IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-10, May 2019 E-ISSN: 2347-2693

K-MEAN++ Applied To Solve Problems of Data Security in Data Science

Sarita Patel[1*], Atul Garg[2], Vandana Tripathi[3]

[1]Department of Information Technology, Hitkarini College of Engineering and Technology, RGPV Bhopal, Jabalpur, India
[2,3]Departments of Electronics and Communication, RGPV Bhopal, Jabalpur, India

[*]*Corresponding Author: spatelvijay@gmail.com, Tel.: 9827391180,9753711292*

DOI: https://doi.org/10.26438/ijcse/v7si10.122126 | Available online at: www.ijcseonline.org

***Abstract*—** In this paper, we describe an application K-MEAN++ clustering algorithm and Data Encryption Standard algorithm for security of information and large volumes data. Data are highly complex multidimensional signals, with rich and complicated information content Data science. For this reason they are difficult to analyze through a unique automated approach. However a K-MEAN++ scheme & Data Encryption Standard are helpful for the understanding of security of data content in Data science. In any system that captures, stores, analyzes, manages, and presents data that are linked to location and like Image satellite sensors acquire huge volumes of imagery to be processed and stored in big archives. Technically, a data science is a data modelling that includes mapping software and its application to data set , land surveying, aerial photography, mathematics, geography, and tools that can be implemented with Data science software Building a hierarchy is a fruitful area if one likes the challenge of having difficult technical problems to solve. Some problems have been solved in other technologies such as database management. However, Data science throws up new demands, therefore requiring new solutions. In this paper we have examine difficult problems, and to be solved and gives some security methods to solve the problem of data security using clustering algorithm.

***Keywords*—** *K-MEAN++, Data science, Security Data Encryption Standard algorithm.*

I. INTRODUCTION

The purpose of this paper is to present K-MEAN++ clustering algorithm for choosing the initial values for the k-means clustering algorithm. K-MEAN++ **is** a way of avoiding the poor clustering found by the k-means algorithm. The k-means problem is to find cluster centers that minimize the sum of squared distances from each data set being clustered to its cluster center that is closest to it. Although finding an exact solution to the k-means problem for arbitrary input, using this approach we find an approximate solution is used to finds reasonable solutions quickly.

However, the k-mean algorithm has two major theoretic limitations:

- First, it has been shown that the worst case running time of the algorithm is super-polynomial in the input size.
- Second, the approximation found can be arbitrarily bad with respect to the objective function compared to the optimal clustering.
- In K-MEAN++ clustering algorithm addresses the second of these obstacles by specifying a procedure to initialize the cluster centres before proceeding with the standard k-means optimization iterations. With the K-MEAN++ initialization, the algorithm is guaranteed to find a solution that is O (log k) competitive to the optimal k-means solution.

Our goal is to use this algorithm to segment or classification of data in an automated fashion and supply the points and the number of clusters you expect to get, and the algorithm returns the same points, organized into clusters. And the combination of both clustering algorithm and Data Encryption Standard algorithm are provide the overall structure of database and methodology required for the analysis of Data science problems. In other word we can say to facilitate the analysis of large amounts of data, and extract features of data. Large data are partitioned into a number of smaller, more manageable data tiles. K-MEAN++ algorithm uses an SQL query language that enables the image and data mining task, and these features to be used in the mining process, and any additional constraints. K-MEAN++ data can be accessed from within MatLab. In addition, K-MEAN++ has the MatLab command tool, which provides for easy transfer of images and for data processing. . The rich statistical functionality of MatLab, together with the K-MEAN++ & Data Encryption Standard algorithm provides cluster data and perform security algorithm are provide interface and the scalability of its data mining engine, allows

