RESEARCH ARTICLE OPEN ACCESS

A Novel Solution for Initial Cluster Centre Problem for K-mean Algorithm over Cloud

[1]Mohd Iqbal, [2]Manish Sharma
[1]*Research Scholar, Professor Center for cloud infrastructure and security Suresh Gyan Vihar University, Jaipur*
Corresponding Author: Manish Sharma

ABSTRACT:- k-means algorithms are one of the most applied algorithm due to its simplicity and efficiency, k-means algorithm performance are generally good but also has got some limitation, it takes initial cluster centre as an arbitrary data points, number of cluster are given by the user, missing data problem, inefficient with the large data set, various research work has been done to improve efficiency of the k-means algorithms. Our main work to deal with the initial cluster centre problem various research work has been done in this direction but still there is not any global method of the initializing the initial cluster centre or mean of the k-means algorithms, in our work we generate the initial cluster centre by using the idea of M-way tree, this method is efficient as compare to the traditional k-means algorithms.
Keyword: -Data Mining, Clustering, k-Means, Efficiency, Optimization

Date Of Submission: 15-09-2019 Date Of Acceptance: 03-10-2019

I. INTRODUCTION

Size of the databases increases as the internet base transaction speed-up, all the data related to the internet transaction, medical related data and from some other application are stored in the databases which may be helpful in the future decision making, .it is tough to produce unknown and hidden information from any large data repository answer of this "Data Mining". Data mining is also helpful in increasing revenue as well as in cost cutting, by using some "data mining" sophisticated technique. Data Mining may be defined as a way of exploring hidden patterns or hidden knowledge from any large databases which is helpful in decision making.

Since the 1960 databases as well as information-technology is changing from old file system to the new sophisticated database system, Since the 1970 old age of computer research work is moving from the hierarchical base system to the relational database system, in relational data base system data are store in the form of the table and some other sophisticated database technique also introduced. As the time passing on some efficient query processing language introduced with the help of that query speed increases query optimization and user-interface quality also improved with the help of that user can access any data into fraction of time, in 1970, relation database was a standard database, which decreases query processing time. In 1980, databases functionality increases in this era extended database, object oriented database, deductive model and application based database such as spatial, scientific and multimedia database were also popular. Data Mining uses sophisticated tools and automatic algorithms to predict the future trend which helpful in decision making in business as well as some other sector. For example, data mining helps in targeted marketing, using previous information which is available to the businessman or any other sector, other application of data mining is in the field of the banking sector where data mining help in fraudulent customer identification as well as in identifying credit-card fraud. Other example of data mining is in the field of the retail business where data mining helps in predicting the future trend of buying behaviour of the customer.

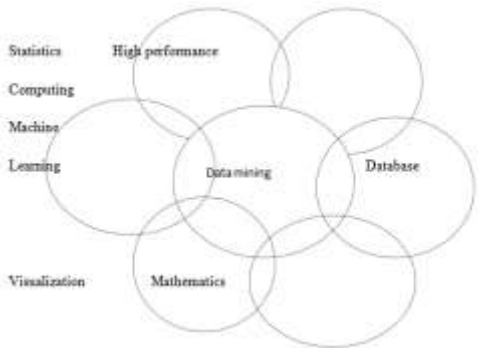

Figure1. Data mining in different-different fields

II. LITERATURE REVIEW

Paper [23] address these problem of the k-means algorithms, in the paper [23] B.Mahela and team work they proposed a method, in this method they use imputation method to fill the missing values, this method they applied in the two

algorithms C 4.5, this algorithms they mix more missing values, in k-means algorithms they fill the missing values with the data imputation, experimental results shows that k-means data imputation method is better.

Dan Li and team work [24], this paper also deal with the missing value in dataset, missing values of the data in the dataset are general in the fields of research. Reasons behind of the missing values in the data seat are equipment's unavailability and not proper functioning of the equipment, due to this an accurate and efficient result does not come.

Missing values generally deal by the following three methods:

- Ignoring methods.
- Parameter estimation.
- Imputation.

Imputation method is that method in which missing values are filled by using some similar cluster information values. Generally missing values in dataset are filled by using the k-means algorithm but problem with k-means algorithms is that it having chances of getting stuck in the local minima. I n this paper, they purpose a method in which by using the fuzzy k-means algorithms, fuzzy k-means algorithms use a belongingness function of each data point to each cluster some extends which make this algorithms less sceptical to the local minima problem. Experiment results shows that this algorithms are better in performance as compare to the k-means algorithms.

