CHAPTER

# BIO-INSPIRED ALGORITHMS FOR BIG DATA ANALYTICS: A SURVEY, TAXONOMY, AND OPEN CHALLENGES

1

**Sukhpal Singh Gill, Rajkumar Buyya**

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

## 1.1 INTRODUCTION

Cloud computing is now the spine of the modern economy, which offers on-demand services to cloud customers through the Internet. To improve the performance and effectiveness of cloud computing systems, new technologies, such as internet of things (IoT) applications (healthcare services, smart cities etc.) and big data, are emerging, which further requires effective data processing to process data [1]. However, there are two problems in existing big data processing approaches, which degrade the performance of computing systems such as large response time and delay due to data being transferred twice [2]: (1) computing systems to cloud and (2) cloud to IoT applications. Presently, IoT devices collect data with a huge amount of volume (big data) and variety and these systems are growing with the velocity of 500 MB/seconds or more [3].

For IoT based smart cities, the transfer of data is used to make effective decisions for big data analytics. Data is stored and processed on cloud servers after collection and aggregation of data from smart devices on IoT networks. Further, to process the large volume of data, there is a need for automatic highly scalable cloud technology, which can further improve the performance of the systems [4]. Literature reported that existing cloud-based data processing systems are not able to satisfy the performance requirements of IoT applications when a low response time and latency is needed. Moreover, other reasons for a large response time and latency are: geographical distribution of data and communication failures during transfer of data [5]. Cloud computing systems become bottlenecked due to continually receiving raw data from IoT devices [6]. Therefore, a bio-inspired algorithm based big data analytics is an alternative paradigm that provides a platform between computing systems and IoT devices to process user data in an efficient manner [7].

**Big Data Analytics for Intelligent Healthcare Management.** https://doi.org/10.1016/B978-0-12-818146-1.00001-5

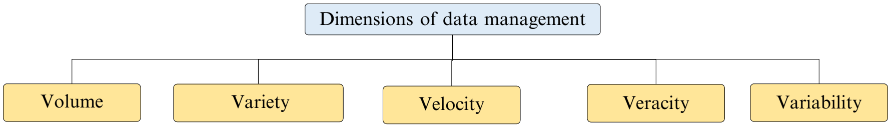

**FIG. 1.1**

Dimensions of data management.

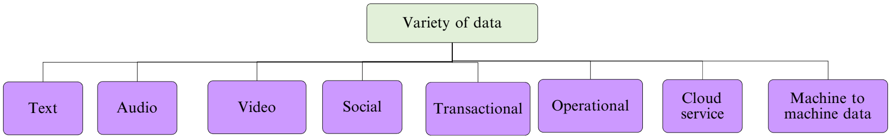

**FIG. 1.2**

Variety of data.

### 1.1.1 DIMENSIONS OF DATA MANAGEMENT

As identified from existing literature [1–6], there are five kinds of dimensions of data, which are required for effective management. Fig. 1.1 shows the dimensions of data management for big data analytics: (1) volume, (2) variety, (3) velocity, (4) veracity, and (5) variability.

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. Fig. 1.2 shows the variety of data, which can be text, audio, video, social, transactional, operational, cloud service, or machine to machine data (M2M data).

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

The rest of the chapter is organized as follows. In Section 1.2, we present the big data analytical model. In Section 1.3, we propose the taxonomy of bio-inspired algorithms for big data analytics. In Section 1.4, we analyze research gaps and present some promising directions toward 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 term, which is a combination of “big data” and “deep analysis” as shown in Fig. 1.3. Every minute, a large amount of user data is being transferred from one device to another device, which needs high processing power to perform data mining for the extraction of useful information from the database. Fig. 1.3 shows the model for big data analytics, which shows that an OLTP (on-line transaction processing) system creates data (txn data). A data cube represents a big data, out of

