algorithm helps to create the statistical model for optimization of data. It follows two steps: i) expectation and ii) minimization. Sara et al. [10] proposed a modified fuzzy K means algorithm with the use of expectation minimization to improve the shape and convergence of local maxima in clusters. Authors have given an advanced techniques of applying K means clustering algorithm by methods like expectation minimization, fuzzy K means and harmonic K means [12].

The rise in trend of machine learning algorithms has helped to solve a number of real world problems. With the increase in the amount of data, it is important to channelize the use of these algorithms in an efficient manner. Authors have applied machine learning to a large chunk of cellular data, which has various inherent patterns of usage. It is important to use the right set of algorithms and tools to obtain important information from this data. Analysis of cellular data using Hive helps to store and manage a large volume and variety of data. While machine learning helps to get important insights from this data for example, popular technologies among customers and network ranges in different regions of the world. The use of such techniques can help telecom companies to solve important problems like network congestion, signal problems etc.

3. APACHE HIVE

Apache Hive is an open source data warehousing framework which is based on Hadoop [11]. It supports SQL like declarative language in which queries are executed using the MapReduce algorithm of Hadoop. Hive is being used in various applications over thousands of jobs on a Hadoop cluster. These applications start from simple tasks to those related to machine learning and business intelligence.

- Built on the top of Hadoop, Apache Hive offers the following features:
- Enables tasks like extraction, transformation and loading for data warehousing
- Data analysis
- Reporting
- Easy access to data via SQL
- Execution of queries via MapReduce or Spark
- Easy access to files stored in HDFS or HBase, which are also parts of Hadoop ecosystem.
- Handles a vast variety of formats of data

Hive is best used to support jobs that require processing of large amounts of data. Hive also provides capabilities of serialization and de-serialization. It also provides Hive metastore to provide flexibility in designing schemas for any database.

In this paper, Hive queries have been used to analyze openCellid records to obtain insights on various features extracted from the dataset. The analysis is made on the basis of various location codes, mobile country codes, latitude and longitude of a

particular place. These features have been listed in the next section, along with the results of analysis.

4. DATA ANALYSIS

4.1 DATASET USED

The dataset used carried 880MB of data and collected from opencellid.org. This data consists of many attributes related to mobile networks, from which we extract the following features.

Table.1. Dataset used for Simulation

Feature	Detail
Radio	Could be GSM, UMTS (for 10k users)
	GSM, UMTS, LTE or CDMA (for 50k users)
MCC	Mobile country code
LAC	In the range of 1 to 65535
Cellid	Different ranges supported for GSM, CDMA
LON	Coordinates of longitude of country
LAT	Coordinates of latitude of country

4.2 EXPERIMENTS

Python has been used for implementation, using scikit-learn as the toolkit. The dataset is divided into two chunks: the first dataset has 10,000 records while the second dataset has 50,000 records. K means clustering is done based on the following parameters: Number of clusters - 3 and features for clustering - Radio, MCC, LAC, Cellid, LON and LAT. The values of these features are normalized to a common scale using the $(x - \text{standard deviation}) / \text{mean}$, where x is the feature value.

Calculation of Distance: Euclidean distance has been used to measure the distance of each feature value from the centroid value. Initial centroids have been designated randomly. Later on, with updation of clusters for every data point, these centroid positions are reassigned, followed by recalculation of Euclidean distance.

After applying k-means clustering algorithm, the clustered dataset is triggered in Apache Hive, and various features obtain insights of the dataset. The results of analysis have been shown for dataset with 10000 and 50000 records, based on the 3 clusters the dataset is divided.

Table.2. Number of mobile phone users in each cluster

Cluster	Count
0	42351
1	1645
2	5994

Table.3. Mean of clustered features on 10k records

	Cluster 0	Cluster 1	Cluster 2
Radio_GSM	1	0	0.761905
Radio_UMTS	0	1	0.23805
mcc	257.68181	256.24891	699.16666

lac	9736.1724	16981.97	3074.1071
lon	11.055883	9.339424	-62.257275
lat	50.203952	48.879107	-28.653642

Similarly, the values for 50k records for cluster 0, 1 and 2 are as shown in the Table.4.

