DOI: http://dx.doi.org/10.26483/ijarcs.v8i7.4032

ISSN No. 0976-5697

Volume 8, No. 7, July – August 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

DATA MINING APPROACH FOR BIG DATA ANALYSIS:A THEORITICAL DISCOURSE

Mudassir Makhdoomi
Islamia College of Science & Commerce
Srinagar, Kashmir J&K-India.

Abstract: BIG data is one of the most emerging topic studied today. Everyone is talking about big data, and it is believed that science, business, industry, government, society, etc. will undergo a thorough change with the influence of big data. The volume of data being produced is increasing at an exponential rate due to our unprecedented capacity to generate, capture and share vast amounts of data. Existing algorithms can be used to extract information from these large volumes of data. However, these algorithms are computationally expensive. In this paper we are discussing some of the major challenges and issues posed by big data and the potential solution to those challenges.

Keywords: Big data analytics; Unstructured data; volume; veracity; velocity

I. INTRODUCTION

BIG data is an emerging field of research and one of the most debated topics nowadays. Big data will have a significant impact on science, business, industry, government, society etc. and they will undergo a thorough change in coming years. Big data is commonly explained through the facets of Volume, Velocity, Veracity, and Variety. It is assumed that either all or any one of them needs to be met for the classification of a problem as a Big Data problem. Due to incredible potential of computing & communication devices and our unprecedented capacity to generate, capture and share vast amounts of data, the volume of data being produced is increasing at an exponential rate. Today when data is streaming at rates faster than that can be handled by traditional algorithms & the quality of data is still a major concern which needs to be either tackled at the data pre-processing stage or by the learning algorithms. In addition to the quantity aspect of data, another major concern is that large variety of data is available for a given object under consideration and to select the relevant information from this enormous data is challenging.

In order to extract information from these large volumes of data, one can make use of already data mining techniques like clustering, classification, decision trees, neural networks etc. However, the computational requirements of these techniques are usually proportional to the amount of data being processed and are therefore computationally expensive. Most of the time the algorithms processing large volumes of data demand prohibitive computational resources. Thus relying on existing techniques remains no option as the problems become increasingly challenging and demanding. Therefore, demand for optimized techniques powered by advanced computing capabilities to manipulate and explore big data increases.

II. BIG DATA CHARACHTERISTICS

Big Data is characterized by the huge-volume, heterogeneous, autonomous sources with distributed and decentralized control, and it seeks to unravel the complex and evolving relationships among data [1]. These are the major characteristics of big data thus making it a tough challenge for finding useful patterns from the Big Data.

One of the main characteristics of the Big Data is the huge volume of data represented by heterogeneous and varied dimensionalities. This is because different users collect data

using different schemas, and the nature of different applications also results in different representations of the data. Conventional database systems are not adequate to handle the big data as it exceeds their processing capacity. In order to gain value from this data, we must choose an alternative way to process it. Some of researchers have defined big data as “Big data exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it within a tolerable elapsed time for its user population”. Another good definition is given as “Big data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze”[2]. These definitions suggest that big data is evolving and is changing as technology is advancing over time and that the meaning of big data vary depending on the organization where it is used, the tools and technologies that are used by that organization [3]. So, we can say that what historically big data was or what is big data today won’t be big data tomorrow.

III. CHALLENGES OF BIG DATA

Some of the striking characteristics of big data are high dimensionality, heterogeneity and large sample size. These features raise several unseen challenges like high dimensionality brings noise accumulation, spurious correlations, and incidental homogeneity and if high dimensionality is combined with large sample size that incurs heavy computational cost and algorithmic complexities. The enormous samples in Big Data are typically combined from multiple sources at different times and technologies. Privacy and data ownership of data is another new challenge faced by users. Apart from these Timeliness of data is another main

characteristic of data which has been always considered crucial. These create issues of heterogeneity, variations in experiments, biases in data analysis, and thus calls for development of more adaptive and robust procedures [4].

The first thing that comes in mind when one thinks about Big Data, is its enormous size. Management of this enormous and rapidly increasing volume of data has been a challenging issue from the beginning. These challenges were mitigates since past by technological advances like making faster processors. But current scenario is changing because volume of data is increasing many folds faster than CPU speeds and other resources. Due to limitations of power, clock speeds have largely been constrained and processors are being built with increasing numbers of cores. Parallel processing techniques that have been used in past for data processing cannot be directly applied for intra-node parallelism, since the architecture looks very different [5].

