Bio-inspired Algorithms for Big Data Analytics: A Survey, Taxonomy and Open Challenges

Sukhpal Singh Gill and Rajkumar Buyya

Cloud Computing and Distributed Systems (CLOUDS) Laboratory
School of Computing and Information Systems
The University of Melbourne, Australia
sukhpal.gill@unimelb.edu.au rbuyya@unimelb.edu.au

Abstract

Presently, various governments and organizations are focusing towards digitization of technical and academic documents, which overloads the digital libraries. However, it is difficult to manage a huge amount of data (big data) with current data processing techniques. In literature, bio-inspired algorithms based models and architectures are developed by various industry and academic groups to facilitate data analytics for big data. This chapter depicts a broad methodical literature analysis of bio-inspired algorithms for big data analytics. The current status of bio-inspired algorithms is categorized into three different categories: ecological, swarm-based and evolutionary. This chapter compares the existing models and architectures, explores the current trends and identifies the existing challenges in the development of big data analytical technique. This research work will also help to choose the most appropriate bio-inspired algorithm for big data analytics in a specific type of data along with promising directions for future research.

Keywords**: **Bigdata, Data Management, Bio-inspired Optimization, Big Data Analytics, Cloud Computing.

1.1 Introduction

Cloud computing paradigm utilizes the Internet to provide on-demand services to cloud users and emerged as a backbone of the modern economy. The emerging big data and Internet of Things (IoT) applications such as smart cities, healthcare services etc. are increasing, which needs fast data processing to improve the performance of computing systems [1]. However, these applications are facing large delay and response time because computing systems need to transfer data to the cloud and then cloud to an application, which affect its performance [2]. The data collected from different IoT devices have a large variety and volume (also known as Big Data), which also needs fog servers with high processing power. As a result of regular capturing and collection of datasets, they grow with the velocity of 250 MB/minute or more [3].

The continuous exchange of data in IoT environments is using for efficient decision making and real-time analytics for smart cities. Data is stored and processed on cloud servers after collection and aggregation of data from smart devices of IoT networks. Further, on-demand highly scalable cloud platforms are required to process the volume of data with large magnitude [4]. Cloud data processing cannot meet the requirements of an IoT application when low latency is required because sources of data are distributed across different sites [5]. Due to a large amount of data processing at the cloud, computing system does not process at the required speed which leads to communication failures. End devices are sending raw data continually to the cloud which makes cloud bottleneck [6]. Therefore, a bio-inspired algorithm based big data analytics is an alternative paradigm which provides a platform between end devices and cloud computing data centers to process user data in an efficient manner [7]. Big data analytics is proficient in filtering and processing the substantial amount of arriving data, creating the data processing architecture distributed and thus scalable.

1.1.1 Dimensions of Data Management

The literature reported that there are five different types of dimensions of data, which are required to manage effectively. Figure 1 shows the dimensions of data management for big data analytics: volume, variety, velocity, veracity and variability.

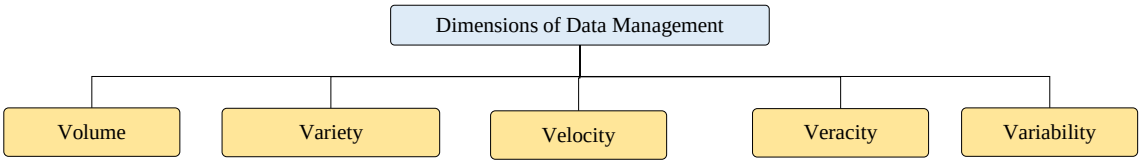

Figure 1: Dimensions of Data Management

The *Volume* represents the magnitude of data in terms of data sizes (terabytes or petabytes). For example, Facebook processes a large amount of data such as millions of photographs and videos. *Variety* refers to heterogeneity in a dataset, which can be different types of data. Figure 2 shows the variety of data, which can be text, audio, video, social, transactional, operational, cloud service or machine to machine data (M2M data).

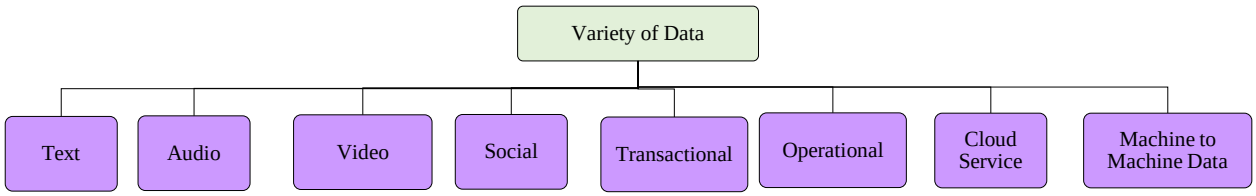

Figure 2: Variety of Data

Velocity refers to the rate of data generation and analysis for processing of a huge amount of data. For example, velocity can be 250 MB/minute or more [3]. *Veracity* refers to biases, noise and abnormality in data, while *variability* refers to the variation in the rate of data flow for generation and analysis.

The rest of the chapter is organized as follows. In Section 1.2, we present the big data analytical model. After that, we discuss the existing related studies in Section 1.3. Based on existing research work, we propose the taxonomy of bio-inspired algorithms for big data analytics. In Section 1.4, we analyse research gaps and present some promising directions towards future research in this area. Finally, we summarize the findings and conclude the chapter in Section 1.5.

1.2 Big Data Analytical Model

Big Data Analytics is a word, which is a combination of “big data” and “deep analysis” as shown in Figure 3. Every minute, a lot of user data is transferring from one device to another device, which needs high processing power to perform data mining for the extraction of useful information from the database. Figure 3 shows the model for big data analytics, which shows that an OLTP (On-Line Transaction Processing) system creates data (txn data). Data cube represents a big data, out of which required information can be extracted using data mining. Initially, different types of data are coming from different users or devices and the process of data cleansing is performed to remove the irrelevant data and stores the clean data into the database [12]. Further, data aggregation is performed to store the data in an efficient manner because incoming data contains a different variety of data and report for same is generated for easy use in future. The aggregated data is further stored in data cubes using large storage devices. For deep analysis, feature extraction is performed using data sampling, which

generates the required type of data. The deep analysis involves visualization of data, preparation of data, model learning (e.g. Decision Tree, Neural network/Bayesian network/Support vector machine/K-nearest-neighbour, Linear regression) and model evaluation [13].