Table.4. Mean of clustered features on 50k records

	Cluster 0	Cluster 1	Cluster 2
Radio_GSM	1	0.92462	0
Radio_UMTS	0	0.07538	0.67968
Radio_LTE	0	0	0.00337
Radio_CDMA	0	0	0.31694
mcc	238.776	534.465	255.905
lac	14132.73	14296.655	18172.757
lon	7.269221	88.5605	-30.535457
lat	49.073005	-5.033754	43.976964

4.3 ANALYSIS IN HIVE

We analyze the clustered dataset in Hive and determine which is the most popular mobile technology used. There are different cellular network standards on the basis of which this analysis is done: GSM, UMTS, LTE and CDMA. Analysis shows that most of the users make use of GSM while the least number of mobile phone users uses LTE. The grouping of these users has been done according to the three clusters is given in Table.5.

Table.5. Number of mobile phone users in each category

	Cluster 0	Cluster 1	Cluster 2
GSM	42351	0	1521
UMTS	4075	0	123
CDMA	0	1900	0
LTE	0	20	0

Using Hadoop's MapReduce technology, Hive is able to achieve this analysis within seconds. It is observed that Cluster 0 and 2 have the maximum number of people using GSM while Cluster 1 has the maximum number of users using CDMA. These patterns of usage among customers in various clusters explain the popularity of different radio systems in cell phones. Users in cluster 0 use Global System for Mobile Communication, which uses Time Division Multiple Access. TDMA works by assigning time slots to conversation streams and switching between each conversation. SIM cards are the most important component of these radio systems. Whereas users in Cluster 1 prefer using CDMA or Code Division Multiple Access. CDMA allows more users to connect at a given time, as compared to GSM. These systems do not use SIM cards and each phone is specifically made to work on a carrier's network. It is observed that many users are switching to UMTS. Universal Mobile Telecommunications System is another network based on the standards of GSM. Thus it is gaining popularity among users according to the analysis. LTE, the 4G mobile communications standard is the least popular technology according to this analysis. These patterns can be observed in Fig.2.

Based on this information, we perform another analysis on the clustered dataset using Hive. Using the mobile country code (mcc), we analyze the number of users in each region according to the three clusters. The Table.6 shows number of users per cluster for some of the mobile country codes.

Table.6. Users per cluster for some country codes

mcc	Cluster 1	Cluster 2	Cluster 3
202	92	120	0
208	5972	1475	0
232	1196	29	0
334	5	9	0
262	11325	689	0
244	888	36	0
404	25	40	97
732	0	0	46

This sample analysis shows that the maximum number of GSM and CDMA users lie in country codes 262 and 208. Cluster 2, which has the minimum number of users is sparsely scattered over existing country codes. Such kinds of analysis can be very useful in getting insights about the popularity of mobile phone networks among people.

4.4 COMPARISON WITH WEKA

Further, we clustered the dataset containing 50k records in Weka and obtained a count of different instances in two clusters. It is observed that the model is built in comparatively longer time as compared to that built using scikit-learn toolkit. The same dataset is used to build a model in Weka and analysis is done to obtain the following results. These have been tabulated in Table.7 - Table.9.

Table.7. Number of mobile phone users in each cluster in Weka

Cluster	Count of instances
0	6118
1	43872

Table.8. Mean of clustered features on 50k records in weka

	Cluster 0	Cluster 1
Radio_GSM	0	1
Radio_UMTS	0.6862	0
Radio_LTE	0.0004	0.00033
Radio_CDMA	0.3106	0
mcc	264.7865	248.5763
lac	18325.789	14106.117
lon	-30.7269	10.437
lat	42.8399	47.2172

Thus Weka is very slow to build our model as compared to Scikit that we used in this paper. Thus with the increase in amount of data and use of new technologies, we can analyze data in a very efficient manner. Further, with such a large amount of data

present in our dataset, it is difficult for traditional data processing tools to gain good insights after data analysis. Use of big data technologies like Apache Hive makes it possible to analyze such large amounts of data within seconds.

