International Journal of Research in Advent Technology (IJRAT) Special Issue
E-ISSN: 2321-9637
Available online at www.ijrat.org
*International Conference "*INTELINC 18*", 12th & 13th October 2018*

Data Mining for Big Data: Issue and Challenges

Prof.Dr. K. Venkatachalapathy[1] , C.Krubakaran[2]
[1] *Professor and Head, Dept. Of Computer and Information Science,Faculty of Science, Annamalai University, Annamalai Nagar, Chidambaram,*
[2]*Assistant Professor, Department of Information Technology, Bharathiyar College of Engineering & Technology, Karaikal, India*
[1]*omsumeetha@rediffmail.com* ,[2]*kirubabcet2018@gmail.com*

Abstract— Recent decade has seen an extraordinary insurgency in computerized innovation with condition of craftsmanship advances and gadgets to deal with, store and transmit data crosswise over wired and remote media on a worldwide premise. With expanding research in computerized systems, top notch media content are being seen in every single business application. Superior quality information might be a picture, sound or video and give an inside and out insight about the subject under examination. Constant utility of top notch information has conjured the ideas of huge information and distributed computing in late time. Extraction of helpful learning from these information acquired from different sources shapes the center idea of information mining. This audit paper gives a far reaching review of later and condition of workmanship procedures in mining helpful data from enormous information particularly utilizing bunch registering approaches.

Keywords—Data mining, big data, high definition data, cloud computing, cluster analysis, evaluation methods

I. INTRODUCTION

Computerized innovation has experienced a fast change and might be named appropriately as transformation with customary information taking care of, putting away and correspondence conventions and gadgets getting an entire makeover with late systems and procedures. For instance, high determination satellite symbolism being utilized to gauge changes in the territory, climate conditions and so forth on an ongoing premise has summoned utilization of superior quality picture and video procurement innovations and their ceaseless transmission to earth stations has required appropriate stockpiling instruments for putting away the mass information. This has prompted the approach of huge information which is an arrangement of composite information from numerous securing sensors and sources. A further subordinate of above idea has made ready for distributed computing where clients could get to information whenever at wherever on the globe through specialist organizations. Be that as it may, regardless of all the above viewpoints, the target of this audit paper lies in managing information mining [7] or extraction of valuable and significant data from the composite pool of information accessible in a capacity database or cloud condition. Information acquired from different sources are accessible as a pool and deliberate extraction of data through mining changes the concentrates into organized data giving much required clearness and comprehension to the client. Further, the acquired data is viewed as important as mining includes extraction by distinguishing reasonable examples which are like each other in the composite pool. Uses of information mining change over a far reaching beginning from relapse examination to expectations of securities exchanges, forex administrations, inconsistency [26] and interruption location in systems and frameworks, assessment of undertaking hazard administrations, preparing of smart systems and information warehousing. On a general sense, data mining could be seen as the intersection of several fields such as statistical modelling, data base management systems and services, machine learning techniques etc. as depicted in figure 1.

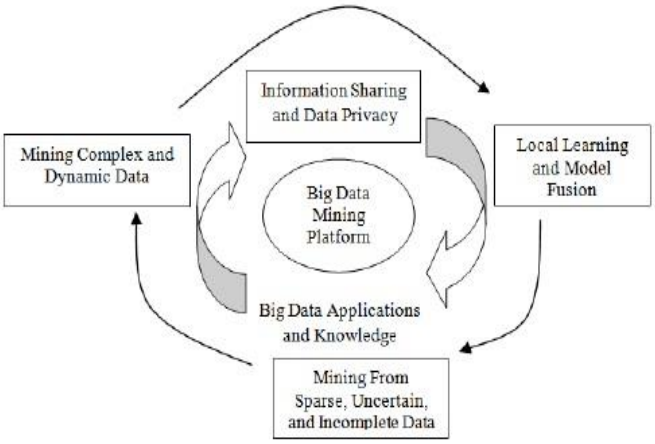

Figure 1 Conceptual illustration of data mining

As observed from the above figure, all information dealing with instruments are covered on a typical information mining stage which translates the yields from every

framework into organized organizations for additionally handling and comprehension. Information mining is accomplished through a few approaches like affiliation based systems [1], grouping based strategies [19], and forecast based methods [16], choice trees [23], and successive examples [18]. A deliberate overview of writing maintaining a strict concentration towards late patterns in information mining in light of group models have been displayed in this paper.