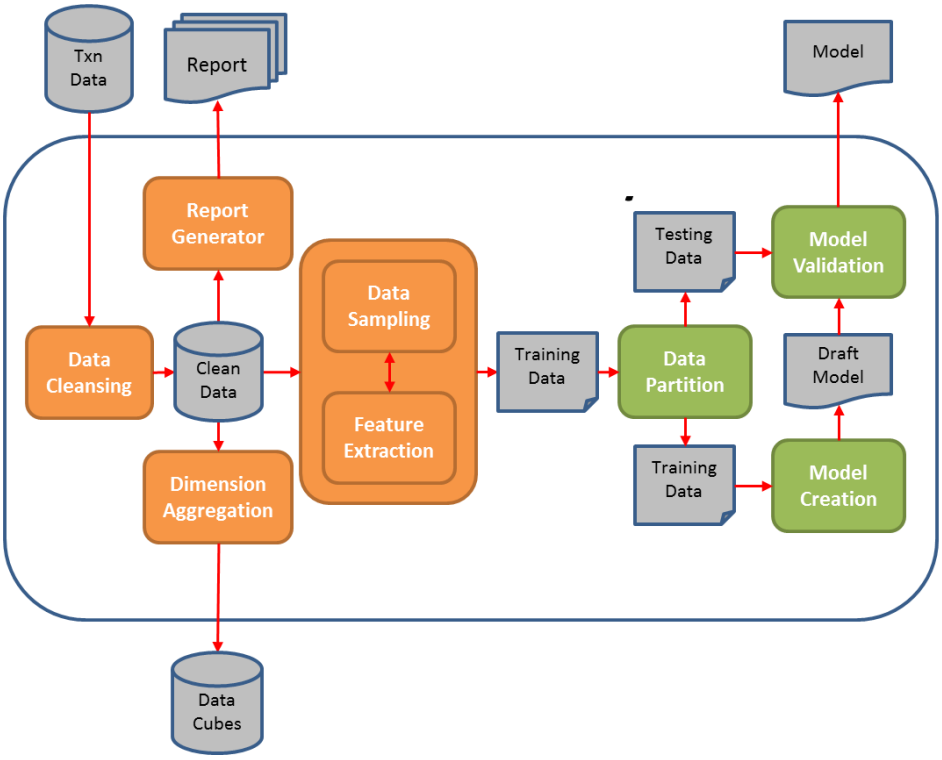

Figure 3: Big Data Analytical Model

Figure 4 shows the process of big data, which has two main components: data management and analytics. There are five different stages to process big data: 1) acquisition and recording (to store data), 2) extraction and cleaning (cleansing of data), 3) integration and aggregation (compiling of required data), 4) modeling and analysis (study of data) and data interpretation (represent data in required form).

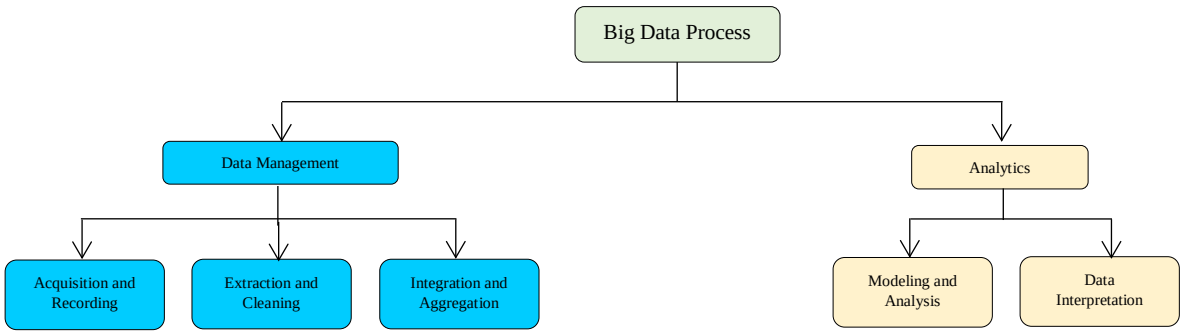

Figure 4: Big Data Process

1.3 Bio-inspired Algorithms for Big Data Analytics: A Taxonomy

This section presents the existing literature of bio-inspired algorithms for big data analytics. The bio-inspired algorithms for big data analytics are categorized into three different categories: ecological, swarm-based and evolutionary. Figure 5 shows the taxonomy of bio-inspired algorithms for big data analytics along with Focus of Study (FoS).

