

Opinion Dynamics based Affinity Propagation Clustering

Pankajdeep Kaur* and Awal Adesh Monga*

ABSTRACT

An opinion in the social society is made in order to make choices of making inter relations and communication among the humans. This capability of humans can be used as a clustering technique by which the data points can make their choices of making clusters by the process of opinions. These opinions are made by the process of message passing between the data points. This is named as opinion dynamics based affinity propagation. We have implemented this proposed algorithm on training dataset in order to test it and get the output. Performance of the algorithm is quite impressive and hence the scalability is also measured. This paper includes the implementation and results produced on the dataset.

Keywords: Big Data, Cluster analysis, CODO, Affinity based Propagation, Opinion Dynamics

I. INTRODUCTION

Big Data is the collection of all structured and unstructured data. It consists of all types of data which needs to be managed properly and efficiently so that the future requirements of the user can be easily met. This is done basically to increase the market value of data and also the reputation of the organization in order to survive in the world of increasing competition. For this survival organizations need to have management of the huge data called as the Big Data Management (BDM). BDM includes four phases in it -Big Data Generation, Big Data Acquisition, Big Data Storage, and Big Data Analysis.

1.1 Cluster Analysis

Cluster analysis is a multiple variable method which goals to differentiate a sample of subjects on the basis of a set of calculated variables into a number of different groups such that all the similar subjects are placed in the same group [2]. An example which goes with the clustering is in the field of psychiatry, where the classification of patients on the basis of clusters or symptoms can be helpful in the identification of an appropriate form of therapy. In marketing, it can be useful to find distinct groups of customers so that advertising can be efficiently targeted.

A cluster is a subset of subjects which are “similar or correlated”[3]. A subset of objects so that the Eucladian distance between any two subjects in the cluster is less than the predicted distance or Eucladian distance between any data point or subject in the cluster. The clustering can be represented as a set of subsets

$$C = C_1, \dots, C_k \text{ of } S$$

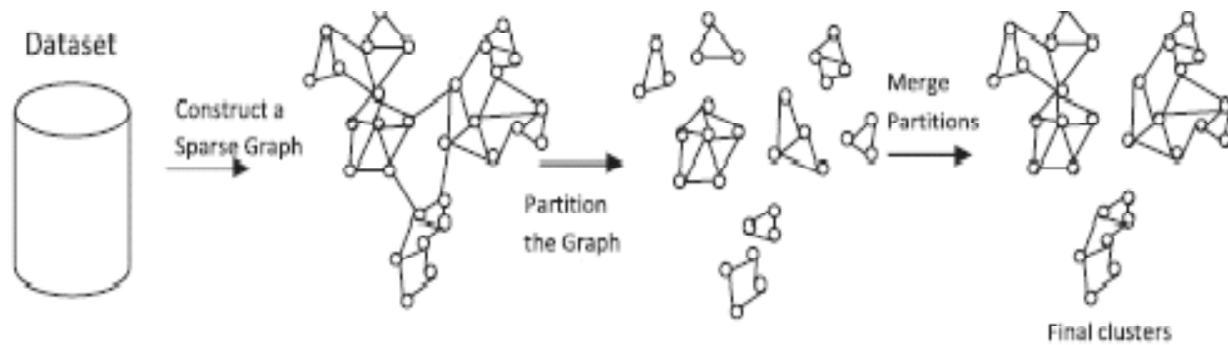
such that:

$$S = \bigcup_{i=1}^k C_i \text{ and}$$

* Department of Computer Science, Guru Nanak Dev University, RC, Jalandhar, E-mail: pankajdeepkaur@gmail.com; awaladesh@gmail.com

$$C_i \cup C_j = \emptyset \text{ for } i \neq j$$

And any instance of S' belongs to only one subset [1].



Clustering [4]

1.1.a Importance of Cluster Analysis

- Accelerating the growth of data volumes.
- Improving market values of the product.
- Improving the probability of fraud detection.
- Improving the reliability of data.
- Helps in the process of reusing the data easily.

1.1.b Main two types of Clustering Techniques

- **Hierarchical clustering** – Connection based clustering is known as *hierarchical clustering*. It is based on the idea of subjects which are more related to nearby subjects than to subjects which are farther. This clustering connects data points or objects to form inter related intra related clusters on the basis of some distance. A cluster can be defined mainly and probably by the maximum distance required to connect parts of the same.
- **k-means clustering** – In centroid-based clustering, clusters are referred by a central vector, which may or may not be a member of the data set. When the number of clusters is fixed to some number k , k -means clustering gives a definition as an optimization problem.