A. Big Data

Evolution of state of art data acquisition techniques have made Development of condition of workmanship information procurement methods have made information elucidation and ensuing examination and expectations to be more exact and exact anyway at the cost of expanding volumes of information measure. Enormous information [29] are portrayed by expanding limit which could be exceptionally well found if there should be an occurrence of a one moment medium lucidity video taking up 10MB of memory with a similar one moment video in superior quality taking up to 150MB of memory utility.

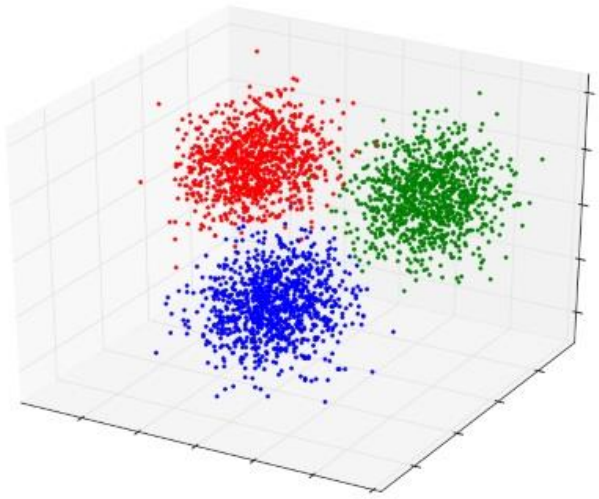

Figure 2 A high dimensional data model [8]

Figure 2 depicts the plot of a high dimensional data model visualized using a Plot Viz tool which represents data from a medical analysis of patients with different symptoms and medical anomaly conditions. Considering the criticality of data clearness got from top quality information, huge information has developed and summoned examine premiums in huge sums as of late. Effective strategies to deal with, process, investigate and store huge information are being inquired about every now and then [28] [29]. Enormous information are as a rule for the most part described by three 'V's in particular volume, speed and assortment [3]. They by and large don't fit into the current regular information taking care of designs like social database administration frameworks (RDBMS). A portion of the powerful devices for mining huge information are KEEL, SPMF, Rattle, Weka, Orange and so forth., [4].

B. Issues in Data mining from big data

A deliberate audit of writing presents the accompanying issues seen in mining information from huge information. An essential stream procedure of an information mining process is outlined in figure 3. The information source which realizes the data sources are pre prepared utilizing reasonable systems [1] and given to design recognizable proof square where the mining procedure happens bringing about learning revelation at the yield. The yields are at times alluded to as learning disclosure database (KDD). The above all else issue in mining enormous information lies in the information source. The information source might be solitary or accumulation of data from different sources which are various in nature. They may likewise be in a homogenous or heterogeneous [28] condition requiring the calculations utilized for mining additionally to be versatile in agreement to the information source write. For instance, a simple picture based information base mining technique won't have the capacity to suit a similar extraction strategy for a pool of video information. The same applies on account of composite information source demonstrate. Consequently the created mining apparatus or calculation which drives it ought to gain from the information source write and locate a reasonable technique to deal with the information. Unmistakable composite information sources may incorporate recordings, pictures, x – beams, versatile configuration reports, designs, phone messages, messages and so on., the nature of information in enormous information frameworks could be an organization of both organized and unstructured segments and thus dealing with these information writes at the same time stances to be a genuine test. The second issue distinguished from the writing [4] lies in giving an appropriate mining to UI to enhance the execution of the translated information. As portrayed in figure 2, the lucidity of the model is very exceptionally poor which is incredibly enhanced through solid mining instruments. In any case, post mining procedure ought to have an unmistakable and effortlessly justifiable UI which empowers the end client to comprehend the yield of the mining apparatus at all time conceivable. The mining apparatus ought to have the capacity to give proper information portrayal composes to suitable huge information models.

