Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 12, December-2019

Clustering of Data using Affinity Algorithm

Rachana B S
Asst. Professor, Dept..of CSE
PES University, R R Campus
Bangalore, India

Pushpa G
Asst. Professor, Dept..of CSE
PES University, R R Campus
Bangalore, India

Abstract:- **The term Big Data is used for denoting the collection of datasets that are extremely large and complex making it difficult to process using traditional data processing applications. The datasets clustering has become a challenging issue in the field of big data. The most widely used procedure to identify clusters is known as k-means. The k-means algorithm finds clusters with the least inertia for a given k. A drawback of this k-means is that if k is not known. This paper presented a new algorithm called affinity propagation which is based on the passing of the message between data points. The number of clusters to be determined or estimated before running the algorithm is not required in this proposed affinity algorithm.**

Keywords:- K-means clustering, exemplars, responsibility, availability.

I.INTRODUCTION

Data mining is a step in the Knowledge Discovery in Databases (KDD) process consisting of the application of data analysis and discovery of algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data[1]

A cluster is a collection of the data object that is similar to one another are in the same cluster and dissimilar to the objects are in other clusters. Organizing this increasing data and learning valuable information from data makes clustering techniques to be used in many research and application areas such as artificial intelligence, biology and so on. Cluster analysis is a tool that is used to observe the characteristics of the cluster and to focus on a particular cluster for further analysis. Clustering is an unsupervised problem[2] and it deals with finding a structure in the collection of unlabeled data.

k-means clustering is one of the most popular algorithms for cluster analysis in data mining[3]. K-means is an unsupervised iterative algorithm that is very simple and very fast, so in many practical applications, the K-means method is proved to be an effective way that produces good clustering results. But it is very suitable if k value is known
in prior.The k-means algorithm finds clusters with the least inertia for a given k . A drawback of this k-means is that if k is not known. This paper presented a new algorithm called affinity propagation which is based on the passing of the message between data points. The number of clusters to be determined or estimated before running the algorithm is not required in this proposed affinity algorithm[4].

The rest of this paper is organized as follows: Section 2 provides background and related work, Section 3 presents the proposed algorithm for clustering, Section 4 presents results, Section 5 presents the conclusion.

II. BACKGROUND AND RELATED WORK

The k-means term was first coined used by James MacQueen in 1967. The standard k-means algorithm [5, 6] is an iterative refinement approach that minimizes the sum of squared distances
between each point and its assigned cluster center.

K-Means algorithm is unsupervised which is used in data mining and pattern recognition. Aiming at minimizing cluster performance index, square-error and error criterion.

Algorithm:
K-Means clustering:
1.choose k centers randomly, where the value k is fixed in advance.
2.Repeat the step1.
3.select each point to the nearest cluster center.
4. Euclidean distance is used to determine the distance between each data object and the cluster centers
5.Recompute the cluster centers of each cluster
6.Iterative this process repeatedly until the criterion function becomes the minimum.

In [8] approximate nearest neighbor search is used instead of the exact nearest neighbor search in the assignment step for each point. This approach was further improved in [9] in terms of convergence speed.

In the literature [10], it proposed a systematic method to find the initial cluster centers, This centers obtained by this method are consistent with the distribution of data. This method will produce more accurate clustering results than the standard k-means algorithm.

III. PROPOSED ALGORITHM

Let { x_1..... x_n }be a set of data points, with no assumptions and let *s* be a function that quantifies the similarity between any two points, such that $s(x_i, x_j) > s(x_i, x_k)$.
The diagonal of S represents the input preference, which means how particular input is to become an exemplar. When S is set to the same value for all the inputs it controls how many classes the algorithm produces.
Fewer classes are produced if the value is close to the minimum possible similarity. Many classes are produced if the value is larger than the maximum possible similarity. It is

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 12, December-2019

initially set to the median similarity of all pairs of inputs. To update two matrices the algorithm proceeds by alternating the message passing.

- The responsibility matrix R. In this matrix, r(i,k) reflects point k is to be an exemplar for point i.
- The availability matrix A. In this matric a(i,k) gives how appropriate it would be for point i to choose point k as its exemplar.