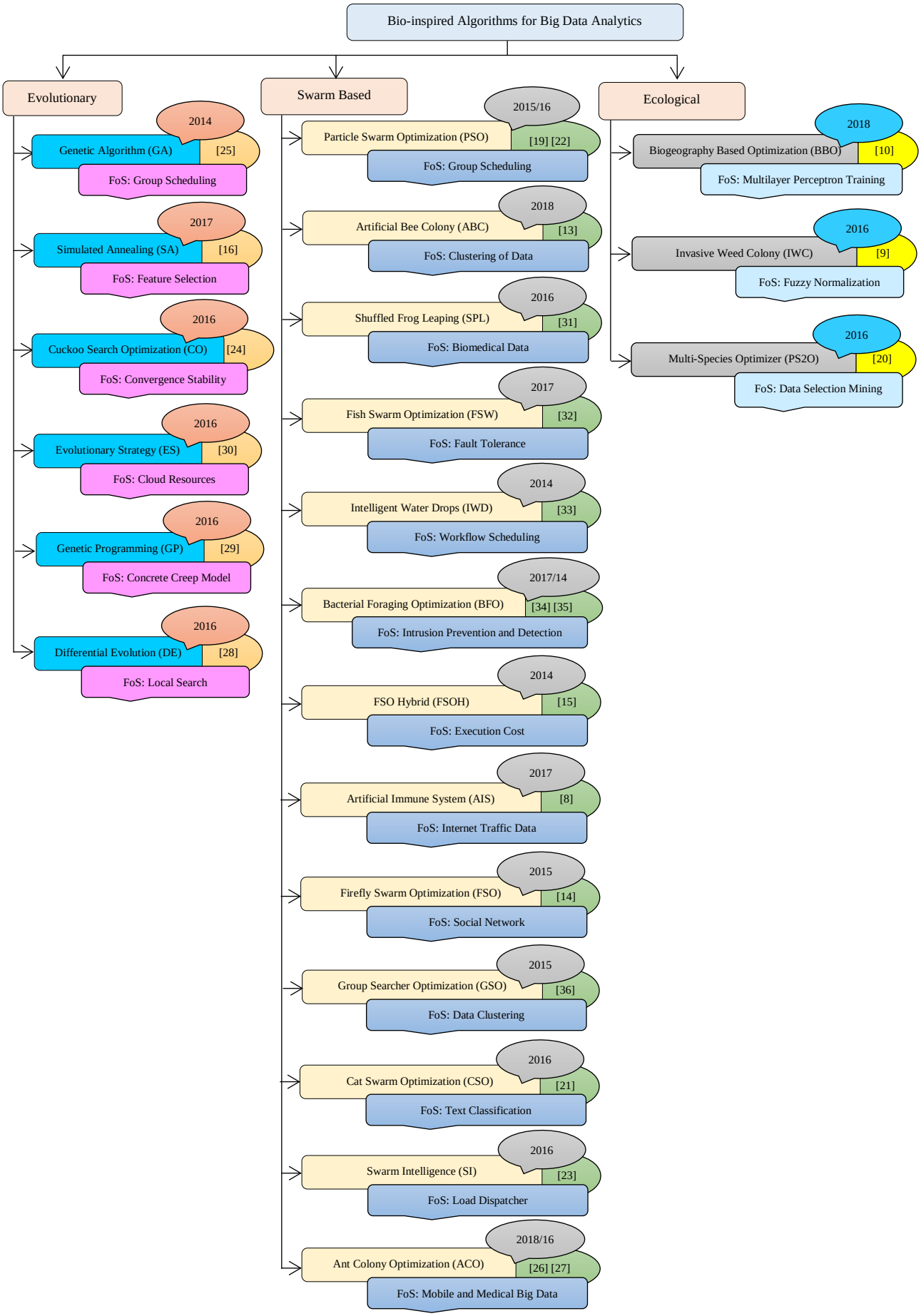

Figure 5: Taxonomy of Bio-Inspired Algorithms for Big Data Analytics

1.3.1 Evolutionary Algorithms

Kune et al. [25] proposed a Genetic Algorithm (GA) based data-aware group/family scheduling approach for big data analytics, which focuses on bandwidth utilization, computational resources and data dependencies. Moreover, GA algorithm decoupled data and computational services are provided as cloud services. Experimental results show that GA algorithm performs better in terms of turnaround time because GA algorithm processes data using parallel processing. Gandomi et al. [29] proposed multi-objective Genetic Programming (GP) algorithm based approach for big data mining, which is used to develop the concrete creep model to provide unbiased and accurate predictions. GP model works with high and normal strength concrete with a wide range of structural properties. Elsayed and Sarker [28] proposed a Differential Evolution (DE) algorithm based big data optimization approach, which uses local search to improve the exploitation capability of DE algorithm. This approach optimizes the big data 2015 benchmark problems with both multi and single-objective problems but it has large computational time. Kashan et al. [30] proposed an Evolutionary Strategy (ES) algorithm based big data analytics technique, which processes data efficiently and accurately using parallel scheduling of cloud resources. Further, ES algorithm minimizes the execution time by partitioning a set of jobs into disjoint groups, in which same resources process all the jobs in the same group.

Mafarja and Mirjalili [16] proposed a Simulated Annealing (SA) algorithm based big data optimization technique, which uses Whale Optimization Algorithm (WOA) to design different feature selection techniques to improve the exploitation by probing the most promising regions. SA algorithm helps to improve the classification accuracy and selects the most useful attributes for classification tasks. Further, Barbu et al. [17] proposed an SA algorithm based Feature Selection (SAFS) technique for big data learning and computer vision. Based on a criterion, SAFS algorithm removes variables and tightens a sparsity constraint, which reduces the problem size gradually during the iterations that makes it mainly fit for big data learning. Tayal and Singh [18] proposed FSO and SA based Hybrid (FSOSAH) approach for multi-objective stochastic dynamic facility layout problem using big data analytics. Saida et al. [24] proposed Cuckoo Search Optimization (CO) algorithm based big data analytics approach for clustering of data. Further, different datasets from the UCI Machine Learning Repository is used to test the performance of CO algorithm and performs better in terms of computational efficiency and convergence stability.

1.3.2 Swarm Based Algorithms

Ilango et al. [13] proposed an Artificial Bee Colony (ABC) algorithm based clustering approach for big data, which identifies the best cluster and performs the optimization for different dataset size. ABC algorithm approach minimizes the execution time and improves the accuracy. Map-reduce based Hadoop environment is used for implementation and experimental results show that ABC algorithm performs better than Particle Swarm Optimization (PSO) and Differential Evolution (DE) in terms of execution time. Raj and Babu [14] proposed a Firefly Swarm Optimization (FSO) algorithm for big data analytics for making new connections in social networks to compute the probability of staying in social network. In this technique, a mathematical model is introduced to test the stability of network and reduces the cost of big data management. Wang et al. [15] proposed an FSO algorithm based Hybrid (FSOH) approach for big data optimization, which focused on six multi-objective problems. It reduces execution cost but it has high computational time complexity.

Wang et al. [19] proposed a Particle Swarm Optimization (PSO) algorithm based big data optimization approach to improve online dictionary learning and introduced an atom-updating stage of the dictionary-learning model. PSO algorithm reduces the heavy computational burdens and

improves the accuracy. Hossain et al. [22] proposed a Parallel Clustered PSO algorithm (PCPSO) based approach for big data-driven service composition. PCPSO algorithm handles huge amounts of heterogeneous data and process data using parallel processing using MapReduce in the Hadoop platform. Lin et al. [21] proposed a Cat Swarm Optimization (CSO) algorithm based approach for big data classification to select features in a text classification experiment for big data. CSO algorithm uses the term frequency-inverse document frequency to improve the accuracy of feature selection.

Cheng et al. [23] proposed a Swarm Intelligence (SI) algorithm based big data analytics approach for economic load dispatch problem and SI algorithm handles the high dimensional data, which improves the data processing with accuracy. Banerjee and Badr [26] proposed Ant Colony Optimization (ACO) algorithm based approach for mobile big data using rough set. ACO algorithm helps to select an optimal feature for resolved decisions, which aids to manage big data of social network (tweets and posts) effectively. Pan [27] proposed Improved ACO algorithm (IACO) based big data analytical approach for management of medical data such as patient data, operation data etc., which helps doctors to retrieve the required data in a little span of time.

Hu et al. [31] proposed a Shuffled Frog Leaping (SFL) algorithm to perform the selection of the feature for optimized high-dimensional biomedical data. For high-dimensional biomedical data, SFL algorithm maximizes the predictive accuracy by exploring the space of possible subsets to obtain the set of features and reduces the irrelevant features. Manikandan and Kalpana [32] proposed a Fish Swarm Optimization (FSW) algorithm for feature selection in big data. FSO algorithm reduces the combinatorial problems by employing the fish swarming behaviour and this is effective for diverse applications. Social interactions among big data have been designed using the movement of fish and their movements for searching food. This algorithm performs effectively in terms of fault tolerance and data accuracy. Elsherbiny et al. [33] proposed Intelligent Water Drops (IWD) algorithm for workflow scheduling to manage big data effectively. The workflows simulation toolkit is used to test the performance of IWD algorithm and experimental results show that proposed algorithm is performed effectively in terms of cost and makespan as compared to FCFS, Round Robin and PSO algorithm.

Neeba and Koteeswaran [34] proposed a Bacterial Foraging Optimization (BFO) algorithm to classify the informative and affective content from the medical weblogs. MAYO clinic data is used as a medical data source to evaluate the accuracy to retrieve the relevant information. Ahmad et al. [35] proposed a BFO algorithm for Network-traffic (BFON) to detect and prevent from intrusions during the transfer of big data. Further, it controls the intrusions using resistance mechanism. Schmidt et al. [8] proposed an Artificial Immune System (AIS) algorithm based big data optimization technique to manage and classify flow-based Internet traffic data. To improve the classification performance, AIS algorithm used Euclidian distance and experimental results show that this technique gives more accurate results as compared to Naïve Bayes classifier. George and Parthiban [36] proposed Group Search Optimization (GSO) algorithm based big data analytics technique using FSO to perform data clustering for the high dimensional dataset. This technique replaces the worst fitness values in each iteration of the GSO with the updated values from FSO to test the performance of clustering of data.