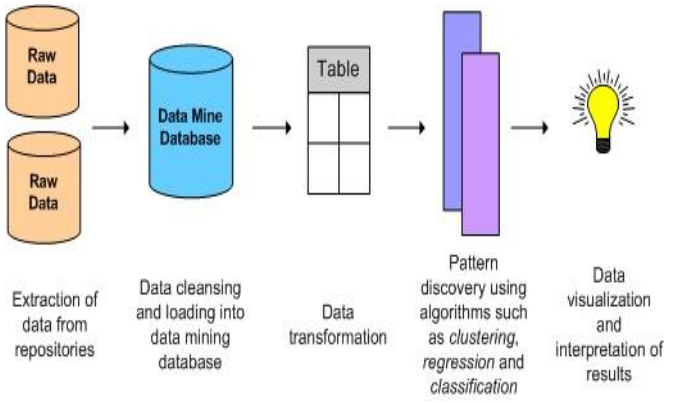

Figure 3 Flow process of data mining

The first and foremost issue in mining big data lies in the data source. The data source is may be singular or collection of information from various sources which are diverse in nature. They may also be in a homogenous or heterogeneous [28] environment necessitating the algorithms used for mining also to be adaptable in accordance to the data source type. It is to be noted that irrespective of data volume or variety, mining is classified into static and dynamic as depicted in figure 4.

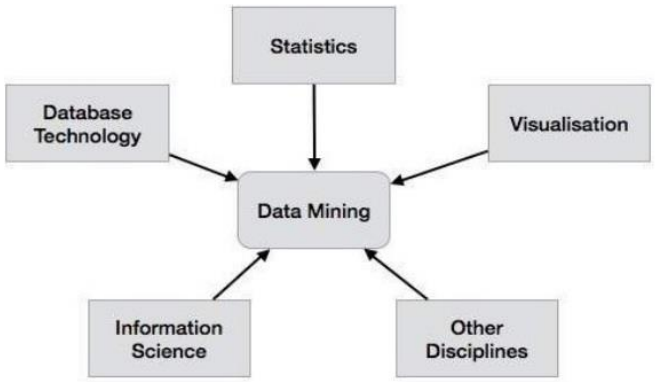

Figure 4 Classification of data mining

Security [22] has been seen to be an issue of significant worry in executing a proficient information mining strategy since information mining typically extricates data from the unstructured pool of information into a more important organization which could be effectively deciphered by the client. Amid this procedure, it is additionally to be noticed that the likelihood of revealing some crucial, secret and private data of people from the unstructured configuration into a reasonable organization are very high in this way damaging the classification strategies of the framework. Thus the mining calculation ought to be planned with the end goal that it deals with not revealing private and secret data amid the mining procedure [23]. Another issue recognized as a testing undertaking in mining is distinguishing proof of reasonable method for mining huge information. Ordinary insight based and machine learning based techniques are accessible in substantial numbers for mining typical information. In any case, the flexibility of this calculation for dealing with huge information for the most part of the scopes of tera bytes is a significant sketchy and testing issue. Writing likewise displays substantial number information diminishment procedures like main part investigation (PCA) [25] accessible for lessening the component vector measurements or the information estimate. This is typically accomplished by expulsion of excess data from the component vector. In any case, in the event of enormous information mining, the subject of the example or rationale behind which certain data ought to be barred for diminishing the information vector measure is very testing. End of data from huge information without appropriate learning based principles or preparing techniques could bring about loss of imperative information rendering the mining strategy to be wasteful [39].

II. REVIEW ON CLUSTER BASED MINING METHODS

Among the distinctive mining techniques accessible, this survey paper focuses on group based strategies for mining of information. A few ongoing calculations have been contemplated and discoveries are methodically exhibited in this area.