Another major challenge is Timeliness, as data grow in volume, we need real-time techniques to summarize and filter what is to be stored, since in many instances it is not economically viable to store the raw data. The main challenge is to provide quick response time to complex queries at scale over high-volume streams. Another common problem is to find data patterns in huge datasets that satisfy certain conditions. This type of search is likely to occur repeatedly during data analysis. But this is quite impractical to scan the entire dataset repeatedly to find suitable data elements. Rather index structures are created to improve search process.

One of the biggest challenges is that of data Privacy and data ownership. Managing privacy is of prime importance in current scenario, as it is not just a matter of technical importance but also an issue of sociological importance which must be addressed. A simple example is that of location based services which collect data about user location by requiring them to share it with their service providers. There are numerous privacy concerns in these types of situations where users are made vulnerable. Such information can be used by attackers to track down the identity of the user. Several other types of critical private information like religious preferences, health issues, buying preferences can also be revealed by observing users movement and usage patterns over time. Such information is believed to reveal the people's identities as the movement of patterns and person's identity is usually correlated. Today another issue is that most of the services online require its users to share private information and most of the users are unaware about the concept of sharing of data and privacy control and know nothing beyond record-level access. In addition, data is dynamically changing and is getting enormous with each passing day; none of the existing techniques result in any useful content being released in this situation. Privacy is but one facet of data ownership.

In order to get the full benefit of the potential of big data, we need to consider the scale of data not just from the perspective of system but also by the perspective of humans. We have to ensure that humans can comprehend it and are not overwhelmed by this ocean of data. In spite of the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a difficult time finding. Ideally, analysis of big data will not be all complex computations rather it will be designed in such a way to have humans involved and in the loop. The incorporation of the concept of visual analytics in to big data analysis is attempting to achieve this goal. Big data analysis should incorporate human input at each and every stage because it usually takes multiple experts from different domains to actually comprehend the situation at hand. It should be possible to accept inputs from multiple distributed experts as it is too expensive to assemble all the experts at one place and support collaborative inputs. This process will involve not just sharing of data but also the algorithms, techniques and different experimental results. Systems with high visualization capabilities should be used as it is important to make users understand the results of the queries in such a way that are best understood in the particular domain and are at the right level of detail. Early system provided users with tabular presentations but today analysts need to present results with the help of powerful visualization which present results in a lucid way, which help in user understanding and collaboration. Complex analysis of data like drill down, rollup, slice and dice operation should be matter of few clicks for users and understand its provenance. Various popular methods can be used to harness human ingenuity problems like crowd sourcing, rating artifacts, leader-boards etc.[6] .

Meanwhile levels of heterogeneity in type, structure, semantics, organization, granularity, and accessibility are other major obstacles faced while developing big data application. Also, there is a high level of repetition of data in datasets. Reduction of redundancy and data compression is effective in indirect cost reduction and the potential values of the data are not affected. Due to unprecedented growth of data compared with the relatively slow advances of storage systems we are facing lot of challenges, one being the current storage system are not capable of handling such enormous volume of data. Therefore, for the big data analytics we need to form some principles regarding the storage of data as to which data shall be stored and which data is to be discarded. Since most of the existing database systems lack the flexibility in terms of expansion thus they could not meet the performance requirements needed for big data. Most of the big data service providers cannot effectively maintain and analyze such enormous amount of data because of the limitations of capacity and professional tools to handle such data. Also as the volume of data increases the analytical demands like the storage, processing and transmission of data will definitely consume more and more amount energy [7]. Therefore, we need to ensure expandability and accessibility by establishing system-level power consumption control and management mechanisms. The algorithms that are used for big data analysis must effectively be able to process increasingly expanding and more complex datasets. We need to have experts from varied

fields for carrying out an interdisciplinary research to harvest the potential of big data. Comprehensive big data network architecture must be established to help scientists and engineers in various fields access different kinds of data and fully utilize their expertise, so as to cooperate to achieve the analytical objectives [8].

IV. EXISTING DATA MINING TECHNIQUES

In general, data mining is a process that is used to search through huge collection of data in order to find useful patterns in data. The goal of this technique is to find patterns that were hidden and previously unknown. Once this useful data is unraveled it can be used to make decisions for development. There are three main steps in Data Mining process: Exploration, Pattern identification and Deployment.