1.3.3 Ecological Algorithms

Pouya et al. [9] proposed Invasive Weed Optimization (IWO) algorithm based big data optimization technique to solve the multi-objective portfolio optimization problem. Further, uniform design and fuzzy normalization method are used to transform multi-objective portfolio selection model into a single-objective programming model. IWO algorithm manages big data with lesser execution time than PSO. Pu et al. [10] proposed a hybrid Biogeography-Based Optimization (BBO) algorithm for

multilayer perceptron training under the challenge of big data analysis and processing. Experimental results show that BBO is effective to provide training to multilayer perceptron and performs better in terms of convergence than GA and PSO algorithm. Fong et al. [20] proposed multi-species optimizer (PS2O) algorithm based approach for data stream mining big data to select features. An incremental classification algorithm is used in PS2O algorithm to classify the collected data streams pertaining to Big Data, which enhanced the analytical accuracy within reasonable processing time.

Figure 6 shows the evolution of bio-inspired algorithms for big data analytics based on existing literature as discussed above.

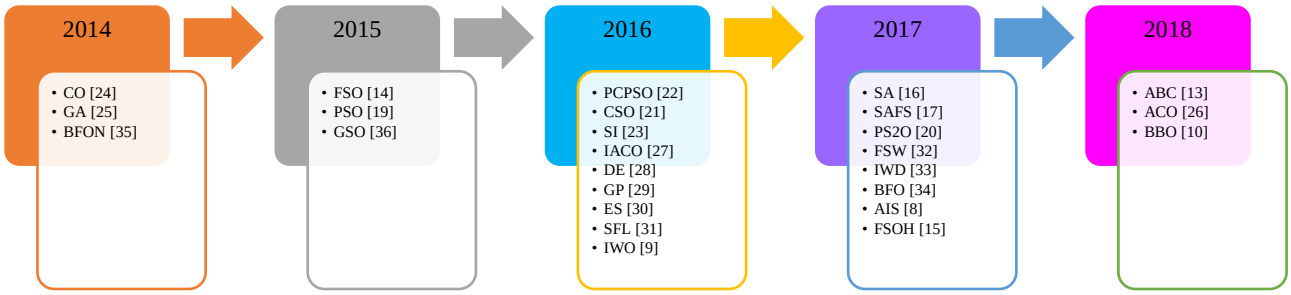

Figure 6: Evolution of Bio-Inspired Algorithms for Big Data Analytics

Figure 7 shows the systematic map, which helps in recognizing the important type of bio-inspired algorithms that were highlighted from 2014 to 2018.

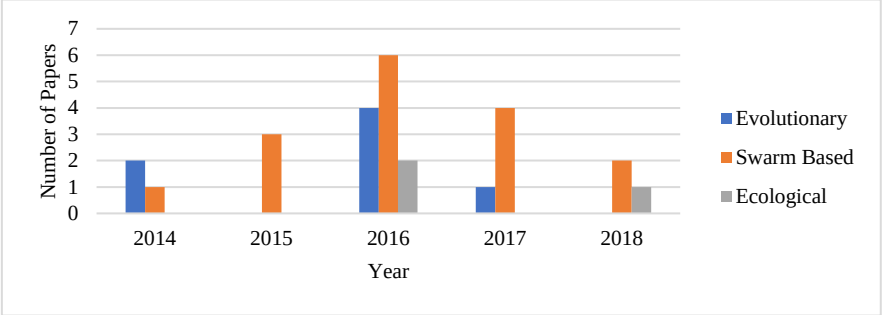

Figure 7: Time Count of Bio-Inspired Algorithms for Big Data Analytics

The literature reported that there are five types of analytics for big data management using bio-inspired algorithms: text analytics, audio analytics, video analytics, social media analytics and predictive analytics as shown in Figure 8.

Text analytics is a method to perform text mining for an extraction of required data from the database such as news, corporate documents, survey responses, online forums, blogs, emails, and social network feeds. There are four methods for text analytics: information extraction, text summarization, question answering and sentimental analysis. The *information extraction* technique extracts a structured data from unstructured data, for example, an extraction of tablet name, type, expiry date from patient's medical data. *Text summarization* method extracts a concise summary of various documents related to some specific topic. *Question answering* method uses a natural language processing to find answers to the questions. *Sentiment analysis* method examines the viewpoint of people regarding events or products.

Audio analytics or speech analytics is a process of an extraction of structured data from an unstructured audio data and healthcare or call center are the examples of an audio analytics. Audio analytics is of two types: Large-Vocabulary Continuous Speech Recognition (LVCSR) and phonetic-based technique. LVCSR performs indexing (to transliterate the speech content of audio) followed

by searching (to find an index-term). The phonetic-based technique is dealing with phonemes or sounds and performs phonetic indexing and searching. *Video analytics* visualize, examine, and extract meaningful information from video streams like CCTV footage, live streaming of match etc. Video analytics can be performed at end devices (edge) or centralized systems (server).

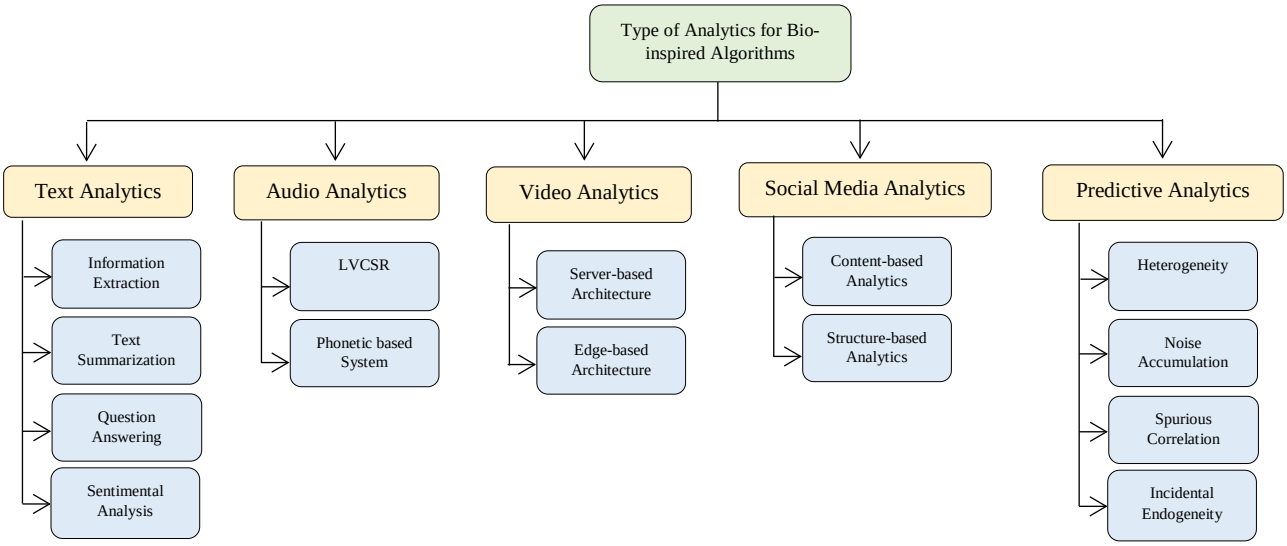

Figure 8: Type of Analytics for Bio-inspired Algorithms

Social media analytics examine the unstructured or structured data of social media websites (a platform, which enables an exchange of information among users) like Facebook, Twitter etc. There are two types of social media analytics: content-based (data posted by users) or structure-based (synthesizing the structural attributes). *Predictive analytics* is a method, which uses historical and current data to predict the future outcomes, which can be done based on: heterogeneity (data is coming from different sources), noise accumulation (an estimation error during interpretation of data), spurious correlation (uncorrelated variable due to huge size of dataset) or incidental endogeneity (predictors or explanatory variables, which are independent of the residual term).