Table.9. Comparison of time taken to build the clustered models

Platform	Time to build model (in ms)
Scikit	3.10e ⁻⁰⁶
Weka	0.28

5. CONCLUSION AND FUTURE WORK

It is observed that traditional tools like Weka take a lot more time to build clustered models of datasets as compared to modern frameworks. Moreover, while Weka analyzes the given dataset over limited parameters, Hadoop ecosystem tools like Apache Hive give us an opportunity to handle a vast amount of data and analyze it using its fast query processing capability. Python frameworks are thus more efficient as compared to traditional machine learning tools like Weka. In this paper, we have also used Apache Hive to draw various conclusions from Opencellid dataset. This dataset consists of various features like latitude, longitude, location codes etc. Using these features, we can extract a lot of useful information and help in finding problems like network congestion, patterns in mobile phone usage etc. As a part of future work, we plan to analyze more such real world datasets. Also, we would extend the work to a number of applications by making use of modern technologies like neural networks, social network analysis etc.

REFERENCES

- [1] M. Chen, A. Ludwig and K. Li, "Clustering in Big Data", Available at: <https://pdfs.semanticscholar.org/2ab0/d4ded091959f0ed7140b85c90bef49d9ab1b.pdf>.
- [2] M. Hajeer and D. Dasgupta, "Handling Big Data using a Data-Aware HDFS and Evolutionary Clustering Technique", *IEEE Transactions on Big Data*, 2017.
- [3] A. Elsayed, O. Ismail and M. El-Sharkawi, "MapReduce: State-of-the-Art and Research Directions", *International Journal of Computer and Electrical Engineering*, Vol. 6, No. 1, pp. 34-39, 2014.
- [4] R. Loochach and K. Garg, "Effect of Distance on K-means Clustering Algorithm", *International Journal of Computer Applications*, Vol. 5, No. 2, pp. 7-9, 2012.
- [5] I. Foster, C. Kesselman, J. Nick and S. Tuecke, "The Physiology of the Grid: an Open Grid Services Architecture for Distributed Systems Integration", Technical Report, Department of Information Science, University of Southern California, 2002.
- [6] Aayushi Bindal and Anup Pathak, "Survey on K-means Clustering and Web-Text Mining", *International Journal of Science and Research*, Vol. 5, No. 4, pp. 1049-1052, 2016.
- [7] Min Huang, Lei Yu and Ying Chen, "Improved K-Means Clustering Center Selecting Algorithm", *Information Engineering and Applications*, pp. 373-379, 2012.

- [8] S. Seo and K. Obermayer, "Self-Organizing Maps and Clustering Methods for Matrix Data", *Neural Networks*, Vol. 17, No. 8-9, pp. 1211-1229, 2004.
- [9] Y. Rani and D. Rohil, "A Study of Hierarchical Clustering Algorithm", *International Journal of Information and Computation Technology*, Vol. 3, No. 10, pp. 1225-1232, 2013.
- [10] C. Selvi and E. Sivasankar, "A Novel Optimization Algorithm for Recommender System using Modified Fuzzy C-means Clustering Approach", *Soft Computing*, pp. 1-16, 2017.
- [11] S. Mehta and V. Mehta, "Hadoop Ecosystem: An Introduction", *International Journal of Science and Research*, Vol. 5, No. 6, pp. 557-562, 2017.
- [12] R. Yadav and A. Sharma, "Advanced Methods to Improve Performance of K-Means Algorithm: A Review", *Global Journal of Computer Science and Technology*, Vol. 12, No. 9, pp. 47-52, 2012.
- [13] R.V. Singh and M.P.S Bhatia, "Data Clustering with Modified K-means Algorithm", *Proceedings of International Conference on Recent Trends in Information Technology*, pp. 55-59, 2011.