In this paper we will see how big data can be analysed. Different organizations have proposed their own technologies for analysing Big Data. These technologies were developed according to the requirements of the organizations. But the basis was to analyse the Big data. A brief introduction to the Big Data Management Big Data Analysis is given in section I. Section II deals with the basics of Affinity based analysis of Big Data and discussion regarding the opinion dynamics based affinity propagation. Section III deals with the experimentation and the results obtained by implementing this on the diabetes dataset. In section IV is the conclusion part of the paper where we have discussed the results and the efficiency of our proposed algorithm.

II. PROPOSED ALGORITHM

2.1. Affinity based propagation

Affinity propagation (AP) is a method of clustering based on the concept of message passing between data points or communication among the data points or between clusters. The main advantage of applying this method on dataset is that we need not to define the number of clusters before running the algorithm.

Let x_1 through x_n be a set of data points, with no assumptions or exceptions made about their internal structure, and let s be a function that shows or quantifies the similarity between any two points, such that $s(x_i, x_j) > s(x_i, x_k)$ iff x_i is more similar or more compatible to x_j than to x_k .

The algorithm processes by communicating two alternate message passing steps, to update two matrices:

- First one is responsibility matrix \mathbf{R} has values $r(i, k)$ that tells or quantify how well compatible is x_k to serve as the similar or exemplar for x_i , in comparison to the other candidate exemplars for x_i .
- Second is the availability matrix \mathbf{A} which contains values $a(i, k)$ that is used to show how appropriate and best it would be for x_i to pick x_k as its similar or exemplar, taking into account other points' of preference for x_k as an exemplar.

Both matrices are started and initialized to all zeroes, and can be seen as log-probability tables. The algorithm then proceeds by following the given updates iteratively or repeatedly:

- First one to set and update around responsibility matrix[8]:

$$r(i, k) \leftarrow s(i, k) - \max_{k' \neq k} \{a(i, k') + s(i, k')\}$$

- And then the updating process goes to the availability matrix[8]

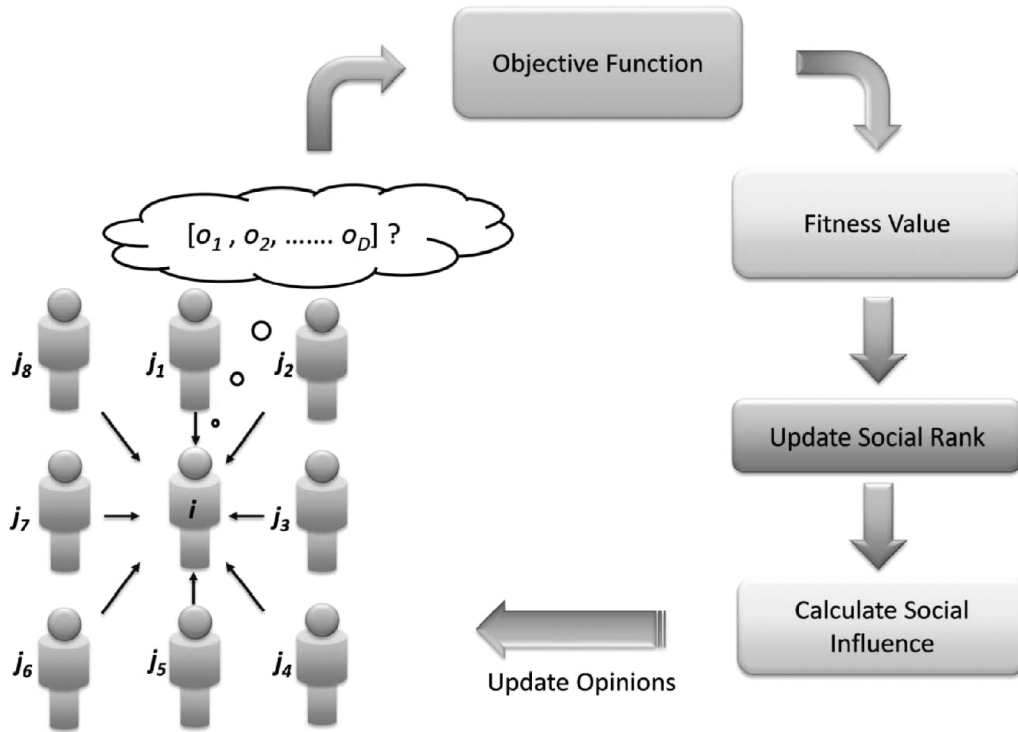
$$a(i, k) \leftarrow \min \left(0, r(k, k) + \sum_{i' \neq (i, k)} \max(0, r(i', k)) \right) \text{ for } i \neq k$$

