International Journal of Electrical Electronics & Computer Science Engineering
Volume 5, Issue 1 (February, 2018) | E-ISSN : 2348-2273 | P-ISSN : 2454-1222
Available Online at www.ijeecse.com

Review on the Cluster based Data Mining Techniques in Big Data

C. Krubakaran[1], Prof. (Dr.) K. Venkatachalapathy[2]
[1]Assistant Professor, Department of IT, Bharathiyar College of Engineering & Technology, Karaikal, India
[2]Professor, Department of Information & Computer Science, Annamalai University, Chidambaram, Tamilnadu, India
[1]kirubabcet2018@gmail.com, omsumeetha@rediffmail.com

Abstract: ***Recent decade has seen a great revolution in digital technology with state of art technologies and devices to handle, store and transmit information across wired and wireless media on a global basis. With increasing research in digital techniques, high definition media content are being observed in almost all commercial applications. High definition data may be an image, audio or video and give an in-depth detail about the subject under study. Real time utility of high definition data has invoked the concepts of big data and cloud computing in recent time. Extraction of useful knowledge from these data obtained from multiple sources forms the core concept of data mining. This review paper provides a comprehensive survey of recent and state of art techniques in mining useful information from big data especially using cluster computing approaches.***

Keywords: Data Mining, Cluster Computing, High Definition Multimedia, Cloud Computing, Big Data.

I. INTRODUCTION

Digital technology has undergone a rapid transformation and may be termed suitably as revolution with conventional data handling, storing and communication protocols and devices getting a complete makeover with recent techniques and strategies. For example, high resolution satellite imagery being used to forecast changes in the terrain, weather conditions etc. on a real time basis has invoked use of high definition image and video acquisition technologies and their continuous transmission to earth stations has necessitated suitable storage mechanisms for storing the bulk data. This has led to the advent of big data which is a compilation of composite data from multiple acquisition sensors and sources. A further derivative of above concept has paved the way for cloud computing where users could access data any time at any place on the globe through service providers. However, in spite of all the above perspectives, the objective of this review paper lies in dealing with data mining [7] or extraction of useful and meaningful information from the composite pool of data available in a storage database or cloud environment. Data obtained from multiple sources are available as a pool and systematic extraction of information through mining transforms the extracts into structured information providing much required clarity and understanding to the user. Further, the obtained information is considered meaningful as mining involves extraction by identifying suitable patterns which are similar to each other in the composite pool. Applications of data mining vary over a wide ranging starting from regression analysis to predictions of stock markets, forex services, anomaly [26] and intrusion detection in networks and systems, evaluation of project risk managements, training of intelligent networks and data 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.

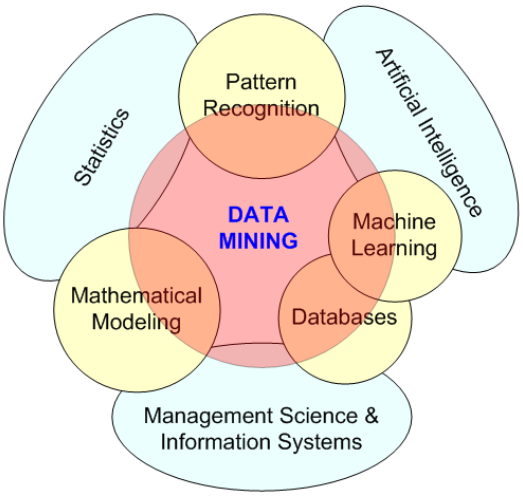

Fig. 1. Conceptual Illustration of Data Mining

As seen from the above figure, almost all data handling mechanisms are overlapped on a common data mining platform which interprets the outputs from each system into structured formats for further processing and understanding. Data mining is achieved through several methodologies like association based techniques [1], clustering based methods [19], prediction based techniques [16], decision trees [43], sequential patterns [18]. A systematic survey of literature keeping a strict focus towards recent trends in data mining based on cluster models have been presented in this paper.