for easy and powerful customization of the data analysis process. Data Encryption Standard is the process of encoding a message or information in such a way that only authorized parties can access it and those who are not authorized cannot. Data Encryption Standard does not itself prevent interference, but denies the intelligible content to a would-be interceptor. DES is the archetypal block cipher that takes a fixed-length string of plaintext bits and transforms it through a series of complicated operations into another cipher text bit string of the same length. In the case of DES, the block size is 64 bits. The subject of data science system has moved a long way from the time when it was thought to be concerned only with digital mapping. Whereas digital mapping is limited to solving problems in cartography, data science is much more concerned with the modelling, analysis and management of geographically related resources. However, there is a widespread lack of awareness as to the true potential of Data science systems in the future. When the necessary education has been completed, will the systems be there to handle the challenge? It has to be said that the perfect Data science system has not yet been developed. Today's database technology is barely up to the task of allowing the handling of geographic data by large numbers of users with adequate performance. Serious questions have been raised as to whether the most popular form of database, the relational model, will be able to handle the all types of data with adequate response. Certainly, if this data is accessed via the approved route of SQL calls, the achievable speed is orders of magnitude less than that which can be achieved by a model structure built for the task. It is a common problem with systems that contain parts that are front ended by different languages that it is not possible to integrate them properly. Modern query languages such as SQL are not sufficient in either performance or sophistication for much of the major development required in a Data science system - but then one would argue that they were not intended for this. A problem which has to be addressed is spatial queries within the language, since trying to achieve this with the standard set of predicates provided is extremely difficult and clumsy. An example of a spatial query is to select objects "inside" a given polygon. If the route adopted is to provide two databases in parallel, a commercial one driven by SQL and a geometry database to hold the graphics, and then there is a problem constructing queries that address both databases. The rest of the paper is organized as follows: Section 2 gives the details of related work. Methodology introduced in section 3. We discuss our experiments and the in Section 4. Conclusions are presented in section 5.

II. RELATED WORK

A great deal of research has been focused the use *o*f data science in the spatial analysis of an archaeological cave site[1],Tran, YH and Tran, QN. 2017.Estimating public opinion in social media content using aspect-based opinion mining. In: International Conference on Mobile Networks and Management, 101–115. [2], Guo, K, Shi, L, Ye, W and Li, X. 2014. A survey of Internet public opinion mining. In: International Conference on Progress in Informatics and Computing, Shanghai, China, 16–18 May2014, https://doi.org/10.1109/PIC.2014.6972319. [3], Ece AKSOY, Turkey, presented there is no universally applicable clustering technique in discovering the variety of structures display in data sets. Also, a single algorithm or approach is not adequate to solve every clustering problem. In data science environment used Self Organizing Maps (SOM) algorithm which is the best and most common spatial clustering algorithm in recent years. data science): Current Issues and Future Challenges in June 8, 2009[4], according to Peter Folger**,** Geospatial information is data referenced to a place a set of geographic coordinates which can often be gathered, manipulated, and displayed in real time. A Geographic Information System (data science) is a computer system capable of capturing, storing, analyzing, and displaying geographically referenced information. Global Positioning System (GPS) data and their integration with digital maps have led to the popular hand-held or dashboard navigation devices used daily by millions Challenges to coordinating how geospatial data are acquired and used collecting duplicative data sets. Implementation of the Extended Fuzzy C-Means Algorithm in Geographic Information Systems [5], Ferdinand Di Martino, Salvatore Sessa, in 2009, focused on density cluster methods have elevated computational complexity and are used in spatial analysis for the determination of impact areas. We propose the extended fuzzy c-means (EFCM) algorithm like alternative method because it has three advantages: robustness to noise and outliers, linear computational complexity and automatic determination of the optimal number of clusters. We can use the EFCM algorithm in spatial analysis for the determination of circular buffer areas. These areas can be considered on the geographic map as a good approximation of classical hotspots. Applications to other frameworks like crime analysis, industrial pollution, etc. shall be tried in future works. Issues of data science data management [6] 2007, this paper deals with current issues of spatial data modeling and management used by spatial management applications. Paper describes ways of solving this problem. Now we can summarize the problem of the data science and CAD integration. Because of the different characteristics of the data science/CAD worlds, firstly there's need to decide for some suitable 3D data model, which could maintain complex and structured data types. This model also must be able to maintain the large-scale 3D models produced by CAD as well as low-scale objects used by data science. Comparative Analysis of k-mean Based Algorithms [7] 2010, in this paper they make analysis of k-mean based algorithms namely global k-means, efficient k-means, k-means++ and x-means over leukemia and colon datasets.

III. METHODOLOGY

In this paper we propose combination of clustering algorithm K-MEAN++ and security Data Encryption Standard algorithm to solve the data security problem of Data science. Data Science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured, similar to data mining.

Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data. It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science.

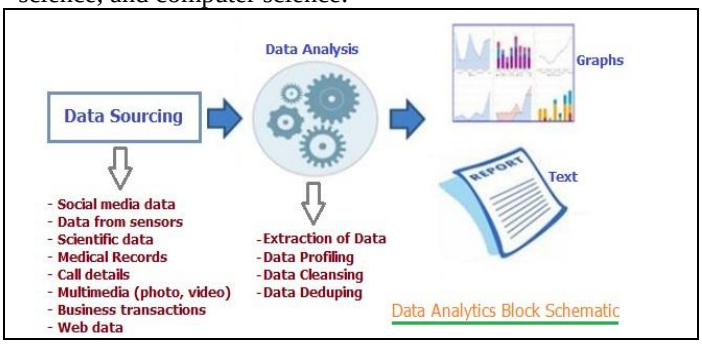

Fig 2.a Data scienec with K-MEAN++ algorithm.