Various leveled bunching methods [5] [31] depend on grouping tree based methodologies and these bunch trees are called dendrograms and are more reasonable for little example information sources. They are additionally split down into base up and top down methodologies with the previous otherwise called agglomerative and the last otherwise called disruptive bunching. Further, the previous is found to blend at least two comparable bunches in back to back cycles while the last parts the comparable groups into more group trees. At each bunch arrange k-implies grouping [16] has been utilized. K implies grouping calculation has been all the more regularly observed to be misused for ideal mining applications in the writing [6] [18] [20].

Two noteworthy benefits of K imply being broadly utilized as a part of writing is that it permits parallelization and works adequately independent of information arrange. A few calculations have been accounted for in the writing which has centered towards adjusting existing K implies calculations to enhance the mining productivity. Two stage grouping strategies [12] have been tested in the writing where the principal stage manages calculation of bunches in a deliberate way to deal with create groups with accuracy. A positioning based K implies calculation [8] has been found to enhance the speed of mining when contrasted with the current bunch based K implies which makes it perfect to be utilized for enormous information plans. Trial comes about demonstrate 476s time utilization for mining of up to 500 records utilizing the positioning strategy with a 487s detailed for a similar record

number utilizing grouping with regular K implies. A change [14] in time reaction and decrease of many-sided quality for mining has been accomplished utilizing calculation of uniform information indicates what's more K implies bunching application to the mining framework. A triple calculation have been introduced in the writing [22] which tends to the issues of choosing the ideal number of groups and evacuation of dead point of confinement [21], lessening of computational and time multifaceted nature.

Framework based strategies [32] have additionally been tested in the writing by building an arrangement of lattice cells took after by calculation of cell thickness. The cells beneath a foreordained limit are wiped out. In view of limiting a goal work, a bunch is made with neighboring comparative gatherings. Sting and Clique calculations [33] are conspicuously utilized as a part of matrix based bunching. From considers in the writing it is watched that STING is an inquiry autonomous approach and parallelization is very infeasible. Then again, CLIQUE calculations allow bunch arrangements of any self-assertive shapes and create groups from thick subspaces [34] [35] with apriori approach. Parcel based approach have likewise been broadly looked into in the writing [38] [39] where an arrangement of p allotments are made at first took after by development of items starting with one gathering then onto the next through an iterative migration procedure. The iterative migration method indeed is found to K implies, Fuzzy C means and K medoids [40].For enormous information, inspecting utilizing CLARA (grouping huge applications) have been used alongside K medoids in a cross breed blend to deliver adaptability because of expanding information measure. Preferred standpoint of K medoids saw in exploratory discoveries from writing [31] demonstrates that they have a minimum calculation time as the estimation of separation between sets happens just once dissimilar to their K implies bunching partners [22-24]. Then again Fuzzy C implies goes under the class of delicate figuring methods and takes a shot at the standard of minimization of target work by introducing a participation framework in view of the chose number of bunches. The participation grids are refreshed in view of the registered bunch focuses and the procedure ends when the figured enrollment network esteems turn out to be not exactly a predefined edge. Else the calculation repeats towards limiting the participation framework.

Imperative construct grouping calculations [47] work with respect to processing an underlying arrangement subject to some client characterized limitations like requirements on singular items, hindrance objects, bunching parameters and so on., [38]. Machine learning based strategies are widely found in the writing [34] as fake neural systems [39] through extraction of emblematic principles strategy which is a change of existing ANN calculation. More exactness and precision have been seen in the test comes about utilizing a four phase ANN preparing through a back spread preparing standard. Additionally writing shows that the regular control based ANN techniques are not that exact in their bunching yields and are likewise computationally costly. A two phase ANN demonstrate found in the writing legitimizes a commotion lessened bunching yield and in addition its capacity to deal with a scope of chaotic information. Fluffy based strategies [35] have been observed to be broadly used to address the issues of protection and security of mined information relating to classified data of people in the pool of information. A protection safeguarding information mining system has been utilized as a part of the exploration works in the writing with the qualities being changed over into fluffy or fresh qualities to save classification of data [15] [34]. Hereditary based calculations [36] have additionally been examined in the writing in a crossover blend with K implies grouping to discover worldwide ideal allotments of given information into bunches. Exploratory investigation shows that GA based strategies create exact mining yields with low intra bunch and high entomb group separates and conquer the downsides of ordinary K implies bunching methods of settling of an imperfect arrangement and the requirement for foreordaining the groups [37]. The remainder of grouping based strategy includes projection and sub space bunching [20] which creates countless and particular calculations have been produced to expel the repetitive bunches in light of harsh set hypothesis to set up apriori property in the end procedure [61]. Developmental methodologies [23] [29] [42] [22] incorporate nature enlivened calculations or bio propelled calculation like molecule swarm streamlining (PSO) [27], Ant settlement enhancement (ACO) and recreated tempering systems [23] and so on.