A. Big Data:

Evolution of state of art data acquisition techniques have made data interpretation and consequent analysis and predictions to be more accurate and precise however at the cost of increasing volumes of data size. Big data [29] are characterized by increasing capacity which could be very well seen in case of a one minute medium clarity video taking up 10MB of memory with the same one minute video in high definition taking up to 150MB of memory utility.

International Journal of Electrical Electronics & Computer Science Engineering
Volume 5, Issue 1 (February, 2018) | E-ISSN : 2348-2273 | P-ISSN : 2454-1222
Available Online at www.ijeecse.com

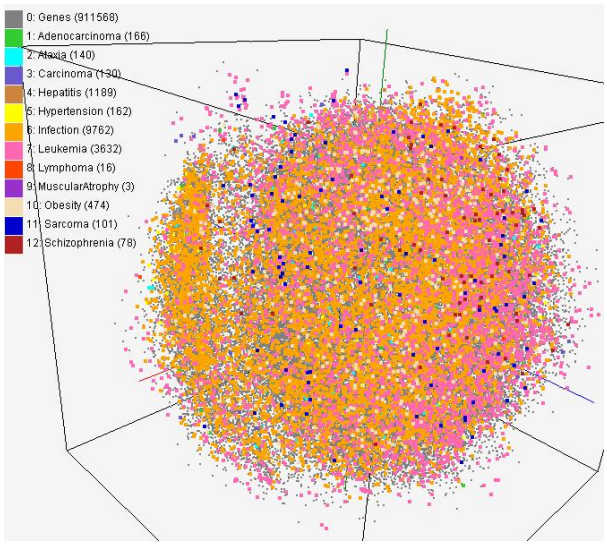

Fig. 2. A High Dimensional Data Model [2]

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 information clarity obtained from high definition data, big data has evolved and invoked research interests in large amounts in recent times. Efficient methods to handle, process, analyze and store big data are being researched from time to time [58] [59]. Big data are being generally characterized by three 'V's namely volume, velocity and variety [3]. They generally do not fit into the existing conventional data handling architectures like relational database management systems (RDBMS). Some of the effective tools for mining big data are KEEL, SPMF, Rattle, Weka, Orange etc., [4].

B. Issues in Data Mining from Big Data:

A systematic review of literature presents the following issues observed in mining data from big data. A basic flow process of a data mining process is illustrated in figure 3. The input data source which brings about the inputs are pre processed using suitable techniques [1] and given to pattern identification block where the mining process takes place resulting in knowledge discovery at the output. The outputs are sometimes referred to as knowledge discovery database (KDD).

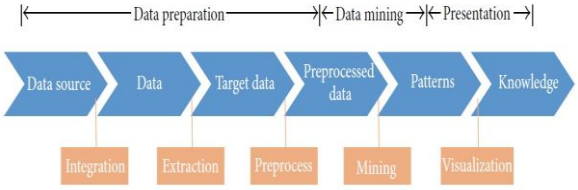

Fig. 3. Process Flow of Data Mining

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.

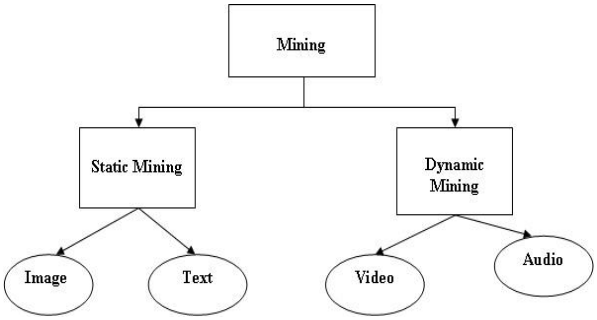

Fig. 4. Classification of Data Mining

For example, a purely image based data base mining method will not be able to accommodate the same extraction method for a pool of video data. The same applies in the case of a composite data source model. Hence the developed mining tool or algorithm which drives it should learn from the data source type and find a suitable method to handle the data. Prominent composite data sources may include videos, images, x – rays, portable format documents, graphics, voice mails, emails etc., the nature of data in big data systems could be a composition of both structured and unstructured components and hence handling these data types simultaneously poses to be a serious challenge. The second issue identified from the literature [4] lies in providing a suitable mining to user interface to improve the performance of the interpreted data. As depicted in figure 2, the clarity of the model is quite very poor which is greatly improved through strong mining tools. However, post mining process should have a clear and easily understandable user interface which enables the end user to understand the output of the mining tool in the least time possible. The mining tool should be able to provide appropriate data representation types for appropriate big data models.