Figure 9 shows the different parameters, which are considered in different bio-inspired algorithms for big data analytics.

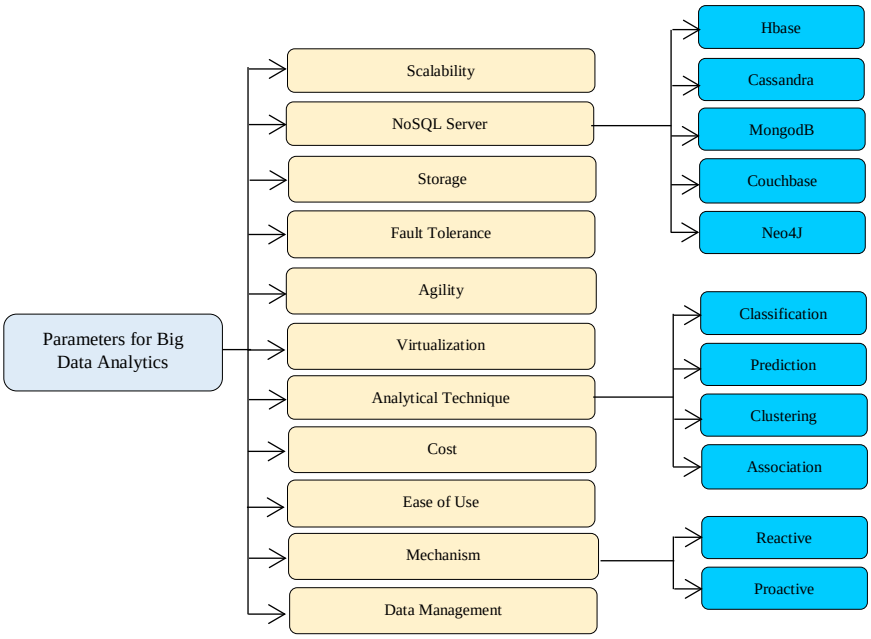

Figure 9: Parameters of different Bio-Inspired Algorithms for Big Data Analytics

There are four different types of *data mining* techniques as studied from literature: classification, prediction, clustering or association. In *classification*, model attributes are used to arrange the data in the different set of categories. *Prediction* technique is used to find out the unknown values. *Clustering* is an unsupervised technique, which clusters the data based on related attributes. An *association* technique is used to establish a relationship among different datasets. There are five types of *NoSQL* database management system (DBMS), which are used in existing techniques: Hbase, Cassandra, MongodB, Couchbase and Neo4J. The two different types of *mechanism* for taking decisions in bio-inspired algorithms for big data analytics: proactive (it is working on forward-looking decisions, which requires for forecasting or text mining) and reactive (perform required decisions based on the requirement for data analytics).

Scalability refers to the mechanism of a system to scale-up or scale-down its nodes based on the amount of transfer of data for analytics. Big data analytics technique uses large *storage* space to store the data to perform the different type of analytics for an extraction of required information. *Fault tolerance* of a system is an ability to process the user data within the required time frame. The type data is coming for analytics is changing continually, so there is a need for *agility*-based big data analytical models to process user data in a required format. *Virtualization* is a technique, which is required for cloud-based systems to create virtual machines for processing of user data in a distributed manner. *Execution cost* is the amount of efforts, which are required to perform big data analytics. *Ease of Use* is defined as the mechanism, which explains that how much easy, the system can perform big data analytics. *Data management* is discussed in *Section 1.2*. Table 1 shows the comparisons of bio-inspired algorithms for big data analytics based on different parameters.

1.4 Future Research Directions and Open Challenges

Recent technological developments such as the Internet of Things, fog computing, edge computing, software-defined clouds are creating new research areas for the bio-inspired algorithm based big data analytics. There is need of the re-evaluation of existing bio-inspired algorithms for big data analytics to address research issues such as energy efficiency, fault tolerance, scalability, processing time etc. In the current scenario, the cloud has emerged as the fifth utility of computing and capturing the significant attention of industries and academia for big data analytics. Virtualization technology is progressing continuously, and new models, mechanisms and approaches are emerging for effective management of big data using cloud infrastructure. Fog computing uses network switches and routers, gateways and mobile base stations to provide cloud service with minimum possible network latency and response time, so fog or edge device can also perform big data analytics at edge device instead of decentralized database or server. Bio-inspired algorithms based big data analytics has several challenges which need to be addressed such as resource management and usability, data processing and elasticity, resilience and heterogeneity in interconnected clouds, sustainability and energy efficiency, data security and privacy protection and edge computing and networking.

1.4.1. Resource Scheduling and Usability

Resource management is the organized method of scheduling available resources to process user data over the Internet. Big data analytics should be performed by allocating virtual resources in an optimized manner and data should be processed with minimum cost and time. Effective resource management in bio-inspired algorithms can improve resource utilization and user satisfaction using the concept of virtualization. There is a problem of under-provisioning and over-provisioning of resources in existing bio-inspired algorithms for big data analytics. To overcome this problem, a Quality of Service (QoS)-aware bio-inspired algorithm based resource management technique is required for efficient management of big data, which optimizes the QoS parameters.