Various algorithms and data mining techniques are used for knowledge discovery from huge collections of data, like: Classification, Regression, Clustering, Association rules, Decision trees etc. Some of them are briefly explained below:

A. Classification

Classification is one of the most commonly used data mining technique which uses a set of pre classified examples to develop a model which can then be used to classify new unclassified data records. The classification process involves two steps: learning and classification. In the learning step data is analyzed by classification algorithm. In classification step test data is used to estimate the accuracy of the rules of classification. Several techniques have been developed for classification and the most popularly used technique for solving real world problems is classification using decision trees, Bayesian classification, support vector machines (SVM),classification based on Association. Another very popular classification technique is by neural networks. These have been used successfully for classification because of their ability to classify complex data. Naive Bayes classifier is a straight-forward probabilistic classifier that applies Bayes theorem and assumes strong independent relationships among the features. K-nearest neighbor is another popular classification technique, which uses the common identification of the nearest neighbor's technique to assign a data item to the class [9].

B. Regression

Regression, which is a supervised mining that can be adapted for prediction of a numerical target. It can be used to model the relationship between one or more independent variables and dependent variables. Regression model evaluated the target value in terms of a function of each data item's predictors. The relationship between the target value and the predictors are then formulated in a model that can be applied to various data sets with unknown target values. There are various types of regression methods like, linear regression, Multivariate Linear regression, Non-linear Regression and Multivariate non-linear regression.

C. Association Rules Mining

Association rule mining is uncovers the relationships of interest and importance between variable in a dataset, i.e. to find frequent item set findings among huge collection of data. This type of analysis helps in decision making, analyzing behavior of customers etc. It has several applications like it can be used as a guide for placing different products inside a store in such a way that it increases sales, information regarding people vising websites and their interests, or to study and discover new relationships among biological data. Some of commonly known types of association rule mining are: Multilevel association rule, Multidimensional association rule and Quantitative association rule etc.

D. Clustering

Clustering is an unsupervised data mining techniques which is used to find similarity among data object and to group them based on that. Grouping is done in such a way that data points in a that belongs to same cluster is more similar to each other than the data that belongs to different clusters. Different types of clustering techniques have been developed like the crisp clustering techniques and fuzzy clustering techniques. Generally clustering algorithms are classified as partitioning clustering algorithms and Hierarchical clustering. K-means is one of the most popularly used clustering algorithms, which uses the partitioning approach to cluster the data into pre specified number of clusters.

E. Anomaly Detection

Anomaly detection is used to identify the outlier points in a data distribution. Outlier points are those points which are considerably different from the rest of data. For example, credit card fraud detection, network intrusion detection, and fault detection etc. are some of the applications for outlier detection techniques. These techniques identify the normal behavior of overall data points and detect the data points which behave differently. There are several types of outlier detection methods which can be divided between univariate methods, multivariate methods, parametric methods, non-parametric methods, graphical-based, statistical-based, the distance-based etc.

F. Rough Sets Analysis

Rough sets analysis is a classification data mining approach which is mainly concerned with the analysis of uncertain and incomplete information. As we know most of the real data sets are complex and usually uncertain and incomplete, so these can be analyzed under this classification scheme. Mathematical computation are used to explore hidden patterns in data in case of Rough sets. Rough sets have several applications like they can be used for data reduction, feature selection and extraction, and generation of decision rules.

V. OPTIMIZATION OF EXISITNG TECHNIQUES FOR BIG DATA ANALYSIS

Optimization is the process of improving the existing techniques by finding the highest achievable performance, by further improving the desired factors and try to reduce the undesired ones. Soft computing methods are most commonly used techniques for optimization of data mining methods and search problems. Data mining methods are used as data preprocessing tools where data can be processed and cleaned from outliers. Then optimization techniques can be applied to get best possible solutions.

Data mining performance is usually influenced by several factors, such as missing values, presence of noise, presence of outliers etc. As discussed earlier there are several characteristics of big data which makes data mining challenging. According to change of volume, velocity, variety and veracity of data extensions need to be introduced in existing data mining techniques because big data mining differs from simple data mining in the way that it needs not just finding of patterns but also involves large scale storage and processing of datasets. Soft computing approach to data mining can be seen as a potential extension of existing data mining techniques for application in big data analysis.