$$a(k, k) \leftarrow \sum_{i \neq k} \max(0, r(i, k))$$

There are steps and behavior which shows that human opinion formation and their interaction process or dynamics can be used to solve complex and huge mathematics related problems. Although if we consider the social and psychiatric point of view these opinion formation steps or opinion dynamics models are very naive and limited in their procedure but in the past research, researchers felt that their research will have a path towards developing new methods, models and tools to understand the real problem solving capability of human beings in a social influence or structure. With this developing idea some of the researchers developed a new algorithm called as CODO (Continuous Opinion Dynamics Optimizer) algorithm. This algorithm was based on the opinions of the clusters and thoroughly their choice of joining the cluster according to their own capability. Further, the effect of these social influence or structure can also be visualized and studied viz. scale free and random graphs. They have also shown the effect of adaptive noise on the formation of clusters in their proposed algorithm and compared the complete performance with a lbest PSO (Particle Swarm Optimization). It is very important to note that their proposed algorithm has only and only a single control parameter (S) unlike the other optimizers which have various controlling parameters. Although, the algorithm in that form was very basic and does not perform quite well in the case of highly multi variate and multimodal problems, but it can be studied, analyzed and made to improve its performance by developing strategies or methods to dynamically update S parameter so that it could attain finer grained search.

2.2. Opinion Dynamic based Affinity Propagation

Due to the increasing complexity of data and associating the analysis of these complex problems with the human behaviour is an interesting and full of exciting area. Lot of theories have been given by various researchers depicting the human bevaieur in the mathematical problems. HOD(Human Opinion Dynamics) is one of the research area which has been growing in order to solve the optimization problems. The main roots of this method lies in Social Impact Theory Optimisation (SITO), because the researchers found that this has not full compatibility and utility to complex and huge problems and also these are based on the



CODO schema [9]

different opinion formation. Thus HOD model became the basis for Continuous Opinion Dynamics Optimizer (CODO) in order to solve its utility issue. Thus the proposed model is based on the procedure of opinion formation of a group of data points or individuals during an iteration and has four important and primitive fundamentals- social arrangements, point view area, social impact and restore order. Social impact and arrangement acts as the are due to which different data point have different interaction capability with others whereas each data point is placed on the node of graph. A notable and an important difference between HOD based optimization from opinion dynamics is that, as in opinion space, collision is possible, i.e. two data points or individuals can have same opinion at the same time while two data points or insects possibly cannot have the same position in the swarm at the same time. Opinions are taken to be continuous here in order to solve our problem of optimization where the parameters of optimizing can have any value within a finite range. Affinity or opinions are influenced by the opinions of its neighbouring data points depending on their social interaction and influence which is defined here as the ratio of social rank of any individual to the distance between them and is given by:

$$w_{ij} = \frac{SR_j(t)}{d_{ij}(t)}$$

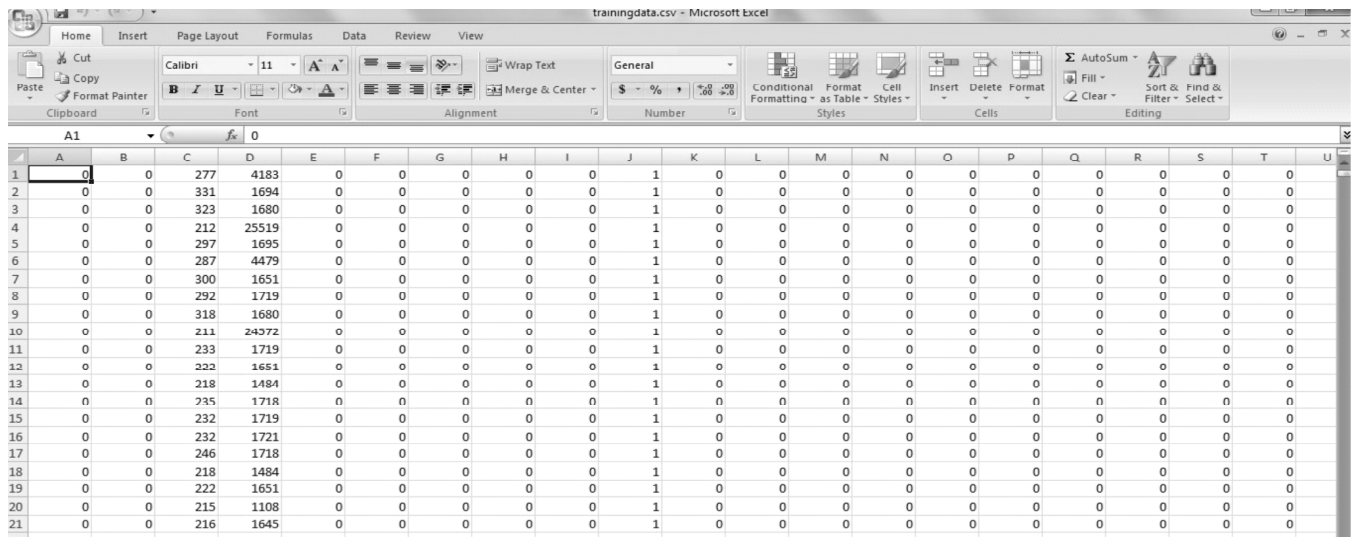
Here, SR (Social Rank) is determined by the inverse of the fitness value of the data point, where fitness value is the error which needs to be minimized. The modification adapted here from the works of Reshamjit et. al. are inspired from the fact that the problem at hand has to predict the value which is a continuous variable and the concept of social score in place of social rank ensured that the difference in individual's fitness is well represented during the opinion update rule. Each individual's opinion is updated by the following rule given as:

$$\Delta o_i = \frac{\sum_{j=1}^N (o_j(t) - o_i(t)w_{ij}(t))}{\sum_{j=1}^N w_{ij}(t)} + \epsilon_i(t), j \neq i$$

Where $o_j(t)$ is the opinion of neighbours of individual i , W_{ij} is the social influence factor, and η is adaptive noise introduced to justify individualization in society after a certain consensus limit is reached.

III. EXPERIMENT AND COMPARISON OF RESULTS

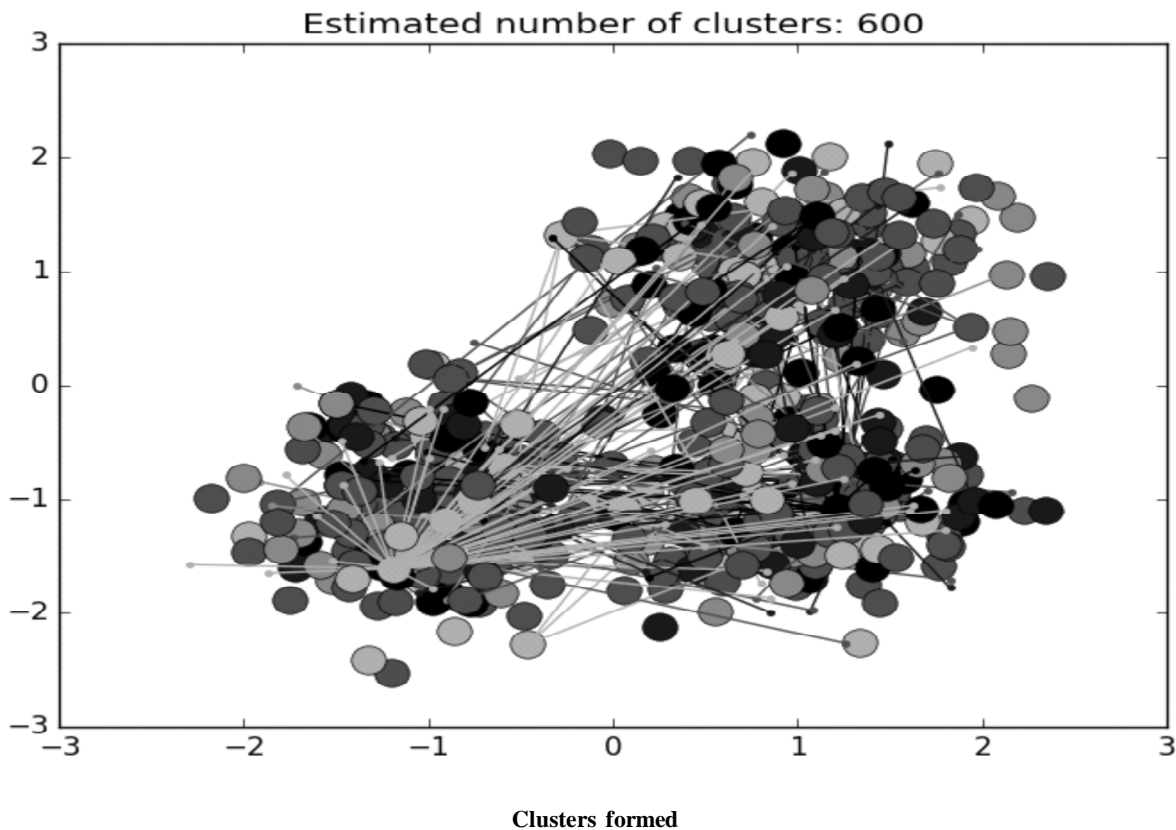
For this research work I have taken training Dataset. This dataset has total of 4391 instances 40 columns having values correspondingly.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	0	0	277	4183	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
2	0	0	331	1694	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
3	0	0	323	1680	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
4	0	0	212	25519	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
5	0	0	297	1695	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
6	0	0	287	4479	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
7	0	0	300	1651	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
8	0	0	292	1719	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
9	0	0	318	1680	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
10	0	0	211	24372	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
11	0	0	233	1719	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
12	0	0	222	1651	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
13	0	0	218	1484	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
14	0	0	235	1718	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
15	0	0	232	1719	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
16	0	0	232	1721	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
17	0	0	246	1718	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
18	0	0	218	1484	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
19	0	0	222	1651	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
20	0	0	215	1108	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
21	0	0	216	1645	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0