In the paper [25] Satish Gajawada and team work, generally database contain some missing values due to which whatever query are applied at those dataset does not provide the good results, to deals with the missing value clustering imputation method are applied but when the number data is more whose values is missing then it is tough to imputed that missing values.In this proposed method they have use K-NN as well as K-means to deals with the missing values, in this method to decreases the number of missing values it suppose some missing data values as imputed values after that they applied the imputed object to fill the missing values, this method applied at the clinical data set and produce good results.

In the paper [26] M.S. R.Malarvizhi, this paper also deals the problem of the missing values of the variable in data set, to deal with the missing value in the data set in this approach two methods applied , K-means clustering approach and K-NN approach, in the K-means clustering imputation approach, means are taken of the non-missing values and replace at the place of the missing values same approach are applied at the K-NN strategy when both method compared with the each other K-NN approach is better than the K-means imputation methods.

III. PROPOSED WORK

As K-means algorithms is one of the most widely employed algorithms for grouping the large set of the data in many partition on the basic of the some similarity between the data set which are grouped in the same cluster, generally in the K-means algorithms distance measure is a method in which distance between the cluster centres and each data point is calculated, to which distance between the data points and cluster centre is less data points cluster with that cluster.

As the K-means algorithms though one of the most employed algorithms it has some limitation.

- Initial cluster centre in the K-means algorithms are taken arbitrarily due to which it efficiency decreases.
- Produce sub-optimal solution.
- In K-means number of cluster to be formed is predetermined.
- Inefficient to deal with the missing value.
- Unable to work with the mix data set.

A main concern of our work is that deals with the problem of initial cluster of the K-means algorithms. In our approach for the initialization of the initial cluster centre we use M-way tree, we use M-way tree approach in such a way, in M-way tree order of the tree is fixed and it work on the array data set, if the number of elements increases more than the order of the tree then whole elements are divided in following parts left-side tree, right side tree and the root of the tree in our approach we applied M-way tree in such a way.

A	1	3	5	7	9	11	13	15	17
B	19	20	23	25	27	29	30	31	33
C	35	37	39	43	44	46	48	49	50

For our experimentation work we has used synthesised data set due to which we do not need to perform the data transformation step to deals with the missing values, manually we has sorted the data in row wise fashion in such a way data in each row are in the increasing order and data in each row is the odd number of elements, our formula to derive the initial cluster centre.

Initial cluster centre=number of elements in a rows /2

As for describing our method in the above table there are 9 elements in each rows, as we consider this division process as an integer so output will be 4, so the initial cluster centre will be the 5^{th} elements of each rows, after initializing the initial cluster centre other step of K-means algorithms will be performed as a dis.

Steps of K-means algorithms are as follows:

1. The first cluster or K-sense algorithms are deliberately taken in the first step.
2. In the second step, the distance is calculated from each data point, each primary cluster centre data points are allocated to the primary cluster centre, which is the lowest distance calculated.
3. Calculation of each cluster is calculated, now the calculation has become the new primary cluster centre.
4. Again each data point and the distance from each cluster centre will take place and new ways will arise repeatedly.
5. Step 3 and 4 will repeat repeatedly and will take balance when it closes.
6. Process will be cancelled when the balance will be taken.

Steps of our modified K-means approach are as follows:

1. Initially the information itself is short.
2. The primary cluster centre will be taken using the M-way tree trip.
3. Distance in second step is calculated from each data point. Data points of each primary cluster centre are allocated to the primary cluster centre, which is the lowest to be calculated.
4. Calculation of each cluster is calculated, now the calculation has become the new primary cluster centre.
5. Again each data point and distance from each cluster centre will take place and new ways will be generated repeatedly.
6. Step III and IV will be repeated again and repeat when the balance will be taken.
7. The process will be cancelled when the process will be cancelled

IV. IMPLEMENTATION

Implementation has done in the programming language java supported NetBeans (IDE), NetBeans start-up was initially college project the main aim of the group of the developer to develop a java based environment which is more flexible than other tools, it is written in the programming language java initially developing group member of the IDE started their own company to develop and distribution of that environment to other people after that an entrepreneur name Roman Stanek was looking for a good idea for investment, he met to original group member of the IDE and they team-up for the distribution as well as the development of the NETBEAN IDE during that period of the time various version of the NETBEANS IDE released during that release working functionality of the NETBEANS IDE increases day-by day after release of the JDK 1.3 people started to develop their own plug-in due to which NETBEANS market was destabilizing or you can say that it was shrinking during the same period of the time Microsoft was looking for the good java development environment and start taking interest development of the NETBEANS IDE, finally NETBEANS comes the open source software which was helpful in various fields, it is a tool which support multi-platform development environment like NETBEAS support many programming language like C++,JAVA,PHP why people prefer NETBEANS as their programming development environment some reason are as follows

1. It is an open source software
2. Easy to use
3. Multilanguage supporting environment
4. Better graphical User Interface

Our work is implemented in the java programming languages, Java Programming language invented due to some limitation in C and C++, java programming language adopted the syntactical approach from C and object oriented approach adopted from the C++ programming language, java is not a new language java development is step by step procedure adopting some attribute form old language.