Table 1: Comparison of Bio-Inspired Algorithms for Big Data Analytics

Technique	Scalability	Storage	Fault Tolerance	Agility	Virtualization	Cost	Ease of Use	Type of Analytics	NoSQL DBMS	Mechanism	Type of Data	Dimension of Data Management	Data Mining Technique
ABC [13]	×	√	×	√	×	×	√	Audio	Cassandra	Reactive	Audio	Volume, Variety, Velocity	Prediction
FSO [14]	√	×	×	×	√	×	×	Social	MongodB	Reactive	Social	Volume, Variability	Classification
FSOH [15]	×	√	√	×	×	√	×	Video	Hbase	Proactive	Video	Variety, Velocity, Veracity	Clustering
SA [16]	×	×	√	√	√	×	×	Text	Neo4J	Proactive	Text	Volume, Velocity	Clustering
SAFS [17]	√	×	×	×	√	×	×	Predictive	Hbase	Reactive	Operational	Volume, Velocity	Prediction
FSOSAH [18]	×	√	×	×	×	√	×	Social	Cassandra	Proactive	Social	Variability, Veracity, Variability	Classification
PSO [19]	√	×	×	√	×	×	√	Audio	Neo4J	Proactive	Audio	Variability, Velocity	Prediction
PCPSO [22]	×	×	×	×	×	×	×	Predictive	Cassandra	Proactive	Cloud service	Variety, Veracity	Association
PS2O [20]	√	×	×	×	×	×	×	Predictive	Hbase	Reactive	M2M Data	Volume, Velocity, Variability	Clustering
CSO [21]	×	×	√	×	√	×	√	Text	Cassandra	Proactive	Text	Volume, Velocity, Variability	Prediction
SI [23]	√	×	√	×	√	×	×	Video	MongodB	Proactive	Video	Variety, Veracity	Prediction
CO [24]	√	×	×	√	×	√	×	Predictive	Cassandra	Proactive	Operational	Variability, Variety, Variability	Association
GA [25]	×	√	×	×	×	√	√	Predictive	Couchbase	Reactive	M2M Data	Veracity, Variability	Prediction
ACO [26]	×	√	×	√	×	×	√	Predictive	MongodB	Proactive	Transactional	Volume, Variability, Velocity	Clustering
IACO [27]	√	×	×	×	√	√	×	Social	Cassandra	Proactive	Social	Veracity, Variability, Velocity	Association
DE [28]	×	√	√	×	×	×	×	Video	Hbase	Reactive	Video	Volume, Velocity, Variety	Clustering
GP [29]	×	×	√	√	×	×	×	Audio	Neo4J	Reactive	Audio	Volume, Veracity, Velocity	Association
ES [30]	×	√	×	√	×	×	√	Predictive	Cassandra	Proactive	Cloud service	Velocity, Variability	Classification
SFL [31]	√	×	×	√	×	×	×	Predictive	MongodB	Proactive	Operational	Velocity, Veracity, Variability	Classification
FSW [32]	√	×	√	×	×	×	√	Audio	Couchbase	Proactive	Audio	Variety, Veracity	Prediction
IWD [33]	×	×	√	×	×	×	×	Text	Hbase	Reactive	Text	Volume, Velocity, Variety	Clustering
BFO [34]	√	×	√	×	×	×	√	Predictive	Cassandra	Reactive	Transactional	Volume, Variability, Velocity	Clustering
BFON [35]	×	√	×	×	√	√	×	Predictive	Hbase	Proactive	Operational	Velocity, Veracity, Variability	Prediction
AIS [8]	×	×	√	√	×	×	√	Predictive	Cassandra	Proactive	Transactional	Volume, Velocity	Association
IWC [9]	√	×	×	×	×	×	×	Text	Couchbase	Proactive	Text	Volume, Variability, Velocity	Classification
BBO [10]	×	√	×	×	√	×	√	Predictive	Cassandra	Reactive	Cloud service	Volume, Variety, Veracity, Variability	Association
GSO [36]	×	√	×	×	×	√	×	Predictive	MongodB	Reactive	Transactional	Volume, Velocity, Veracity	Clustering

1.4.2. Data Processing and Elasticity

Computing systems are also facing a challenge of data synchronization in bio-inspired algorithms because data is processed geographically, which overloads the cloud service. To solve this problem, rapid elasticity can be used to find the overloaded nodes and it adds new instances to handle the current data coming from different sources. Further, there is a need for efficient data backup technique for big data analytics to recover the data in case of server downtime.

1.4.3. Resilience and Heterogeneity in Interconnected Clouds

The prominent cloud providers such as Google, Facebook, Amazon and Microsoft are providing highly available cloud computing services using thousands of servers, which consists of multiple resources such as processors, network cards, storage devices and disk drives for processing of big data. With the growing adoption of cloud, Cloud Data Centres (CDCs) are rapidly expanding their sizes and increasing the complexity of the systems, which increases the resource failures during big data analytics. The failure can be Service Level Agreement (SLA) violation, data corruption and loss and premature termination of execution, which can degrade the performance of big data analytics and affect the business. For next-generation big data analytical models to be reliable, there is a need to identify the failures (hardware, service, software or resource), their causes and manages them to improve their reliability. To solve this problem, a bio-inspired algorithm-based data analytical model is required that introduces replication of services and their coordination to enable reliable analytics in a cost-efficient manner.

1.4.4. Sustainability and Energy-efficiency

To provide a reliable cloud service, it is required to identify that how the occurrences of failures affect the energy efficiency of big data analytical models. Moreover, it is necessary to save the checkpoints with the minimum overhead after predicting an occurrence of a failure. Therefore, user data can be migrated to more reliable servers for efficient processing, which can save the energy consumption and time. Further, consolidation of multiple independent instances (web service or email) of an application can be performed using bio-inspired algorithms to improve the energy efficiency, which improves the sustainability and availability of cloud service to improve big data analytical process.

1.4.5. Data Security and Privacy Protection

It is very difficult to incorporate security protocols in big data analytics due to its distributed environment. One of the main security issues is calling authentication at different levels of data management.

1.4.6. IoT based Edge Computing and Networking

IoT based Fog environment consists of a large number of edge devices in a distributed manner and computation may be consuming more energy than centralized cloud environment, hence it is an important research issue for big data analytics. Existing research reported that fog devices are more capable to reduce latency as compared to the server by experiencing a little larger energy consumption. Fog devices have additional compute and storage power, but it is not possible for these devices to provide the resource capacity of the cloud. Therefore, the bio-inspired algorithm based an efficient big data analytical technique is required to process the user data at an edge device instead of the server, which can reduce execution time and cost.

1.5 Emerging Research Areas in Bio-Inspired Algorithms based Big Data Analytics

In addition to future research directions, there are various hotspot research areas in bio-inspired algorithms based big data analytics which needs to be addressed in future such as containers, serverless computing, blockchain, software-defined clouds, bitcoin, deep learning and quantum

computing. Here we discuss hotspot research areas in the context of b bio-inspired algorithms based ig data analytics.

1.5.1 Container as a Service (CaaS)

Docker is container-based virtualization technology can be used for bio-inspired algorithms based big data analytics in multiple clouds using light weight web server i.e. HUE (Hadoop User Experience) web interface. HUE-based Docker container provides robust and light-weight Container as a Service (CaaS) data processing facility using virtual multi-cloud environment.

1.5.2 Serverless Computing as a Service (SCaaS)

Serverless computing can be used for bio-inspired algorithms based big data analytics without managing the cloud infrastructure and it is effective in processing of user data without configuration of the network and resource provisioning. Serverless Computing as a Service (SCaaS) have two different services: Backend-as-a-Service (BaaS) and Function-as-a-Service (FaaS), which can improve the efficiency, robustness and scalability of big data processing systems and analyse the data in a fastest manner.

1.5.3 Blockchain as a Service (BaaS)

Blockchain is a distributed database system, which can manage the huge amount of data at low cost and provides instant risk-free transaction. Blockchain as a Service (BaaS) decrease the time of processing a transaction dramatically and increases security, quality and integrity of data in bio-inspired algorithms based big data analytics.