Training Dataset of used for implementation

The proposed scheme is tested using Python 3.4 environment. From the simulation of the experiment results obtained are as given,



```

>>> ----- RESTART -----
>>>
[[-1.02873476 -1.19560853]
 [ 1.46327915 -0.72095906]
 [ 0.79938265 -1.40004124]
 .....
 [ 1.36582946 -1.03274419]
 [-0.88630361 -1.50836932]
 [ 0.48503236 0.82502832]]
Estimated number of clusters: 600
Homogeneity: 0.740
Completeness: 0.135
V-measure: 0.228
Adjusted Rand Index: -0.000
Adjusted Mutual Information: -0.002
Silhouette Coefficient: 0.337
>>> ----- RESTART -----
>>>
[[-1.02873476 -1.19560853]
 [ 1.46327915 -0.72095906]
 [ 0.79938265 -1.40004124]
 .....
 [ 1.36582946 -1.03274419]
 [-0.88630361 -1.50836932]
 [ 0.48503236 0.82502832]]
Estimated number of clusters: 600
Homogeneity: 0.740
Completeness: 0.135
V-measure: 0.228
Adjusted Rand Index: -0.000
Adjusted Mutual Information: -0.002
Silhouette Coefficient: 0.337
>>> |
    
```

Other parameters of Opinion based Affinity Propagation

Precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percentage.

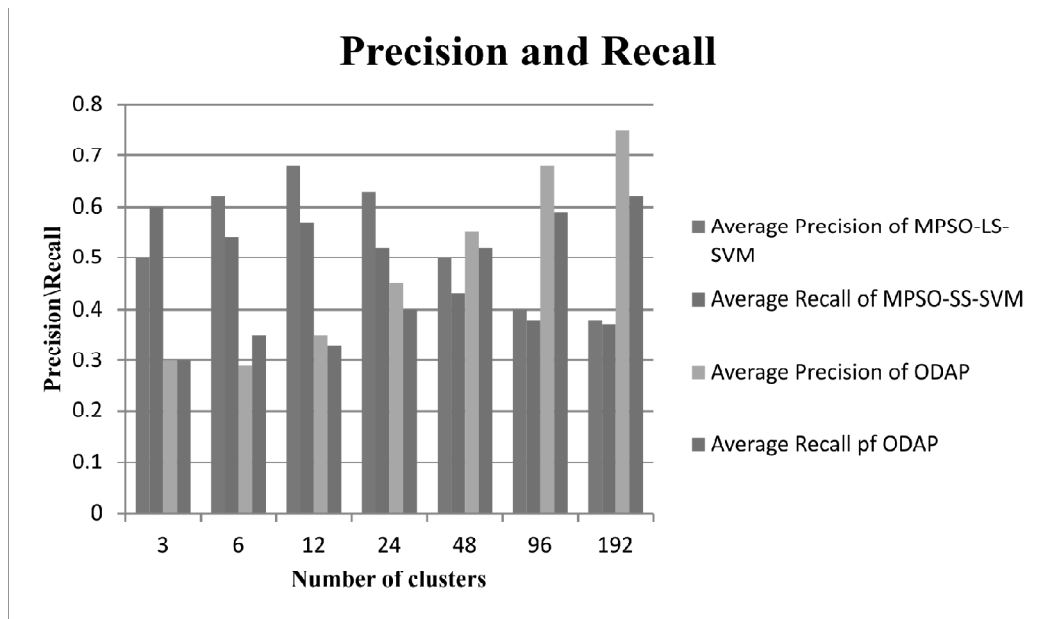
$$Precision = \frac{tp}{tp + fp}$$

Recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually expressed as a percentage.

$$Recall = \frac{tp}{tp + fn}$$

MPSO-LS-SVM is Modified Particle Swarm Optimization with the Least Square Support Vector Machine algorithm.