K-MEAN++ is another variation of k-means, a new approach to select initial cluster centers by random starting centers with specific probabilities is used. With the intuition of spreading the k initial cluster centers away from each other, the first cluster center is chosen uniformly at random from the image data points that are being clustered, after which each subsequent cluster center is chosen from the remaining image data points with probability proportional to its distance squared to the point's closest cluster center.

The algorithm is as follows:

1. Choose one center uniformly at random from among the data points.
2. For each data point x, compute $D(x)$, the distance between x and the nearest center that has already been chosen.
3. Add one new data point at random as a new center, using a weighted probability distribution where a point x is chosen with probability proportional to $D(x)^2$.
4. Repeat Steps 2 and 3 until k centres have been chosen.
5. Now that the initial centres have been chosen, precede using standard k-means clustering.

This method gives out considerable improvements in the final error of *k*-means. Although the initial selection in the algorithm takes extra time, the *k*-means part itself converges very fast. But lowers the computation time too. The method provides with real and synthetic datasets and obtained typically 2-fold improvements in speed, and for certain datasets close to 1000-fold improvements in error. Additionally, we calculate an approximation ratio for their algorithm. The k-means++ algorithm guarantees an approximation ratio $O(\log(k))$ where k is the number of clusters used. This is in contrast to *k*-means, which can generate clustering arbitrarily worse than the optimum.

Data security and Database Organization Data Encryption Standard: is a system for data mining and statistical analysis of large collections of remotely sensed images. And also provides the infrastructure and methodology required for the analysis of land surveying, aerial photography, and mathematics, geography and satellite images. Big data uses an SQL-like query language that enables specification of the data mining task, features to be used in an image mining process, and any additional constraints. The query language allows the user to specify the type of knowledge to be discovered, and the set of image data relevant to the image mining process. Based on this an SQL query statement is constructed to retrieve the relevant image data. This new domain requires expertise in image processing, database organization, pattern recognition, content based retrieval and data mining: image processing indicates the understanding and extraction of patterns from a single image; in this system provides users the capability to deal with large collections of images by accessing into large image databases and also to extract and infer knowledge about patterns hidden in the images, so that the set of relevant images is dynamic, subjective and unknown. It enables the communication between heterogeneous source of information and users with diverse interests at high semantic abstraction. The GUI (graphical user interface) enables browsing and manipulation of the images and associated features, creation of data mining queries, and visualization of the results of the data analysis.

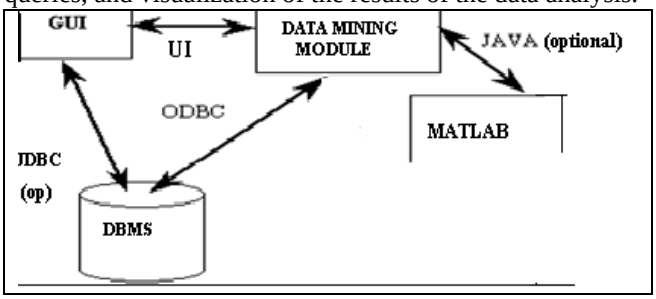

Fig 2.b. Combination of K-MEAN++ & DES with Matlab.

The system is capable of performing similarity searches based on any combination of features. Big data is based on decision tree models. Big data can be accessed from within MatLab by using Java connectivity for images and ODBC connectivity for image and region data .In MatLab, which provides the interactive program for numerical computation and data visualization which along with its programming capabilities provides a very useful tool for almost all areas of science and engineering, which is used extensively in both

academia and industry. The Big data can also display graphics, which are created using a command line interface and shown within MatLab figure window. The combination of MatLab and Big data features creates a unique environment for interactive exploration and analysis of remotely sensed image data.

IV. RESULTS AND DISCUSSION

The K-MEAN++ & Data Encryption Standard algorithm can also display MatLab graphics, which are created using a command line interface and shown within figure window.