Security [52] has been observed to be an issue of major concern in implementing an efficient data mining method since data mining normally extracts information from the unstructured pool of data into a more meaningful format which could be easily interpreted by the user. During this process, it is also to be noted that the probability of uncovering some vital, confidential and private information of individuals from the unstructured format into an understandable format are quite high thereby violating the confidentiality policies of the system. Hence the mining algorithm should be

International Journal of Electrical Electronics & Computer Science Engineering
Volume 5, Issue 1 (February, 2018) | E-ISSN : 2348-2273 | P-ISSN : 2454-1222
Available Online at www.ijeecse.com

designed such that it takes care of not disclosing private and confidential information during the mining process [53]. Another issue identified as a challenging task in mining is identification of suitable technique for mining big data. Conventional intelligence based and machine learning based methods are available in large numbers for mining normal data. However, the adaptability of this algorithm for handling big data usually of the ranges of tera bytes is quite a questionable and challenging issue. Literature also presents large number data reduction techniques like principal component analysis (PCA) [25] available for reducing the feature vector dimensions or the data size. This is usually achieved by removal of redundant information from the feature vector. However, in case of big data mining, the question of the pattern or logic behind which certain information should be excluded for reducing the input vector size is quite challenging. Elimination of information from big data without proper learning based rules or training methods could result in loss of vital data rendering the mining method to be inefficient [39].

Data mining involves extraction of information from a pool of input data. In normal real time applications, expecting that the pool of input data could be in a centralized place is not possible as the inputs may be distributed over several places and across several servers depending on the coverage area. For example, in case of satellite imagery, the inputs may be obtained from several sensors from different satellites and stored in several places distributed over several servers. Hence, the mining algorithm should be trained to work on data from distributed platforms.

Time of data mining [40] is finally an essential issue as handling of big data increases the time consumption in an exponential manner and hence a time efficient mining algorithm should be designed considering the above mentioned constraints.

II. Review On Cluster Based Mining Methods

Among the different mining methods available, this review paper concentrates on cluster based methods for mining of data. Several recent algorithms have been studied and findings are systematically presented in this section.

A comprehensive study of literature [13] indicates several clustering methods categorized into the below mentioned approaches.

- Hierarchical clustering
- Partition based clustering
- Grid based clustering
- Constraint based clustering
- Machine learning based clustering
- Scalable clustering
- Projection and subspace clustering for big data.

Hierarchical clustering techniques [5] [31] are based on clustering tree based approaches and these cluster trees are called dendrograms and are more suitable for small sample data sources. They are further broken down into bottom up and top down approaches with the former also known as agglomerative and the latter also known as divisive clustering. Further, the former is found to merge two or more similar clusters in consecutive iterations while the latter splits the similar clusters into more cluster trees. At each cluster stage k-means clustering [16] has been used. K means clustering algorithm has been more often found to be exploited for optimal mining applications in the literature [6] [18] [20] and works on the objective function of L2 norm obtained by normalizing the intra cluster distances specified as

$$O(c) = \sum \|c_i - c_j\| \quad (1)$$

Where c_i denotes the centroid of cluster *i*.

Two major merits of K means being extensively used in literature is that it allows parallelization and works effectively irrespective of data order. Several algorithms have been reported in the literature which has focused towards modifying existing K means algorithms to improve the mining efficiency. Two phase clustering methods [12] have been experimented in the literature where the first phase deals with computation of clusters in a systematic approach to generate clusters with preciseness. A ranking based K means algorithm [8] has been found to improve the speed of mining when compared to the existing cluster based K means which makes it ideal to be used for big data schemes. Experimental results indicate 476s time consumption for mining of up to 500 records using the ranking method with a 487s reported for the same record number using clustering with conventional K means. An improvement [14] in time response and reduction of complexity for mining has been achieved using computation of uniform data points in addition to K means clustering application to the mining system. A threefold algorithm have been presented in the literature [22] which addresses the issues of deciding upon the optimal number of clusters and removal of dead limit [21], reduction of computational and time complexity.