ODAP is Opinion Dynamic based Affinity Propagation algorithm.

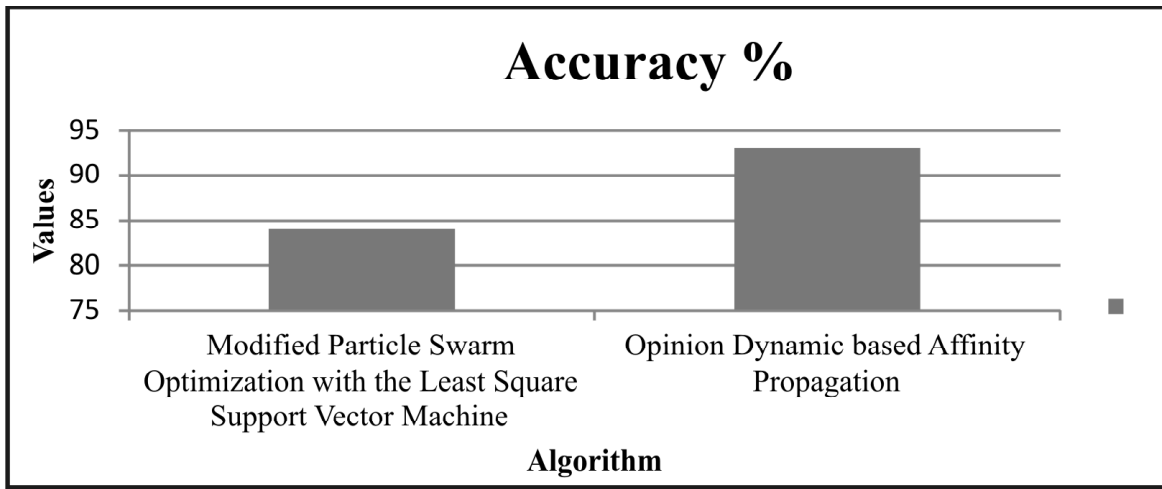


Comparing precision and recall

The above chart clearly shows that the results obtained from the proposed algorithm are far better than the existing algorithm. These results need to be verified in the terms of accuracy.

Accuracy in simple terms is the measurement of the systematic errors. But in case of algorithms accuracy can be defined as the trueness of the algorithm such that the results obtained from the algorithm are close to the true nature as if they are more close to the accurate. Accuracy can't be 100% as some errors always occur in the designing of the algorithm. The actual work done is to make the accuracy near to cent percent.

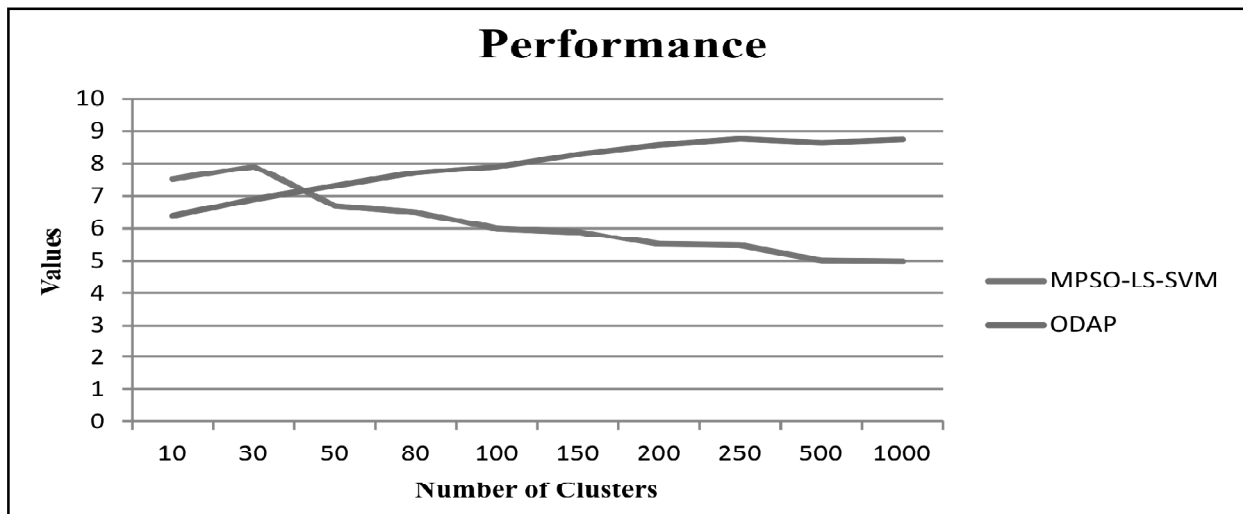
$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$