1.5.4 Software-defined Cloud as a Service (SCaaS)

There is a huge amount of data is coming from different IoT devices and it is necessary to transfer data from source to destination without any loss. Software-defined Cloud as a Service (SCaaS) is a new paradigm, which provides the effective network architecture to move data from IoT devices to cloud datacenter in a reliable and efficient manner by taking intelligent infrastructure decisions. Further, SCaaS offers other advantages for bio-inspired algorithms based big data analytics such as failure recovery, optimisation of network resources and fast computing power.

1.5.5 Deep Learning as a Service (DLaaS)

Deep learning is a new paradigm for bio-inspired algorithms based big data analytics to process the big data with high accuracy and efficiency in real time manner using hybrid learning and training mechanisms. Deep Learning as a Service (DLaaS) uses hierarchical learning process to extract high-level, complex abstractions as data representations for analysis and learning of massive amounts of unsupervised data.

1.5.6 Bitcoin as a Service (BiaaS)

Cryptocurrencies are very popular technology to provide secure and reliable service for huge number of financial transactions. Bitcoin as a Service (BiaaS) performs real-time data extraction from the blockchain ledger and stores the big data in an efficient manner for bio-inspired algorithms based big data analytics. BiaaS based big data analytics provides interesting benefits such as trend prediction, theft prevention and identification of malicious users.

1.5.7 Quantum Computing as a Service (QCaaS)

The new trend of Quantum computing helps bio-inspired algorithms based big data analytics to solve complex problems by handling massive digital datasets in an efficient and fastest manner. Quantum Computing as a Service (QCaaS) allows for quick detection, analysis, integration, and diagnosis from large scattered data sets. Further, QCaaS can search extensive, unsorted data sets to quickly uncover patterns.

1.6 Summary and Conclusions

In this chapter, the survey of bio-inspired algorithms for big data analytics has been presented. We identified the focus of study of bio-inspired algorithms and proposed open research challenges. Based on the identified important open issues and focus of study, taxonomy and comparison of bio-inspired algorithms has also been presented. Bio-inspired algorithms are categorized into three different categories and investigate the existing research works based on their techniques towards addressing the research challenges. Moreover, we proposed some promising research directions based on the analysis, that can be pursued in the future.

Acknowledgements

One of the authors, Dr. Sukhpal Singh Gill [Postdoctoral Research Fellow], gratefully acknowledges the Cloud Computing and Distributed Systems (CLOUDS) Laboratory, School of Computing and Information Systems, The University of Melbourne, Australia, for awarding him the Fellowship to carry out this research work.

References

[1] Khan, S., Liu, X., Shakil, K.A. and Alam, M., 2017. A survey on scholarly data: From big data perspective. Information Processing & Management, 53(4), pp. 923-944.

[2] Gandomi, A. and Haider, M., 2015. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), pp.137-144.

[3] Gill, S.S. and Buyya, R., 2018. Resource Management for Internet of Things using Fog-assisted Cloud: The Next Generation Smart Home Controller, Cluster Computing.

[4] Gill, S.S., Arya, R.C., Wander, G.S. and Buyya, R., 2018. Fog-based Smart Healthcare as a Big Data and Cloud Service for Heart Patients using IoT, International conference on Computer Networks and Inventive Communication Technologies (ICCNCT - 2018).

[5] Chen, C.P. and Zhang, C.Y., 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences, 275, pp.314-347.

[6] Wang, J., Wu, Y., Yen, N., Guo, S. and Cheng, Z., 2016. Big data analytics for emergency communication networks: A survey. IEEE Communications Surveys & Tutorials, 18(3), pp.1758-1778.

[7] Gill, S.S. and Buyya, R., 2018. A Taxonomy and Future Directions for Sustainable Cloud Computing: 360 Degree View. arXiv preprint arXiv:1712.02899.

[8] Schmidt, B., Al-Fuqaha, A., Gupta, A. and Kountanis, D., 2017. Optimizing an artificial immune system algorithm in support of flow-Based internet traffic classification. Applied Soft Computing, 54, pp.1-22.

[9] Pouya, A.R., Solimanpur, M. and Rezaee, M.J., 2016. Solving multi-objective portfolio optimization problem using invasive weed optimization. Swarm and Evolutionary Computation, 28, pp.42-57.

[10] Pu, X., Chen, S., Yu, X. and Zhang, L., 2018. Developing a Novel Hybrid Biogeography-Based Optimization Algorithm for Multilayer Perceptron Training under Big Data Challenge. Scientific Programming, 2018.

[11] Gill, S.S., Chana, I. and Buyya, R., 2017. IoT Based Agriculture as a Cloud and Big Data Service: The Beginning of Digital India. Journal of Organizational and End User Computing (JOEUC), 29(4), pp.1-23.

[12] Singh, I., Singh, K.V. and Singh, S., 2017. Big Data Analytics Based Recommender System for Value Added Services (VAS). In Proceedings of Sixth International Conference on Soft Computing for Problem Solving (pp. 142-150). Springer, Singapore.

[13] Ilango, S.S., Vimal, S., Kaliappan, M. and Subbulakshmi, P., 2018. Optimization using Artificial Bee Colony based clustering approach for big data. Cluster Computing, pp.1-9.
[14] Raj, E.D. and Babu, L.D., 2015. A firefly swarm approach for establishing new connections in social networks based on big data analytics. International Journal of Communication Networks and Distributed Systems, 15(2-3), pp.130-148.
[15] Wang, H., Wang, W., Cui, L., Sun, H., Zhao, J., Wang, Y. and Xue, Y., 2017. A hybrid multi-objective firefly algorithm for big data optimization. Applied Soft Computing.
[16] Mafarja, M.M. and Mirjalili, S., 2017. Hybrid Whale Optimization Algorithm with simulated annealing for feature selection. Neurocomputing, 260, pp.302-312.
[17] Barbu, A., She, Y., Ding, L. and Gramajo, G., 2017. Feature selection with annealing for computer vision and big data learning. IEEE transactions on pattern analysis and machine intelligence, 39(2), pp.272-286.
[18] Tayal, A. and Singh, S.P., 2016. Integrating big data analytic and hybrid firefly-chaotic simulated annealing approach for facility layout problem. Annals of Operations Research, pp.1-26.
[19] Wang, L., Geng, H., Liu, P., Lu, K., Kolodziej, J., Ranjan, R. and Zomaya, A.Y., 2015. Particle swarm optimization based dictionary learning for remote sensing big data. Knowledge-Based Systems, 79, pp.43-50.
[20] Fong, S., Wong, R. and Vasilakos, A.V., 2016. Accelerated PSO swarm search feature selection for data stream mining big data. IEEE transactions on services computing, 9(1), pp.33-45.
[21] Lin, K.C., Zhang, K.Y., Huang, Y.H., Hung, J.C. and Yen, N., 2016. Feature selection based on an improved cat swarm optimization algorithm for big data classification. The Journal of Supercomputing, 72(8), pp.3210-3221.
[22] Hossain, M.S., Moniruzzaman, M., Muhammad, G., Ghoneim, A. and Alamri, A., 2016. Big data-driven service composition using parallel clustered particle swarm optimization in mobile environment. IEEE Transactions on Services Computing, 9(5), pp.806-817.
[23] Cheng, S., Zhang, Q. and Qin, Q., 2016. Big data analytics with swarm intelligence. Industrial Management & Data Systems, 116(4), pp.646-666.
[24] Saida, I.B., Nadjet, K. and Omar, B., 2014. A new algorithm for data clustering based on cuckoo search optimization. In Genetic and Evolutionary Computing (pp. 55-64). Springer, Cham.
[25] Kune, R., Konugurthi, P.K., Agarwal, A., Chillarige, R.R. and Buyya, R., 2014, December. Genetic algorithm based data-aware group scheduling for Big Data clouds. In Big Data Computing (BDC), 2014 IEEE/ACM International Symposium on (pp. 96-104). IEEE.
[26] Banerjee, S. and Badr, Y., 2018. Evaluating Decision Analytics from Mobile Big Data using Rough Set Based Ant Colony. In Mobile Big Data (pp. 217-231). Springer, Cham.
[27] Pan, X., 2016. Application of improved ant colony algorithm in intelligent medical system: from the perspective of big data. CHEMICAL ENGINEERING, 51.
[28] Elsayed, S. and Sarker, R., 2016. Differential evolution framework for big data optimization. Memetic Computing, 8(1), pp.17-33.
[29] Gandomi, A.H., Sajedi, S., Kiani, B. and Huang, Q., 2016. Genetic programming for experimental big data mining: A case study on concrete creep formulation. Automation in Construction, 70, pp.89-97.
[30] Kashan, A.H., Keshmiry, M., Dahooie, J.H. and Abbasi-Pooya, A., 2016. A simple yet effective grouping evolutionary strategy (GES) algorithm for scheduling parallel machines. Neural Computing and Applications, pp.1-14.
[31] Hu, B., Dai, Y., Su, Y., Moore, P., Zhang, X., Mao, C., Chen, J. and Xu, L., 2016. Feature selection for optimized high-dimensional biomedical data using the improved shuffled frog leaping algorithm. IEEE/ACM transactions on computational biology and bioinformatics.