Soft computing techniques are new optimization techniques in artificial intelligence which exploit the tolerance for imprecision i.e. these methods exploit the capability of computers to search huge amount of data in a fast and effective manner. Some of the soft computing methods which can be used to optimize existing data mining techniques are:

A. Fuzzy logic

Fuzzy Logic is a logic system for reasoning that is approximate rather than exact. The fundamental unit of a fuzzy logic is the fuzzy set. Given the universal set X in order to define a fuzzy set A on X, we define a membership function A: X -> [0,1] that maps element x of X into real numbers in [0,1]. A(x) is interpreted as the degree to which x belongs to the fuzzy set A. We sometimes write fuzzy set A as $\{(x, A(x))| x \in X\}$. According to the classical set theory an element is either a member of the set or does not belong to the set but the fuzzy theory allows the gradual assessment of the membership of an element to the set. The fuzzy set theory allows the partial set membership. A classical or crisp set, then, is a fuzzy set that restricts its membership values to {0, 1}, the endpoints of the unit interval. Fuzzy logic has been used effectively for data mining tasks, mainly because of its capability to represent imperfect information, for instance by means of imprecise categories, measures of resemblance or aggregation methods [10]. The methods based on fuzzy logic that are used for data mining are: fuzzy decision trees, fuzzy prototypes and fuzzy clustering [11].

The importance of the use of fuzzy systems for big data analysis is unswerving. Fuzzy approach h provides a better depiction of the problem space by making used of fuzzy labels, and this ability makes it a very efficient approach when both the volume and variety of the dataset increases. The use of linguistic labels also increase the interpretability of data. Focus must be shifted towards the re designing of the state of art algorithms for incorporating fuzzy modeling approach. By incorporating this complete library of methods would be available for experts to completely exploit this novel work area.

Some of the characteristics of fuzzy methods like the management of uncertain and noisy data make them an invaluable tool for Big Data mining tasks. However, their current development is still at in infancy and we need to explore it more. They allow at managing big datasets without damaging the classification accuracy and providing fast response times. Fuzzy models provide a wide amount of advantages when applied to Big Data problems.

B. Neural networks

Artificial neural networks are inspired by the biological nervous systems. Neural network consists of large number of interconnected processing elements working as a single unit to solve specific problem. Artificial neural networks is used for several applications like pattern recognition, data classification etc. One of the important features of these networks is their adaptive nature where learning by example is the main strategy used for solving problems. The problems where little or incomplete information is available this feature of learning make it very appealing in such cases. Artificial neural network is self-adaptive approach as opposed to rest of the methods. This rapid use of ANN in wide variety of applications has been attributed to its ability of solving problems with relative ease of use, robustness to noisy input data or outliers, execution speed and the possibility of being implemented in parallel. Artificial neural network have been effectively employed on varied complex and extraordinary problem domains, such as pattern recognition, classification, signal processing, image processing, robotics, weather predictions, medical diagnosis, stock market analysis, financial forecasting etc. Learning is a basic and essential characteristic of ANN. The ability to learn from the experience through network examples and to generalize the captured knowledge, and to self-update in order to improve its performance is known as learning [12]. During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct response of the input patterns. Recently new concept of deep neural networks has emerged. Deep neural networks (DNNs) and their learning algorithms are known as the most successful methods for big data analysis. Compared with traditional methods, deep learning methods use data-driven and can extract features (knowledge) automatically from data. Deep learning methods have significant advantages in analyzing unstructured, unknown and varied model and cross field big data. At present, the most widely used deep neural networks in big data analysis are feed forward neural networks(FNNs). They work well in extracting the correlation from static data and suiting for data application scenarios based on classification. But limited by its intrinsic structure, the ability

of feed forward neural networks to extract time sequence features is weak. Infinite deep neural networks, i.e. recurrent neural networks (RNNs) are dynamical systems essentially. Their essential character is that the states of the networks change with time and couple the time parameter. Hence they are very suit for extracting time sequence features. It means that infinite deep neural networks can perform the prediction of big data. If extending recurrent structure of recurrent neural networks in the time dimension, the depth of networks can be infinite with time running, so they are called infinite deep neural networks.

C. Evolutionary algorithms

Evolutionary algorithms (EAs) are those techniques which take inspiration from the biological models of evolution and natural selection. The field of Evolutionary Computation consists of several types of evolutionary algorithm. These include *Genetic Algorithms* (GAs), *swarm based approaches etc*.