Comparison of Accuracy

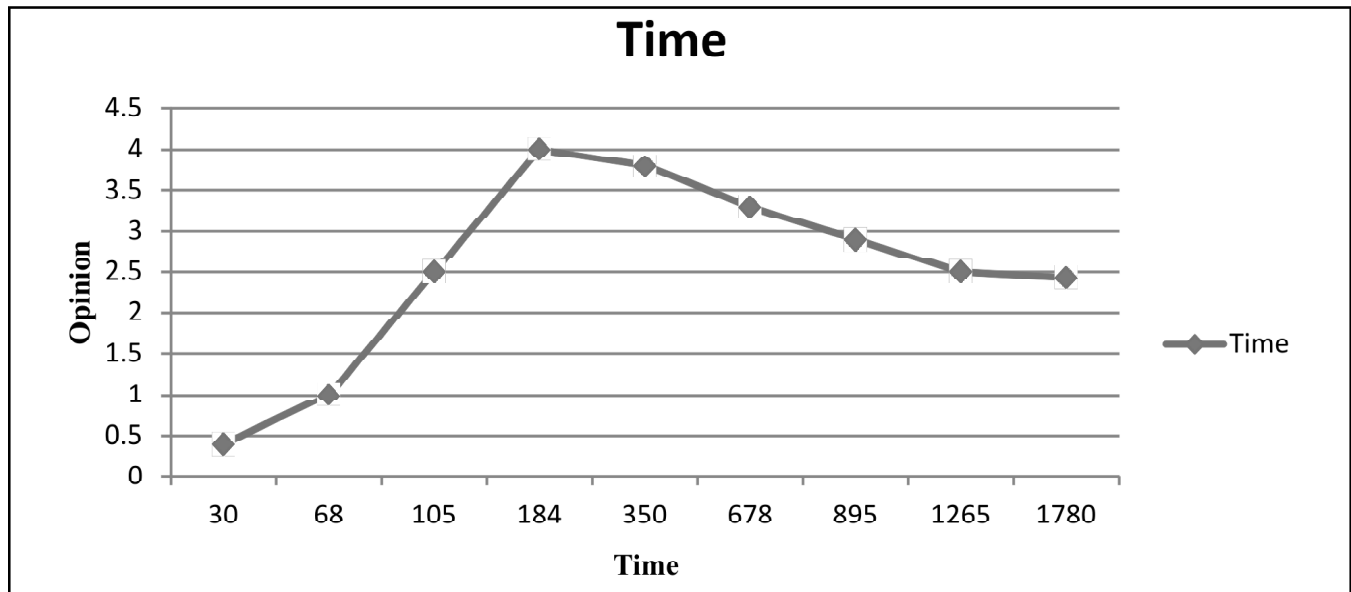
This chart clearly shows that the ODAP algorithm gives more accurate results such that the clusters formed by ODAP algorithm are more precise and accurate than the other.

Performance of the algorithm is capability of the algorithm with which it performs such that using the algorithm gives faster and most reliable results. Performance of the affinity based algorithm depends mainly on 6 factors. Their values determine how much efficient and capable the algorithm is. The capability of an algorithm does not depend on a single parameter. It depends on various other parameters which define the quality of the algorithm and the efficiency of the algorithm.



Performances of the Algorithms

Time is the most crucial factor in case of algorithms. Every time an algorithm is approached the time taken by it for execution is considered at first. So designing an algorithm which takes minimum time for execution even if the data is increasing, is a challenging task. The time consumption the proposed algorithm is shown in the chart below.



Time consumption of proposed ODAP

IV. CONCLUSION

The results obtained after running the program are more surprising than interesting. This algorithm was basically proposed in order to increase the performance of the Affinity Propagation by making use of Hierarchical Technique and thus introducing the Opinion Dynamic based Affinity Propagation Algorithm. The overall objective of the research work has been completed in this file. The results obtained are satisfactory at performance point of view but if we got to see on the basis of time consumption, then ODAP algorithm is not preferred. Although the results are good but the time factor decreases the favoring condition for the algorithm. In case of Big Data Analysis time is essential due to the complexity of data. The basis reason behind the lagging of algorithm is the complexity and variety of data in the given dataset.

1. Effectiveness and efficiency- According to the results obtained from the implementation. It can be clearly seen that Opinion Dynamic based Affinity Propagation is far efficient than existing technique for the analysis of large and huge Dataset.
2. Scalability – According to the results obtained we found that the running time at each and every iteration is increasing proportionally with the number of similarities, but the number of iterations may not seem be the same for all and different data size.