Grid based methods [32] have also been experimented in the literature by building a set of grid cells followed by computation of cell density. The cells below a predetermined threshold are eliminated. Based on minimizing an objective function, a cluster is created with adjacent similar groups. Sting and Clique algorithms [33] are prominently used in grid based clustering. From studies in the literature it is observed that STING is a query independent approach and parallelization is quite infeasible. On the other hand, CLIQUE algorithms permit cluster formations of any arbitrary shapes and generate clusters from dense subspaces [34] [35] with apriori approach. Partition

based approach have also been extensively researched in the literature [38] [39] where a set of *p* partitions are created initially followed by movement of objects from one group to another through an iterative relocation technique. The iterative relocation technique once again is found to K means, Fuzzy C means and K medoids [40].For big data, sampling using CLARA (clustering large applications) have been utilized along with K medoids in a hybrid combination to address scalability due to increasing data size. Advantage of K medoids observed in experimental findings from literature [41] indicates that they have a least computation time as the calculation of distance between pairs occurs only once unlike their K means clustering counterparts [22-24]. On the other hand Fuzzy C means comes under the category of soft computing techniques and works on the principle of minimization of objective function by initializing a membership matrix based on the selected number of clusters. The membership matrices are updated based on the computed cluster centers and the process terminates when the computed membership matrix values become less than a predefined threshold. Else the algorithm iterates towards minimizing the membership matrix.

Constraint based clustering algorithms [47] work on computing an initial solution dependent on some user defined constraints like constraints on individual objects, obstacle objects, clustering parameters etc., [48]. Machine learning based methods are quite extensively found in the literature [44] in the form of artificial neural networks [49] through extraction of symbolic rules technique which is a modification of existing ANN algorithm. More precision and accuracy have been observed in the experimental results using a four stage ANN training through a back propagation training rule. Also literature indicates that the conventional rule based ANN methods are not that precise in their clustering outputs and are also computationally expensive. A two stage ANN model found in the literature justifies a noise reduced clustering output as well as its capability to handle a range of messy data. Fuzzy based techniques [55] have been found to be extensively used to address the issues of privacy and security of mined data corresponding to confidential information of individuals in the pool of data. A privacy preserving data mining technique has been used in the research works in the literature with the attributes being converted into fuzzy or crisp values to preserve confidentiality of information [15] [54]. Genetic based algorithms [56] have also been investigated in the literature in a hybrid combination with K means clustering to find global optimal partitions of given data into clusters. Experimental analysis indicates that GA based techniques produce precise mining outputs with low intra cluster and high inter cluster distances and overcome the drawbacks of conventional K means clustering techniques of settling of a sub optimal solution and the need for predetermining the clusters [57]. The last of clustering based method involves projection and sub space clustering [60] which generates a large number of clusters and specific algorithms have been developed to remove the redundant clusters based on rough set theory to establish apriori property in the elimination process [61]. Evolutionary approaches [23] [29] [42] [62] include nature inspired algorithms or bio inspired algorithm like particle swarm optimization (PSO) [67], Ant colony optimization (ACO) and simulated annealing techniques [63] etc.

Other recent techniques towards handling big data with mining information from them include Hadoop [64] which is a open source fault tolerant system for big data storage and processing. It effectively handles the problem of distributed computing through Map Reduce [65 – 66] and tools like Hive, Mahout and Pig are used for handling the heterogeneous nature of data source. Map Reduce works on a clustering scheme and capable of processing large volumes of data on a parallel and distributed approach through two functions map and reduce.