C programming language play a major role in development of the java programming language, before the java programming language FORTRAN programming language but problem with the FORTRAN programming language was that it was only design for the scientific application.

Another programming language which is efficient BASIC, but Problem with this language is that it structure was not well defined, programming language like BASIC,FORTRAN,COBOL, the basic principal of these languages was goto statement and these languages become complex, as there were various programming languages problem with these programming language were that it were design for the specific purpose so all these programming languages can't be applied on general type of the problem during 1970 computer world need such type of the programming languages which can deal with the all the type of the application so C language was develop which was efficient and simplest one but one of the problem with the c is that when size of the program increases it is really tough to understand the working of the code.

As there were very good programming language was C which was universally accepted then why we need a programming language C++, answer is that C++ because as the program size increases C++ become more complex to deal with that complexity, C++ programming language is another step towards the invention of the java programming languages, C++ got various feature like class, encapsulation and there were some other feature due to which C++ was most easily and structured programming language, all these feature like inheritance, encapsulation, which later adopted by the java languages, C++ was good enough but it was also some limitation like platform independence and internet programming, java was developed by the Sun Microsystems engineer, they took around 18 month to lunch java first working version.

Java is a programming language which is well defined architecture based upon the oops principal, java code is a simple English language instruction, first code is compiled into the byte code after that code is run and it produced the user desired output

V. RESULTS :

Our results shows that our method is better as compare to the traditional K-means algorithms, we are using in our approach M-way tree approach to get initial cluster centre , experimentally when we compare our method with the traditional K-means approach our method is taking less time as well as the iteration as compare to the traditional K-means approach, below chart comparison of the our method has been done with the traditional approach at a particular result of the both method, it may vary but our method is performance in maximum number of the time up to the marks as compare to the traditional one.

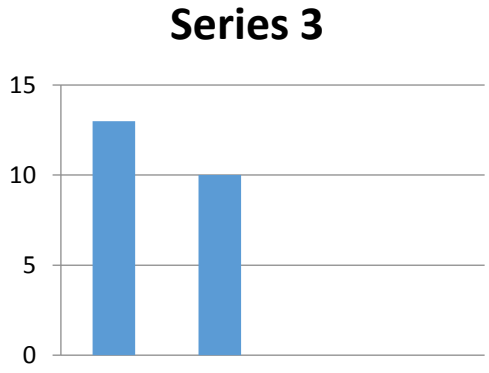

Figure 4: Comparison chart of K-means and modified K-means

5.1 K-means output

when clustering of the data are performed by using the K-means approch as we are forming the same number of clustre in the both the method output of the K-means algorithms shows at a particular point it is taking 31 milisecond and 7iteration , at the in this work we shows form number of the cluster at the last.our approach we are using M-way tree to get the initial after entering the initial cluster centre clustering is done in our method, our method result at particular instance taking 15 millisecond and 3 iteration so this method is efficient as compare to the K-means algorithms it is taking less time as well as the less iteration as compare to the traditional K-means algorithms.

VI. CONCLUSION

As k-means partitioning clustering approach has various disadvantages some of them are as follows number of cluster need to be predetermined, Initial cluster centre problem, inefficient to dealing with data uncertainty, various research work are being done in this approach our main concern is the initial cluster centre problem, as K-means algorithms take initial cluster centre arbitrarily in our proposed approach we have used M-way tree approach to get the initial cluster centre in the K-means algorithms. As in our approach we have use M-way tree to get the initial cluster centre problem to improve the efficiency of the K-means algorithms some other approach are welcomed concerning the initial cluster centre to improve the efficiency of the K-means algorithms. K-means algorithms are applied to deals with different-different application one fields where some method can be developed to improve the efficiency of the K-means algorithms regarding dealing with the data uncertainty because when K-means are applied in the data set which contain data uncertainty, performance of the K-means algorithms degrades dramatically so by applying some data structure K-means performance may be improved even this direction.

REFERENCES:

[1]. Shitl A Raut and S.R.Sathe, “A modified fast K-means clustering algorithms for large scale gene expression data set”, Volume 1, No 4, November 2011.

[2]. Soniya Sharma and Shikha Rai, “Genetic K-means Algorithms-Implementation and Analysis”, International Journal of Recent Technology and Engineering, ISSN: 2277-3878, Volume-1, Issue2, and June 2012.