Both matrices are initialized to all zeroes. Then performs the following updates iteratively:

Responsibility is given by:

$$r(i,k) \leftarrow s(i,k) - \max_{k' \neq k} \{ a(i,k') + s(i,k') \} \quad (1)$$

Availability is given by two formulas:

If the points not on the diagonal of S

$$a(i,k) \leftarrow \min \{ 0, r(k,k) + \sum_{i' s.t. i' \notin \{i,k\}} \max(0, r(i',k)) \} \quad (2)$$

If the points is on the diagonal of S

$$a(k,k) \leftarrow \sum_{i' \neq k} \max(0, r(i',k)) \quad (3)$$

The iterations have to be continued until either the cluster boundaries remain unchanged over several iterations or after some predetermined number of iterations.

Similarity function is the negative euclidian distance squared

$$s(i,k) = -\|x_i - x_k\|^2 \quad (4)$$

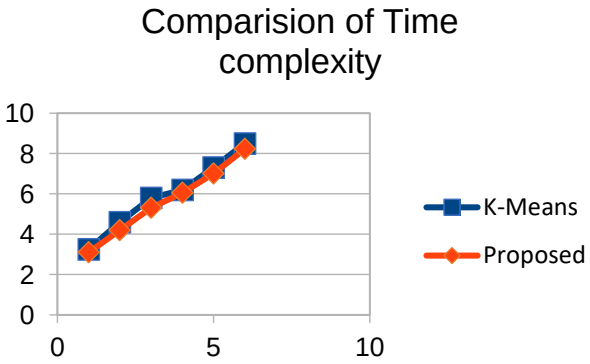
The exemplars are getting from the two final matrices where responsibility + availability' is positive.

IV.RESULTS

We implemented our proposed algorithm in java platform.

Table 1: comparision of time complexity

Algorithm Name	Worst case	Average case	Best case
K-means algorithm	$O(n^i)$ 2<=i<3	$O(n^2)$	O(n)
Proposed algorithm	$O(n^2)$	O(n)	O(n)



V.CONCLUSION

The datasets clustering has become a challenging issue in the field of big data. The most widely used procedure to identify clusters is known as K-Means. The k-means algorithm finds clusters with the least inertia for a given k. A drawback of this k-means is that if k is not known. This paper presented a new algorithm called affinity propagation which is based on the passing of the message between data points. The number of clusters to be determined or estimated before running the algorithm is not required in this proposed affinity algorithm.

REFERENCES

[1] Bhagyashri S. Gandhi, and Leena A. Deshpande, "The survey on approaches to efficient clustering and classification analysis of big data", International Journal of Engineering Trends and Technology (IJETT), vol. 36, no. 1, pp. 33-39, 2016.

[2] BianZhaoQi and ZhangXuegong Pattern Recognition Beijing Tsinghua University Press 2000.

[3] Zhai, D.; Yu, J.; Gao, F.; Lei, Y.; Feng, D. K-means text clustering algorithm based on center selection according to maximum distance. Appl. Res. Comput. 2014, 31, 713–719.

[4] *Delbert Dueck, Frey:"Clustering by passing messages between data points".5814):"(2007).*

[5] Forgy, E. W. "Cluster analysis of multivariate data: Efficiency versus interpretability of classifications". Biometrics, 21:768–780, 1965

[6] MacQueen, J.B. "Some methods for classification and analysis of multivariate observations". In Proc. of 5th Berkeley Symposium on Mathematical Statistics and Probability, volume 1,1967.

[7] Wang, Q.; Wang, C.; Feng, Z.; Ye, J. "Review of K-means clustering algorithm". Electron. Des. Eng. 2012, 20, 21–24.

[8] Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. "Object retrieval with large vocabularies and fast spatial matching. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)",2007.

[9] Philbin, J. "Scalable Object Retrieval in Very Large Image Collections". Ph.D. thesis, University of Oxford, 2010.

[10] *F Yuan, Z. H Meng, H. X Zhang, C. R Dong, "A New Algorithm to Get the Initial Centroids", Proc. of the 3rd International Conference on Machine Learning and Cybernetics, pp. 26-29, 2004.*

[11] *Renchu Guan; Xiaohu Shi; Maurizio Marchese; Chen" Yang; Yanchun Liang (2011). "Text Clustering with Seeds Affinity Propagation". IEEE Transactions on Knowledge & Data Engineering. **23** (4): 627–637.*

[12] Welling M, Kurihara K. Bayesian K-means as a "Maximization-Expectation" Algorithm. In: SDM. SIAM; 2006. p. 474–478.

[13] Teh YW. Dirichlet process. In: Encyclopedia of Machine Learning. Springer, US; 2010. p. 280–287.

Rachana B S
Asst.Professor,Dept..of CSE,
PES University,RR campus,Bangalore

Pushpa G
Asst.Professor,Dept..of CSE
PES University,RR campus,Bangalore