REFERENCES

- [1] Lior Rokach, Oded Maimon, "CLUSTERING METHODS", 2003.
- [2] Rosie Cornish, "Cluster Analysis", 2007.
- [3] Jerzy Stefanowski, "Data Mining - Clustering", Poznan University of Technology, 2009.
- [4] Ari Wibisono, Wisnu Jatmiko, Hanief Arief Wisesa, Benny Hardjono, Petrus Mursanto, "TrafficbigdatapredictionandvisualizationusingFastIncrementalModelTrees-DriftDetection(FIMT-DD)", Science Direct, 2015.

- [5] https://en.wikipedia.org/wiki/Cluster_analysis.
- [6] https://en.wikipedia.org/wiki/Hierarchical_clustering.
- [7] <https://www.google.co.in/imgres?imgurl=http%3A%2F%2Fimage.slidesharecdn.com%2Ffcc-141229090756-conversion-gate01%2F95%2Fa-clusteringbased-approach-to-detect-probable-outcomes-of-lawsuits-13-638.jpg%3Fcb%3D1419844532&imgrefurl=http%3A%2F%2Fwww.slideshare.net%2Fdanielgribel%2Fa-clusteringbased-approach-to-detect-probable-outcomes-of-lawsuits&docid=XwQBM3MkeRpVCM&tbnid=JUIJ6PJTh5ZJM%3A&w=638&h=359&bih=643&biw=1366&ved=0ahUKEwiMupmih7NAhWJK48KHcYKDvIQMwhDKBswGw&iact=src&uact=8>.
- [8] https://en.wikipedia.org/wiki/Affinity_propagation.
- [9] Rishemjit Kaur, Ritesh Kumar, Amol P. Bhondekar, Pawan Kapur, “Human opinion dynamics: An inspiration to solve complex optimization problems”, Scientific Reports, 2013.
- [10] K.Vembandasamy, T.Karthikeyan, “Novel Outlier Detection In Diabetics Classification Using Data Mining Techniques”, International Journal of Applied Engineering Research Volume 11, Number 2, 2016.
- [11] Guazzini A, Cini A, Bagnoli F and Ramasco JJ (2015) Opinion dynamics within a virtual small group: the stubbornness effect. *Front. Phys.* 3:65. doi: 10.3389/fphy.2015.00065.
- [12] Cini A, Guazzini A. Human virtual communities: affinity and communication dynamics. *Adv Complex Sci.* (2013) 16:1350034-1–24. doi: 10.1142/S0219525913500343.
- [13] Athman Bouguettaya, Qi Yu, Xumin Liu, Xiangmin Zhou, Andy Song, “Efficient agglomerative hierarchical clustering”, Science Direct, 2014.
- [14] Xiaojin Zhu, “Clustering”, Advanced Natural Language Processing, 2010.
- [15] Chanchal Yadav, Shuliang Wang, Manoj Kumar, “Algorithms and Approaches to handle large Data – A Survey”, IJCSN International Journal of Computer Science and Network, Vol 2, Issue 3, 2013.
- [16] Patrick Redmond, Prof. John A. Trono, Dave Kronenberg, “Affinity Propagation, and other Data Clustering Techniques”, White Paper, 2012.
- [17] Michael Biddick, “The Big Data Management Challenge”, InformationWeeks, April 2012.
- [18] George Gilbert, “A Guide To Big Data Workload Management Challenges”, GigaOM PRO, May 2012.
- [19] Jinchuan CHEN, Yueguo CHEN, Xiaoyong DU, Cuiping LI, Jiaheng LU, Suyun ZHAO, Xuan ZHOU, “Big data challenge: a data management perspective”, Higher Education Press and Springer-Verlag Berlin Heidelberg 2013.
- [20] Hongtaek Ju, Choong Seon Hong, Makoto Takano, Jae-Hyoung Yoo, Kuang-Yao Chang, Kiyohito Yoshihara, Jee-Yih Jeng, “Management in the Big Data & IoT Era: A Report on APNOMS 2012”, Springer Science+Business Media New York 2013, 22 february 2013.
- [21] Pankaj Deep Kaur and Awal Adesh Monga, “BIG DATA MANAGEMNENT”, International Journal of Advance Foundation And Research In Science & Engineering (IJAFRSE) Volume 1, Special Issue, ICCICT 2015, August 2015.
- [22] Pankaj Deep Kaur, Awal Adesh Monga, “MANAGING BIG DATA: a step towards huge data security”, International Journal of Wireless and Microwave Technology (IJWMT), Volume 6, Issue 2, pp.10-20, March 2016.