D. Genetic Algorithm

Genetic algorithm (GA) is a heuristic based search method that is imitates the process of natural evolution. Genetic algorithms are the class of evolutionary algorithms, which provide solutions to optimization problems using techniques that are inspired by natural evolution, such as inheritance, mutation, selection, and crossover. In Genetic Algorithms we form a population of candidate solutions towards an optimal solution. Genetic Algorithms implement the law of survival of the fittest to optimize the candidate solutions. The technique of GA progresses in the following manner:

1. First step is creation of the initial population of candidate solutions.
2. A fitness value is assigned to each individual from the population using fitness function.
3. Fitness is evaluated and then parents are selected.
4. Reproduction operators: crossover, mutation and selection on parents are used for creating offspring.
5. Based on the fitness evaluation of offspring's they are selected and new population is created.
6. Steps 3, 4, 5 are repeated until a termination condition is met

Genetic algorithms provide a comprehensive search methodology for machine learning and optimization. Algorithm is started with a population-a set of solutions represented by chromosomes. These Solutions from one population are used to form a new population. It is assumed that the new population will be better than the old one. Solutions which are selected to form new solutions (offspring) are selected according to their fitness the more suitable they are the more chances they have to reproduce [13]. The Genetic Algorithm is used in those domains where outcome is unpredictable and the process of outcome contains complex inter related modules. Also they are suitable for those problems where problem specification is very difficult to formulate. Using Genetic algorithms for data mining creates great robust, computationally efficient and adaptive systems. Genetic algorithms can be combined with various data mining techniques like clustering, association rules, classification etc to optimize the big data analysis [14].

E. Swarm based Approaches

Since 1990's, several collective behavior inspired algorithms meaning those algorithms which emulate the behavior of social insects, bird flocking etc. have been proposed. Swarm based techniques have several applications like: Traveling Salesman Problem, Quadratic Assignment Problem, Graph problems, network routing, clustering, data mining, job scheduling etc. Particle swarm Optimization (PSO), Bee swarm optimization (BSO), Ant Colonies Optimization (ACO) are currently the most popular algorithms in the swarm intelligence domain [15].

PSO is a population-based computational method that is used for optimization of a problem iteratively by first initialization with a population of random solutions called particle and then by improving candidate solution. All particles have fitness values associated with them which are then evaluated by the fitness function to be optimized, and have velocities which direct the flying of the particles. This was first designed to simulate the behavior of flock of birds seeking food. In PSO each particle is also associated with a velocity and they fly through the search space with velocities which can be dynamically adjusted according to their behaviors. The search process improves over the time as the particles have the tendency to fly towards the better and better search area. This technique has the ability to learn from the scenario and uses it to solve the optimization problems. PSO searches for optima by updating each generation of particles [16].

Ant Colonies Optimization (ACO) algorithms emulates the behavior of ants seeking a path between their colony and a source of food. ACO performs a model based search. They are interested mainly in the colony survival rather than individual survival. Ants search food by exploring the surrounding area of nest in a random manner. While moving, ants leave a chemical pheromone trail on the ground. Ants are guided by pheromone smell. Ants tend to choose the paths marked by the strongest pheromone concentration. When an ant finds a food source, it evaluates the quantity and the quality of the food and carries some of it back to the nest. During the return trip, the quantity of pheromone that an ant leaves on the ground may depend on the quantity and quality of the food. The pheromone trails will guide other ants to the food source. The indirect communication between the ants via pheromone trails enables them to find shortest paths between their nest and food sources [17].

There are two different kinds of approaches through which swarm intelligence can be applied to data mining techniques [17]. The first approach searches for solution by making

individuals of a swarm move through a solution space for the data mining task. This search based approach is used to optimize data mining techniques e.g., the parameter tuning. The second approach data instances that are placed on a low-dimensional feature space are moved in order to come to a suitable clustering or low dimensional mapping solution of data. This data organizing approach is applied directly to the data samples, e.g., dimensionality reduction of the data. Multi-objective and single objective problems can be solved with the swarm intelligence, especially particle swarm optimization or ant colony optimization algorithms. The two characteristics of particle swarm i.e. the self-cognitive and social learning; make them suitable for data clustering techniques, clustering high-dimensional data, semi-supervised learning based text categorization, and the Web data mining [18]. Several solutions can exist at the same time in case of swarm intelligence. Also due to the solutions getting clustered very fast premature convergence may happen. The soft computing and data mining techniques can be combined to produce results much better and beyond what either method could achieve alone. Swarm optimizations techniques can be applied to big data but there are some problems which need to be handled like handling dynamical data, handling high dimensional data, large scale optimization, multi-objective optimization etc. Since big data analytics is required to manage immense amounts of data and as the dimension of data and the number of objective of problems increase the complexity of problems also increase. Big data involves high dimensional problems and a large amount of data. Swarm intelligence studies the collective behavior in a group of individuals. It has shown significant achievements on solving large scale, dynamical, and multi-objective problems. With the application of the swarm intelligence, more rapid and effective methods can be designed to solve big data analytics problems.