III. Review Of Evaluation Metrics

SNR is used in data compression, and precision and recall are used in text-based information mining. Good metrics will lead the technique in the correct direction while bad ones may mislead the research effort. Currently, some image mining systems measure performance based on the "cost/time" to find the right matches. Others evaluate performance using precision and recall, terms borrowed from text-based retrieval. Although these criteria measure the system's performance to some extent, they are far from satisfactory. One major reason causing the difficulty of defining a good evaluation criterion is the perception subjectivity of data source content. 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.

$$precision = \frac{tp}{tp+fp} \quad (2)$$

$$recall = \frac{tp}{tp+fn} \quad (3)$$

$$accuracy = \frac{tp+tn}{tp+tn+fp+fn} \quad (4)$$

IV. Conclusion

A comprehensive review of cluster based mining algorithms for big data have been extensively presented in this review article. Big data is an evolving technology being implemented in cloud networks for access by clients on a global basis at any given point of time and anywhere in the globe. However, due to the large volumes of data which is characteristic of big data

systems which obtain inputs from multiple systems or sensors, their storage is done either in a centralized method or distributed method in a unreadable or unstructured format. Hence, data mining tools are effectively employed to transform these unstructured or unreadable formats of data which may be composite in nature to structured languages which could be easily interpreted by the end user even on a preliminary visual inspection. Unlike conventional data mining algorithms for normal data, big data mining algorithms have to be systematically framed and built with several constraints and bounds being considered before mining the information. This review has investigate the literature in a broad and exhaustive manner and findings presented systematically in various sections with special emphasis on cluster based techniques due to their noticeable merits and simplicity in computational construction and complexity. A brief overview of evolutionary computing methods for effectively mining data from composite data sources has also been presented with final conclusions given with a list of evaluation metrics needed to quantify the efficiency of the mining algorithm. These findings could be really effective in identifying research formulations or refining the problem objective for future scope of research in these avenues.

V. 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.

[41] Van der Laan, M. J., Pollard, K. S., & Bryan, J., "A new partitioning around medoids algorithm", Journal of Statistical Computation and Simulation, Vol. 73, No. 8, pp. 575–584, 2003.

[42] Mukhopadhyay A, U. Maulik, S. Bandyopadhyay, and C. A. C. Coello, "A survey of multiobjective evolutionary algorithms for data mining: part I", IEEE Transactions on Evolutionary Computation, Vol. 18, No. 1, pp. 4–19, 2014.

[43] Sumathi N, Geetha R, "Spatial data mining: Techniques, trends and its applications", Journal of computer applications, Vol. 1, No. 4, 2008.

[44] Sushmita Mitra, Sankar K. Pal and Pabitra Mitra, "Data mining in soft computing framework: A survey" IEEE transactions on neural networks, Vol. 13, No. 1, 2002.

[45] Kannan A, Mohan V, Anbazhagan N, "An effective method of image retrieval using image mining techniques", International journal of multimedia and its applications, Vol. 2, No. 4, 2010.

[46] Ying Mei Cheng and Sou Sen Leu, "Constraint based clustering and its applications in construction management", International journal of expert systems with applications, Vol. 36, Issue. 3, pp. 5761 – 5767, 2009.

[47] Anthony K H, Tung, Raymond T, Laks V, Lakshmanan S, Jiawei Han, "Constraint based clustering in large databases", proceedings of the eight international conference on data base theory, pp. 405 – 419, 2001.

[48] Kir Wagstaff, Claire Cardie, Seth Rogers, Stefan Schrodl, "Constrained K means clustering with background knowledge", Proceedings of the eighteenth international conference on machine learning, pp. 577 – 584, 2001.

[49] Parekh R, Yang J and Honavar V, "Constructive neural network learning algorithms for pattern classification", IEEE transactions on neural netwokrs, Vol. 11, pp. 436 – 451, 2000.

[50] Cheng Ching Chang and SSu Han Chen, "A comparative analysis on artificial neural network based two stage clustering", Cogent engineering, Vol. 2, Issue. 1, 2015.

[51] Pradeep Kumar, Kishore Indukumari Verma and Ashish Sureka, "Fuzzy based clustering algorithm for privacy preserving data mining", International journal of business information systems, Vol. 7, No. 1, pp. 29 – 40, 2011.