Other ongoing procedures towards dealing with huge information with mining data from them incorporate Hadoop [24] which is an open source blame tolerant framework for huge information stockpiling and preparing. It adequately handles the issue of conveyed registering through Map Reduce [15 – 16] and devices like Hive, Mahout and Pig are utilized for taking care of the heterogeneous idea of information source. Guide Reduce deals with a grouping plan and fit for preparing extensive volumes of information on a parallel and conveyed approach through two capacities outline lessen.

III. REVIEW OF EVALUATION METRICS

SNR is utilized as a part of information pressure, and exactness and review are utilized as a part of content based data mining. Great measurements will lead the procedure in the right bearing while terrible ones may deceive the examination exertion. As of now, some picture mining frameworks measure execution in light of the "cost/time" to locate the correct matches. Others assess execution utilizing accuracy and review, terms acquired from content based recovery. Despite the fact that these criteria measure the framework's execution to some degree, they are a long way from acceptable. One noteworthy reason causing the trouble of characterizing a decent assessment measure is the recognition

subjectivity of information source content. Since, it is basically an identification based issue, the traditional effectiveness parameters like True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) are utilized to depict Precision, Recall and exactness of the recovery framework. Since, it is essentially a detection based problem, the conventional efficiency parameters like True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) are used to describe Precision, Recall and accuracy of the retrieval system.

$$precision = \frac{tp}{tp+fp} \quad (1)$$

$$recall = \frac{tp}{tp+fn} \quad (2)$$

$$accuracy = \frac{tp+tn}{tp+tn+fp+fn} \quad (3)$$

.

IV. CONCLUSION

A far reaching audit of bunch based digging calculations for huge information has been widely exhibited in this survey article. Huge information is an advancing innovation being executed in cloud systems for access by customers on a worldwide premise at any given purpose of time and anyplace in the globe. Be that as it may, because of the substantial volumes of information which is normal for enormous information frameworks which get contributions from numerous frameworks or sensors, their capacity is done either in a concentrated technique or disseminated strategy in a muddled or unstructured organization. Subsequently, information mining instruments are adequately utilized to change these unstructured or indistinguishable configurations of information which might be composite in nature to organized dialects which could be effectively translated by the end client even on a fundamental visual examination. Not at all like traditional information digging calculations for ordinary information, enormous information mining calculations must be methodically confined and worked with a few imperatives and limits being considered before mining the data. This audit has examine the writing in an expansive and comprehensive way and discoveries exhibited methodically in different segments with exceptional accentuation on group based procedures because of their discernible benefits and effortlessness in computational development and multifaceted nature. These findings could be really effective in identifying research formulations or refining the problem objective for future scope of research in these avenues.

REFERENCES

[1] Adams M N, "Perspectives of data mining", International journal of market research, Vol. 52, No. 1, pp. 11 – 19, 2010.

[2] Jong Youl Choi, Seung Hee Bae, Judy Qiu, Geoffrey Fox, Bin Chen and David Wild, "Browsing large scale cheminformatics data with dimensional reduction", proceedings of emerging computational methods for the life sciences workshop, 2010.

[3] Song Z, Kusiak A, "Optimizing product configurations with a data mining approach", International journal of product research, Vol. 47, No. 7, pp. 1733 – 1751, 2009.