[32] Manikandan, R.P.S. and Kalpana, A.M., 2017. Feature selection using fish swarm optimization in big data. Cluster Computing, pp.1-13.
[33] Elsherbiny, S., Eldaydamony, E., Alrahmawy, M. and Reyad, A.E., 2017. An extended Intelligent Water Drops algorithm for workflow scheduling in cloud computing environment. Egyptian Informatics Journal.
[34] Neeba, E.A. and Koteeswaran, S., 2017. Bacterial foraging information swarm optimizer for detecting affective and informative content in medical blogs. Cluster Computing, pp.1-14.
[35] Ahmad, K., Kumar, G., Wahid, A. and Kirmani, M.M., 2014. Intrusion Detection and Prevention on Flow of Big Data Using Bacterial Foraging. Handbook of Research on Securing Cloud-Based Databases with Biometric Applications, p.386.
[36] George, G. and Parthiban, L., 2015, November. Multi objective hybridized firefly algorithm with group search optimization for data clustering. In Research in Computational Intelligence and Communication Networks (ICRCICN), 2015 IEEE International Conference on (pp. 125-130). IEEE.

Key Terminology & Definitions

Bigdata – This is the set of huge datasets, which contains the different type of data such as video, audio, text, social etc.

Data Management - There are five types of dimensions of data management for big data analytics: volume, variety, velocity, veracity and variability

Big Data Analytics – The process of an extraction of required data from unstructured data is called big data analytics and there are five types of analytics for big data management using bio-inspired algorithms: text analytics, audio analytics, video analytics, social media analytics and predictive analytics.

Bio-inspired Optimization - The bio-inspired algorithms are used for big data analytics, which can be ecological, swarm-based and evolutionary algorithms.

Cloud Computing - Cloud computing offers three types of main service models: software, platform and infrastructure. At the software level, the cloud user can utilize application in a flexible manner, which is running on cloud datacenters. Cloud user can access infrastructure to develop and deploy cloud applications at platform level. Infrastructure as a service offers access to computing resources such as processor, networking and storage and enable virtualization-based computing.

Authors Bio:

Dr. Sukhpal Singh Gill is working as Postdoctoral Research Fellow at Cloud Computing and Distributed Systems (CLOUDS) Laboratory, School of Computing and Information Systems, The University of Melbourne, Australia. Dr. Gill joined Computer Science and Engineering Department of Thapar Institute of Engineering and Technology (TIET), Patiala, India, in 2016 as a Faculty. Dr. Gill obtained the Degree of Master of Engineering in Software Engineering from TIET, as well as a Doctoral Degree specialization in "Autonomic Cloud Computing" from TIET. Dr. Gill received the Gold Medal in Master of Engineering in Software Engineering. Dr. Gill was a DST Inspire Fellow [2013-2016] and worked as a SRF-Professional on DST Project, Government of India. He has done certifications in Cloud Computing Fundamentals, including Introduction to Cloud Computing and Aneka Platform (US Patented) by ManjraSoft Pty Ltd, Australia and Certification of Rational Software Architect (RSA) by IBM India. His research interests include Software Engineering, Cloud

Computing, Internet of Things, Big Data and Fog Computing. He has more than 40 research publications in reputed journals and conferences. For further information on Dr. Gill, please visit: www.ssgill.in

Affiliation/Address:

E-mail: sukhpal.gill@unimelb.edu.au
Affiliation

Cloud Computing and Distributed Systems (CLOUDS) Laboratory
School of Computing and Information Systems
The University of Melbourne, Australia

Dr. Rajkumar Buyya is a Redmond Barry Distinguished Professor and Director of the Cloud Computing and Distributed Systems (CLOUDS) Laboratory at the University of Melbourne, Australia. He is also serving as the founding CEO of Manjrasoft, a spin-off company of the University, commercializing its innovations in Cloud Computing. He served as a Future Fellow of the Australian Research Council during 2012-2016. He has authored over 625 publications and seven text books including "Mastering Cloud Computing" published by McGraw Hill, China Machine Press, and Morgan Kaufmann for Indian, Chinese and international markets respectively. He is one of the highly cited authors in computer science and software engineering worldwide (h-index=116, g-index=255, 71000+ citations). Recently, Dr. Buyya is recognized as a "Web of Science Highly Cited Researcher" in both 2016 and 2017 by Thomson Reuters, a Fellow of IEEE, and Scopus Researcher of the Year 2017 with Excellence in Innovative Research Award by Elsevier for his outstanding contributions to Cloud computing. He served as the founding Editor-in-Chief of the IEEE Transactions on Cloud Computing. He is currently serving as Editor-in-Chief of Journal of Software: Practice and Experience, which was established over 45 years ago. For further information on Dr. Buyya, please visit his cyberhome: www.buyya.com

Affiliation/Address:

E-mail: rbuyya@unimelb.edu.au
Affiliation

Cloud Computing and Distributed Systems (CLOUDS) Laboratory
School of Computing and Information Systems
The University of Melbourne, Australia