VI. CONCLUSION

Big Data Analytic has emerged as potential area of research in the field of Computer Science and Information Technology. The exponential increase in the volume of data from different sources has prompted many researches to analyze such data and explore valuable information out of it. The large volume of data from different sources with varied characteristic features poses certain challenges so far as the acquisition, storage and processing of such data is concerned. The already available conventional data processing tools suffer from plenty of limitations in handling Big Data and therefore demand application of alternative data processing techniques to analyze such data for discovery of useful information form it. The already existing techniques however could be complimented by application of soft computing techniques in order to address most of challenges posed. Since Big Data continues to evolve in an unpredictable manner and most of the data is noisy, highly interrelated, heterogeneous and unreliable, it is likely that the performance of data mining techniques will remain sensitive to its unique characteristics.

VII. REFERENCES

[1] Prem, A., & Jayanthi, P. “Big data sources and data mining”. International Education and Research Journal,(2016) 2(4).

[2] Zicari R , “Big Data: Challenges and Opportunities” Akerkar R (ed) Big Data Comput. Chapman and Hall/CRC (2013), p 564.

[3] Che, D., Safran, M., & Peng, Z..” From big data to big data mining: challenges, issues, and opportunities”. In International Conference on Database Systems for Advanced Applications (2013) (pp. 1-15). Springer Berlin Heidelberg

[4] Fan, Jianqing, Fang Han, and Han Liu, "Challenges of big data analysis." National science review (2014) 1.2: 293-314.

[5] Dileep Kumar G., Manoj Kumar Singh,” Effective Big Data Management and Opportunities for Implementation” IGI Global(2016) ,ISBN: 9781522501824.

[6] Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. “Big data and its technical challenges”. Communications of the ACM,(2014) 57(7), 86-94.

[7] Chen, M., Mao, S., Zhang, Y., & Leung, V. C. “Big data: related technologies, challenges and future prospects “.(2014),Heidelberg: Springer pp 2-9.

[8] Chen, M., Mao, S., & Liu, Y. “Big data: A survey”. Mobile Networks and Applications, (2014),19(2), 171-209.

[9] Dina Fawzy1, Sherin Moussa and Nagwa Badr , “The Evolution of Data Mining Techniques to Big Data Analytics: An Extensive Study with Application to Renewable Energy Data Analytics”,(2016) Asian Journal of Applied Sciences (ISSN: 2321 – 089) Volume 04 – Issue 03.

[10] Rokach, Lior. "The Role of Fuzzy Sets in Data Mining." Soft Computing for Knowledge Discovery and Data Mining. Springer US, 2008. 187-203.

[11] Rekha, M., and M. Swapna."Role of fuzzy logic in data Mining." International Journal of Advance Research in Computer Science and Management Studies 2 (2014): 12.

[12] Abdul Hamid, Norhamreeza.”The effect of adaptive parameters on the performance of back propagation." (2012), PhD diss., Universiti Tun Hussein Onn Malaysia.

[13] Bouzouita, Ines, et al. "A Comparative Study of a New Associative Classification Approach for Mining Rare and Frequent Classification Rules." (2011),International Conference on Information Security and Assurance. Springer Berlin Heidelberg.

[14] Hans, Nivranshu, Sana Mahajan, and S. Omkar.,"Big data clustering using genetic algorithm on hadoop mapreduce." (2015), International Journal of Scientific Technology Research 4.

[15] Jevtic, A., Gazi, P., Andina, D. and Jamshidi, M.O.,”Building a swarm of robotic bees”. (2010) In World Automation Congress (WAC),(pp. 1-6). IEEE.

[16] Grosan, C., Abraham, A., & Nicoara, M. “Performance tuning of evolutionary algorithms using particle sub swarms”. In Symbolic and Numeric Algorithms for Scientific Computing, 2005. SYNASC 2005. Seventh International Symposium on (pp. 8-pp). IEEE.

[17] Abraham, A., Grosan, C., & Ramos, V. (Eds.). “Swarm intelligence in data mining”(2007),Vol. 34. Springer.

[18] Cheng, S., Zhang, Q., & Qin, Q.”Big data analytics with swarm intelligence”. Industrial Management & Data Systems (2016), 116(4) 646-666