[4] Bhoj Raj Sharma, Daljeet Kaur and Manju, "A review on data mining: Its challenges, issues and applications", International journal of current engineering and technology, Vol. 3, No. 2, pp. 695 – 700, 2013.

[5] Kriti Srivastava, Shah R, Valia D and Swaminarayan, "Data mining using hierarchical agglomerative clustering algorithm in distirbuted cloud computing environment", International journal of comptuer theory and engineering, Vol. 5, No. 3, pp. 520 – 522. 2013.

[6] Supreet Kaur, Usvir Kaur, "A survey on various clustering techniques with K means clustering algorithm in detail", International journal of computer science and mobile computing, Vol. 2, Issue. 4, pp. 155 – 159, 2013.

[7] Neelamadhab Padhy, Pragnyaban Mishra and Rasmita Panigrahi, "The survey of data mining applications and future scope", International journal of computer science, engineering and information technology, Vol. 2, No. 3, 2012.

[8] Navjot Kaur, Jaspreet Kaur Sahiwal and Navneet Kaur, "Efficient K means clustering algorithm using ranking method in data mining", International journal of advanced research in computer engienering and technology, Vol. 3, No. 3, 2012.

[9] Philippe Hanhart et al, "Benchmarking of objective quality metrics for HDR image quality assessment", EURASIP journal of image and video processing, Vol. 39, 2015.

[10] Seshadrinathan K, Soundararajan R, Bovik A C, Cormack L K, Study of subjective and objective quality assessment of video", IEEE transactions on image processing, Vol. 19, No. 6, pp. 1427 – 1441, 2010.

[11] Wang Z, Li Q, "Information content weighting for perceptual image quality assessment", IEEE transactions on image processing, Vol. 20, No. 5, pp. 1185 – 1198, 2011.

[12] Abdul Nazeer K A and Sebastian M P, "Improving the accuracy and efficiency of K means clustering algorithm", proceedings of the world congress on engineering, Vol. 1, 2009.

[13] Osama Abu Abbas, "Comparison of various clustering algorithms", International Arab journal of information technology, Vol. 5, No. 3, 2008.

[14] Napoleon D and Gangalakshmi, "An efficient K means clustering algorithm for reducing time complexity using uniform distribution data points", proceedings of IEEE international conference on trends in information sciences and computing, 2011.

[15] Don Kulasiri, Sijia Liu, Philip K Maini and Radek Erban, "DifFUZZY: A fuzzy clustering algorithm for complex data sets", International journal of computational intelligence in bioinformatics and systems biology, Vol. 1, No. 4, pp. 402 – 417, 2010.

[16] Kedar Sawant and Snehal Bhogan, "Iteration reduction K-means clustering algorithm", International journal of innovative science, engineering and technology, Vol. 3, No. 5, pp. 501 – 506, 2016.

[17] Madhu Yedla, Srinivasa Rao, Pathakota and Srinivasa T M, "Enhancing K means clustering algorithm with improved initial center", International journal of computer science and information technologies, Vol. 1, No. 2, pp. 121 – 125, 2010.

[18] Bhatia M P S and Deepika Khurana, "Experimental study of data clustering using K means and modified algorithms", International journal of data mining and knowledge management process, Vol. 3, No. 3, pp. 2013.

[19] Sumit Garg and Arvind Sharma K, "Comparative analysis of data mining techniques on educational dataset", International journal of comptuer applications, Vol. 74, No. 5, 2013.

[20] Ahamed Shafeeq B M, Hareesha K S, "Dynamic clustering of data with modified K means algorithm", Proceedings of international conference on information and computer networks, Vol. 27, pp. 221 – 225, 2012.

[21] Laurence Morissette and Sylvain Chartier, "K means clustering technique: General considerations and implementation in Mathematica", Tutorials in quantitative methods for psychology, Vol. 9, No. 1, pp. 15 – 24, 2013.

[22] Jyoti Yadava and Monika Sharma, "A review of K mean algorithm", Intenrational journal of engineering trends and technology, Vol. 4, Issue. 7, pp. 2972 – 2976, 2013.