[52] Atzori, M., Bonchi, F., Giannotti, F. and Pedreschi, D., "Anonymity preserving pattern discovery', The International Journal on Very Large Data Bases, Vol. 17, No. 4, pp. 703 – 727, 2008.

[53] Aggarwal, C.C., Yu, P.S. (Eds), "Privacy-preserving data mining: models and algorithms", Advances in Database Systems, Springer series, Vol. 34, No. 22, p.514, 2008.

[54] Domingo-Ferrer, J. and Vicen, T., "Fuzzy icroaggregation for microdata protection", Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol. 7, No. 2, pp.153–159, 2003.

[55] Neha Bharill, Aruna Tiwari and Aayushi Malviya, "Fuzzy based scalable clustering algorithms for handling big data using apache spark", IEEE transactionso n big data, Vol. 2, No. 4, 2016.

[56] Mamta Mor, Poonam Gupta and Priyanka Sharma, "A genetic algorithm approach for clustering", International journal of engineering and computer science, Vol. 3, Issue. 6, pp. 6442 – 6447, 2014.

[57] Maulik U and Bandyopadhyay, "Genetic algorithm based clustering technique", Pattern recognition, Vol. 33, No. 9, pp. 1355 – 1365, 2000.

[58] Chanchal Yadav, Shuliang Wang and Manoj Kumar, "Algorithm and approaches to handle large data: A survey", International journal of computer science and network, Vol. 2, Issue. 3, 2013

[59] Mehmet Koyuturk, Ananth Grama, and Naren Ramakrishnan, "Compression, Clustering, and Pattern Discovery in very High-Dimensional Discrete-Attribute Data Sets", IEEE Transactions On Knowledge And Data Engineering, Vol. 17, No. 4, 2005.

[60] Jun Wang, Zhaohong Deng, KupSze Choi, Yizhan gJiang, Xiaoqing Luo, Fu-
Lai Chung, Shitong Wang, "Distance metric learning for soft subspace clustering in composite kernel space", Pattern Recognition, Vol. 52, pp. 113-134, 2016.

[61] Singh Vijendra and Sahoo Laxman, "Subspace Clustering of High-Dimensional Data: An Evolutionary Approach", Applied Computational Intelligence and Soft Computing, Vol. 2013, 2013.

[62] Handl J and J. Knowles, "An evolutionary approach to multiobjective clustering", IEEE Transactions on Evolutionary Computation, Vol. 11, No. 1, pp. 56–76, 2007.

[63] Hamid Mohamadi, Jafar Habib , Mohammad Saniee, Abadeh Hamid Saadi, "Data mining with a

simulated annealing based fuzzy classification system", Pattern recognition, Vol. 41, No. 5, pp. 1824 – 1833, 2008.

[64] Barga R S, J. Ekanayake, W. Lu, "Project Daytona: Data Analytics as a Cloud Service" Proceedings of the International Conference of Data Engineering (ICDE 2012), IEEE Computer Society, pp. 1317–1320, 2012.

[65] Elsayed S, O. Ismail, and M.E. El-Sharkawi, "MapReduce: state-of-the-art and research directions", International Journal of Computer and Electrical Engineering, vol. 6, no. 1, pp. 34-39, 2014.

[66] Xu L, C. Jiang, J. Wang, J. Yuan, And Y. Ren, "Information Security in Big Data Privacy and Data Mining", IEEE Access, Vol. 2, pp 1149-1176, 2014.

[67] Amreen Khan, Bawane N G and Sonali Bod, "An analysis of particle swarm optimization with data clustering technique for optimization in data mining", Journal on computer science and engineering, Vol. 2, No. 4, pp. 1363 – 1366, 2010.

[68] Yi Jun Chen Man, Leung Wong, Haibing Li, "Applying ant colony optimization to configuring stacking ensembles for data mining", Expert systems with applications, Vol. 41, Issue. 6, pp. 2688 – 2702, 2014.

[69] Sreeja N K, A. Sankar, "A hierarchical heterogeneous ant colony optimization based approach for efficient action rule mining", Swarm and Evolutionary Computation, Vol. 29, pp. 1-12, 2016.