[3]. Rajshree Dash and Rashmita Dash, “COMPRATIVE ANALYSIS International Journal of Advanced Computer and Mathematics Sciences, ISSN: 2230-9624, Volume 3, Issue 2, 2012, pp 257-265.

[4]. Qian Ren and Xinjing Zhuo, 2011 IEEE International Conference on System Biology (ISB) , 978-1-4577-16669-9.
[5]. Chittu V and N Sumanthi, "A Modified Genetic Algorithms Initializing K-means Clustering,Global Journal of Computer Science and Technology, Volume 11, Issue-2, February 2011.
[6]. Jing Wang,JingdongWang,Qifa Ke§ Gang Zeng and Shipeng L, Microsoft Research Silicon Vally (Asia).
[7]. Adil M Bhagirov, Julien Ugon and Dean Webb, "Fast Modified Global K-means Algorithms for Incremental Cluster Construction", Pattern Recognition 44(2011), 864-876.
[8]. Z.H.S. chan and N Kasabov, "Efficient Global Clustering Using the Greedy Elimination method" electronics letters, 9th December 2004, Volume 40, No 25.
[9]. Kohei Arai and Ali Ridho Barakbah,, "Hierarchical k-means: an algorithms for centre initialization for K-means algorithms", Reports of the Faculty of Science and Engineering, Volume 36, No 1, 2007.
[10]. Yogul Kumar G. Sahoo, "New initialization method for originate cluster centre for K-means algorithms", International Journal of Advanced Science and Technology, Volume 62, (2014), pp. 43-53.
[11]. Bin Zang et.al, "k-harmonic mean data clustering algorithms", Hp laboratory palo alto,(1999).
[12]. D Nepolian and S Pawalacodi, " An enhanced K-means algorithms to improve efficiency using Normal distribution data points", International Journal of Science and Engineering", Volume 13 No-7, January 2011.
[13]. Raed T.Aldadooh and Wesam Asohour, "DIMK-Distance-based Initialization Method for K-means algorithms", Intelligent System and Application, 2013, 2, 41-51.
[14]. Wang Shunye, Cui Yeqin, Jin Zuotao and Liu Xinyuan, "K-means Algorithms in the Optimal Initial Centre Based on Dissimilarity", Journal of Chemical and Pharmaceutical Research, 2013, 5(12): 745-749.
[15]. Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, D.Payetko, Ruth Silverman and Angela Y. Wu, "An Efficient K-means Clustering Algorithms: Analysis and Implementation" IEEE Transaction on Patterns Analysis and Machine Intelligence, Volume 24, No-7,July 2011.
[16]. Ahmad Safeek and Harshika, "Dynamic Clustering of Data with Modified K-means Algorithms", International Conference on Information and Computer Networks", IPCSIT Volume 27,2012.
[17]. G Kumara Swamy and Amitabh Wahi, "New Algorithms for Selection of the Better K-Value Using Modified Hill Climbing in K-means Algorithms", Journal of Theoretical and Applied Information Technology, 30th September 2013, Volume 55, No.13.
[18]. Nimrat Kour Sidhu and Rajneet Kour, "Redefining and Enhancing K-means Algorithms", International Journal of Innovative Research in Computer and Communication Engineering, Volume 1,Issue 3,May 2013.
[19]. Krista Rizman Zalik, "An Efficient K-means Clustering Algorithms", Pattern Reorganization Letters 29(2008),1385-1391.
[20]. [20] Sanjiv K Bhatiya, "Adaptive K-means Clustering", American Association for Artificial Intelligence, 2004.
[21]. Amir Ahmad and Lipika Dey, "A k-means Clustering Algorithms for Mixed Categorical and Numerical Data", Data and Knowledge Engineering 63(2007) 503-527.
[22]. Ziang Huang, "Extension to the K-means Algorithms for Clustering Large Data Set with Categorical Values", Data Mining and Knowledge Discovery,2,283-304, 1998.
[23]. B. Mehala, P. Ranjit Jeba Thangaiah, and K. Vivekanandan, "Selecting Scalable Algorithms to deal with Missing Values", International Journal trend and Engineering, Volume 1, No 2,May 2009.
[24]. Dan Li, Jitender Deogun, William Spaulding and and Bill Shuart2, "Towards Missing Data Imputation: A Study of Fuzzy K-means Clustering Method", RSCTC 2004,LNAI 3066,pp. 573-579,2004.

Manish Sharma" A Novel Solution for Initial Cluster Centre Problem for K-mean Algorithm over Cloud " International Journal of Engineering Research and Applications (IJERA), vol. 9, no. 9, 2019, pp. 60-64