[23] Freitas A.A. "A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery", In: Ghosh A., Tsutsui S. (eds) Advances in Evolutionary Computing. Natural Computing Series. Springer, Berlin, Heidelberg, 2003.

[24] Tan K C, Yu Q and Lee T H, "A distributed evolutionary classifier for knowledge and discovery in data mining", IEEE transactions on systems, man and cybernetics, Vol. 35, No. 2, pp. 131 – 142, 2005.

[25] Keogh E, Chakrabarti K, Pazzani M and Mehrotra S, "Dimensionality reduction for fast similarity search in large time series databases", Knowledge and informtion systems, Vol. 3, No. 3, pp. 263 – 286, 2001.

[26] Gogoi P, Bhattacharyya D K, Borah Ba nd Kalita J K, "A survey of outlier detection methods in network anomaly identification", The comptuer journal, Vol. 54, No. 4, pp. 570 – 588, 2011.

[27] Sun Y, J. Han, X. Yan, and P. S. Yu, "Mining knowledge from interconnected data: a heterogeneous information network analysis approach," in Proceedings of the VLDB Endowment, pp.2022–2023, 2012.

[28] Wu X, Zhu X, Wu G Q and Ding Q, "Data mining with big data", IEEE transactions on knowledge and data engineering", Vol. 26, No. 1, pp. 97 – 107, 2014.

[29] Tan K C, Teoh Jm Yu K and Goh C, "A hybrid evolutionary algorithm for attribute selection in data mining", Expert systems with applications, Vol. 36, Issue. 4, pp. 8616 – 8630, 2009.

[30] Xiao Feng Yin, Li Pheng Khoo and Yih Tng Chong, "A fuzzy C means based hybrid evolutionary approach to the clustering of supply chain", Computers and industrial engineering, Vol. 66, Issue. 4, pp. 768 – 780, 2013.

[31] Pooya Daie and Simon Li, "Hierarchical clustering for structuring supply chain network in case of product variety", Journal of manufacturing systems, Vol. 38, pp. 77 – 86, 2016.

[32] Suman and Pink Rani, "A survey on STING and CLIQUE grid based clustering methods", International journal of advanced research in computer science, Vol. 8, No. 5, pp. 1510 – 1512, 2017.

[33] Jyoti Yadav, Dharmender Kumar, "Subspace clustering usign CLIQUE: an exploratory study", International Journal of advanced research in computer engineering and technology, Vol. 3, Issue. 2, 2014.

[34] Anne Patrikaienen and Marina Meila, "Comparing subspace clusterings", IEEE transactions on knowledge an data engineering, Vol. 18, No. 7, pp. 902 – 916, 2006.

[35] Lu Y., Sun Y., Xu G., Liu G., "A Grid-Based Clustering Algorithm for High-Dimensional Data Streams", In: Li X., Wang S., Dong Z.Y. (eds) Advanced Data Mining and Applications, Lecture Notes in Computer Science, vol 3584. Springer, Berlin, Heidelberg, 2005.

[36] Pradeep Rai and Shubha Singh, "A Survey of Clustering Techniques", International Journal of Computer Applications, Vol. 7,No. 12, pp. 1-5, 2010.

[37] Raymond T. Ng and Jiawei Han, "CLARANS: A Method for Clustering Objects for Spatial Data Mining", IEEE Transaction on Knowledge and Data Engineerring, Vol. 14,No. 5, 2002.

[38] Swarndeep Saket J and Sharnil Pandya, "An overview of partitioning algorithms in clustering", International Journal of advanced research in computer engineering and technology, Vol. 5, Issue. 6, pp. 1943 – 1946, 2016.

[39] Velmurugan T and Santhanam T, "A survey of partition based clustering algorithms in data mining: An experimental approach", Information technology journal, Vol. 10, No. 3, pp. 478 – 484, 2011.

[40] Hae Sang Park and Chi Hyuck Jun, "Simple and fast algorithm for K medoids clustering", Expert systems with applications, Vol. 36, pp. 3336 – 3341, 2009.