Table 1

Results over different variations of k-means algorithm using a tree image classified according colors.		
(Total number of records present in dataset = 70)		
Clustering Algorithm	Correctly Classified	Average Accuracy
K-MEAN with DES	68	94.88
K-MEAN++ with DES	70	95.83

The combination of K-MEAN++ and security Data Encryption Standard features create a unique graphics environment for data mining and these data are stored in database in form of vector and MatLab provides the good database connectivity in data science. In order to facilitate the analysis of large amounts of data, we extract features of the data. Large amount of data are partitioned into a number of smaller (segmentation), more manageable image tiles. Partitioning allows fetching of just the relevant tiles when retrieval of only part of the data is requested, and provides faster segmentation of data tiles. Individual data tile is processed to Encrypt and extract the feature vectors.

PERFORMANCE EVALUATIONS

MatLab connectivity: is an interactive computing Environment for graphics, data analysis, statistics, and mathematical computing. Data was then transferred from MatLab to MS Access using the database connectivity (ODBC) tools as provided by the MatLab Database Toolbox. The data was then transferred from MatLab across the LAN to the SQL Server 7. This process is repeated for matrices with 4 columns per row then 253 columns per row.

Each matrix contained 1000 rows. Once the MatLab process was complete, MatLab was closed and MS Access opened. A process was then run that gathered the timestamp information for each row written to the MS Access tables and the SQL Server 7 tables. The SQL Server 7 tables were then emptied, and the row data in MS Access was written to SQL Server 7.It contains a superset of the S object-oriented language and system originally developed at AT&T Bell Laboratories, and it provides an environment for high-interaction graphical analysis of multivariate data, modern statistical methods, data clustering and classification, and mathematical computing. In total, MatLab contains over 3000 functions for scientific data analysis. Big data can be accessed from within MatLab by using Java connectivity for images and ODBC connectivity for image and region data. In addition, Big data has the MatLab command tool, which provides for easy transfer of images, and for data processing. The K-MEAN++ display MatLab graphics, which are created using a command line interface and shown within figure window. The combination of MatLab and K-MEAN++ features creates a unique environment for interactive exploration and analysis of remotely sensed image and data. The rich statistical functionality of MatLab, together with the approach user interface and the scalability of its data mining engine, allows for easy and powerful customization of the data analysis process.

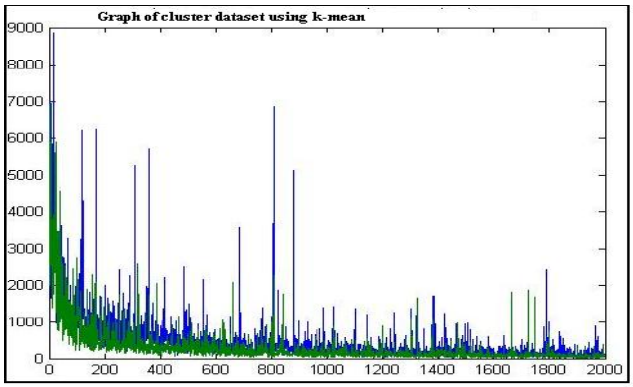

Figure 3.a Graph for combination with DES and k-mean

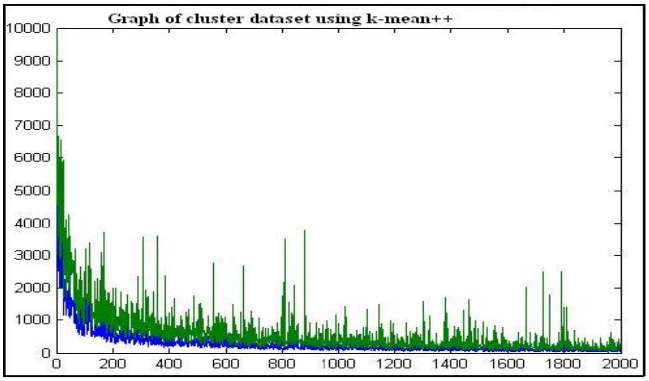

Figure 3.b Graph for combination with DES and k-mean++

V. CONCLUSION AND FUTURE SCOPE

In this paper we presented a K-MEAN++ and Data Encryption Standard algorithm using MatLab framework, provides powerful security with numeric engine and technical programming environment which makes it easy to create and manipulate vectors, MatLab has become the language of technical computing, which explores state-of-the-art data mining and databases technologies to retrieve integrated spectral and spatial information from Data science system. A scalable data warehouse containing a huge amount

of images may be a better database architecture for fundamentally distributed data management and mining system such as NASA Earth Observing System (EOS). Meanwhile, performance analysis for clustering on and retrieving from large volumes of images is critical for the system to succeed in practical applications. And the results of experiments on the basic of images show that the proposed approach can greatly improve the efficiency and performances of image retrieval, as well as the convergence to user's retrieval concept. Clustering algorithm has been widely used in computer vision such as image segmentation and Data Encryption Standard algorithm is able to distinguish between pixel, region and tile levels of features, providing several security feature extraction algorithms for each level. In addition, current implementation provides data and image based search. A segmentation process can be used to segment an image into non-overlapping regions on which we can further apply the texture feature extraction.

In future we use Asymmetric algorithm in place of symmetric algorithm to improve reliability, security and performance of data in data science in distributed environment and performance gains related to advances in data science.

REFERENCES

[1] Hao, X, An, H, Zhang, L, Li, H and Wei, G. 2015. Sentiment Diffusion of Public Opinions about Hot Events: Based on Complex Network. *Plos One*, 10(10): e0140027. DOI: https://doi.org/10.1371/journal.pone.0140027.

[2] M.C. Burl, C. Fowlkes, and J. Roden, "Mining for image content," in Systemic, Cybernetics, and Informatics / Information Systems: Analysis and Synthesis, Orlando, FL, July 1999.

[3] L.-K. Soh and C. Tsatsoulis, "Data mining in remotely sensed images: a general model and an application," in Proceedings of IEEE

[4] IGARSS 1998, vol. 2, Seattle, Washington, USA, Jul 2012, pp. 798-800.

[5] J. Zhang, H. Wynne, M. L. Lee, "Image mining: issues, frameworks, and techniques," in Proceedings of 2nd International Workshop on Multimedia Data Mining, San Francisco, USA, Aug 2001, pp.13 – 20.

[6] G.B.Marchisio andJ.Cornelison,"Content-based search and clustering of remote sensing imagery," in Proceedings of IEEE IGARSS 1999, vol. 1, Hamburg, Germany, Jun 1999, pp. 290 – 292.

[7] A.Vellaikal, C.-C.Kuo, and S. Dao, "Content-based retrieval of remote sensed images uses vector quantization," in Proc. of SPIE Visual Info. Processing IV, vol. 2488, Orlando, USA, Apr 1995, pp.178 – 189.

[8] Ying Liu, Dengsheng Zhang, Guojun Lu, Wei-Ying Ma.A Survey of content-based image retrieval with high- Level Semantics. Pattern Recognition, Volume 40, Issue 1, January 2007, Pages 262-282.

[9] Muhammad Atif Tahir, Ahmed Bouridane, Fatih Kurugollu. Imultaneous feature selection and feature weighting Using Hybrid Tabu Search/K-nearest neighbor classifier. Pattern Recognition Letters, Volume 28, Issue 4, 1 March 2007.

[10] Sarbast Rasheed, Daniel Stashuk, Mohamed Kamel.Adaptive Fuzzy k-NN classifier for EMG signals Decomposition. Medical Engineering & Physics, Volume 28, Issue 7, September 2006, Pages 694-709.

[11] J. Amores, N. SEbE, P. Radeva.Boosting the distance Estimation: Application to the K-Nearest Neighbor Classifier. Patter Recognition Letters, Volume 27, Issue 3, February 2006, Pages 201-209.

[12] Man Wang, Zheng-Lin Ye, Yue Wang, Shu-Xun Wang. Dominant sets clustering for image retrieval. M. Wang et al. /Signal Processing 88 (2008) 2843–2849., Venables W. N. and Ripley B. D. (2000), S Programming, Springer, New York..

[13] Edwards,D.,2005, Excavations at Khirbet Cane, Israel, http://anticompetitive/cane.

[14] M.C. Burl, C. Fowlkes, and J. Roden, "Mining for image content," in Systemic, Cybernetics, and Informatics / Information Systems: Analysis and Synthesis, Orlando, FL, July 1999.

[15] Zlatanova S.: Large-scale data integration An Introduction to the Challenges for CAD and GIS Integration, Directions magazine, July 10, 2014.

[16] Van Ostracism P.: Bridging the Worlds of CAD and GIS, Directions magazine, June 17, 2004.

[17] David Arthur and Sergei Vassilvitskii: k-means++: The Advantages of Careful seeding, Proceedings of the eighteenth Annual ACM-SIAM Symposium on discrete algorithms. pp. 1027—1035, 2007.

[18] Zhang Y, Mao J. and Xiong Z.: An efficient Clustering Algorithm, In Proceedings of Second International Conference On Machine Learning And Cyber netics, November 2003.

[19] IEEE Trans. on Knowledge and Data Engineering, 14, No.5, Sept/Oct 2009.

[20] M. E. Hellman, "DES will be totally insecure within ten years", *IEEE Spectrum*, vol. 16, no. 7, pp. 32-39, July 1979.