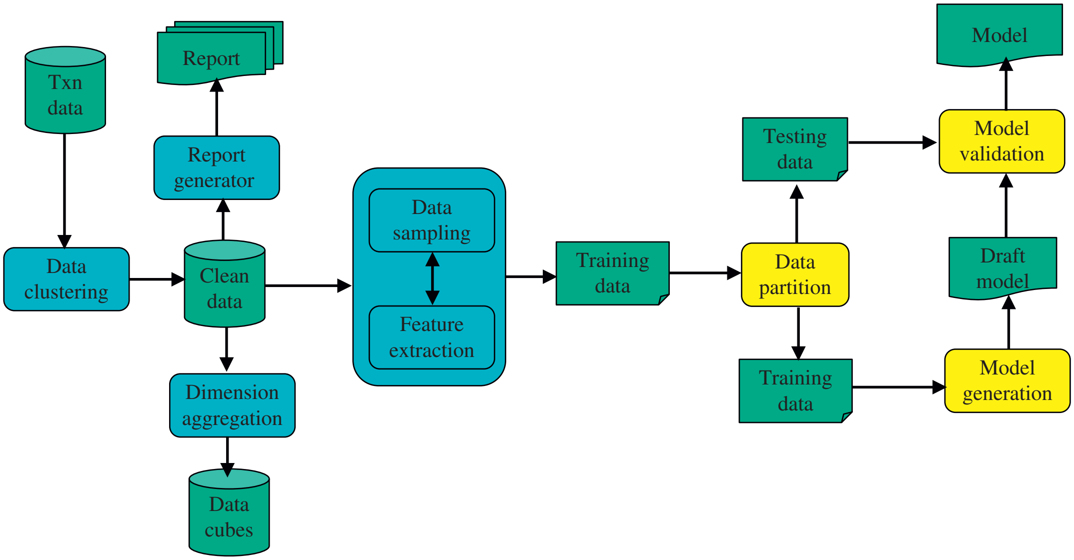

FIG. 1.3

Big data analytical model.

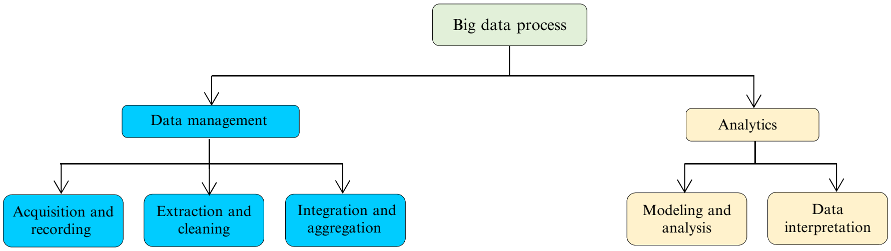

FIG. 1.4

Big data process.

which required information can be extracted using data mining. Initially, different types of data come from different users or devices and the process of data cleansing is performed to remove the irrelevant data and stores the clean data in the database [8]. Further, data aggregation is performed to store the data in an efficient manner because incoming data contains a variety of data and a report 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 includes data visualization, model learning (e.g., K-nearest-neighbor, Linear regression), and model evaluation [9].

Fig. 1.4 shows the process of big data, which has two main components: data management and analytics. There are five different stages in processing 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 (5) data interpretation (represent data in required form).

## 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 categories: ecological, swarm-based, and evolutionary. Fig. 1.5 shows the taxonomy of bio-inspired algorithms for big data analytics along with focus of study (FoS).

### 1.3.1 EVOLUTIONARY ALGORITHMS

Kune et al. [10] proposed a genetic algorithm (GA) based data-aware family scheduling approach for analytics of big data, which focuses on bandwidth utilization, computational resources, and data dependencies. Moreover, the GA algorithm decoupled data and computational services are provided as cloud services. The results demonstrate that the GA algorithm gives effective results in terms of turnaround time because the GA algorithm processes data using parallel processing. Gandomi et al. [11] proposed a multiobjective 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. The GP model works with high and normal strength. Elsayed and Sarker [12] proposed a differential evolution (DE) algorithm-based big data analytics approach, which uses local search to increase the exploitation capability of the DE algorithm. This approach optimizes the big data 2015 benchmark problems with both multi- and single-objective problems but it exhibits large computational time. Kashan et al. [13] proposed an evolutionary strategy (ES) algorithm-based big data analytics technique, which processes data efficiently and accurately using parallel scheduling of cloud resources. Further, the ES algorithm minimizes the execution time by partitioning a group of jobs into disjointed sets, in which the same resources execute all the jobs in the same set.

Mafarja and Mirjalili [14] proposed a simulated annealing (SA) algorithm-based big data optimization technique, which uses the whale optimization algorithm (WOA) to architect various feature selection approaches to reduce the manipulation by probing the most capable regions. The proposed approach helps to improve the classification accuracy and selects the most useful features for categorization tasks. Further, Barbu et al. [15] proposed an SA algorithm-based feature selection (SAFS) technique for big data learning and computer vision. Based on a criterion, the SAFS algorithm removes variables and tightens a sparsity constraint, which reduces the problem size gradually during the iterations and this makes it mainly fit for big data learning. Tayal and Singh [16] proposed big data analytics based on the FSO and SA-based hybrid (FSOSAH) technique for a stochastic dynamic facility layout-based multiobjective problem to manage data effectively. Saida et al. [17] proposed the cuckoo search optimization (CO) algorithm-based big data analytics approach for clustering data. Further, different datasets from the UCI Machine Learning Repository are considered to validate the CO algorithm through experimental results and these datasets perform better in terms of computational efficiency and convergence stability.

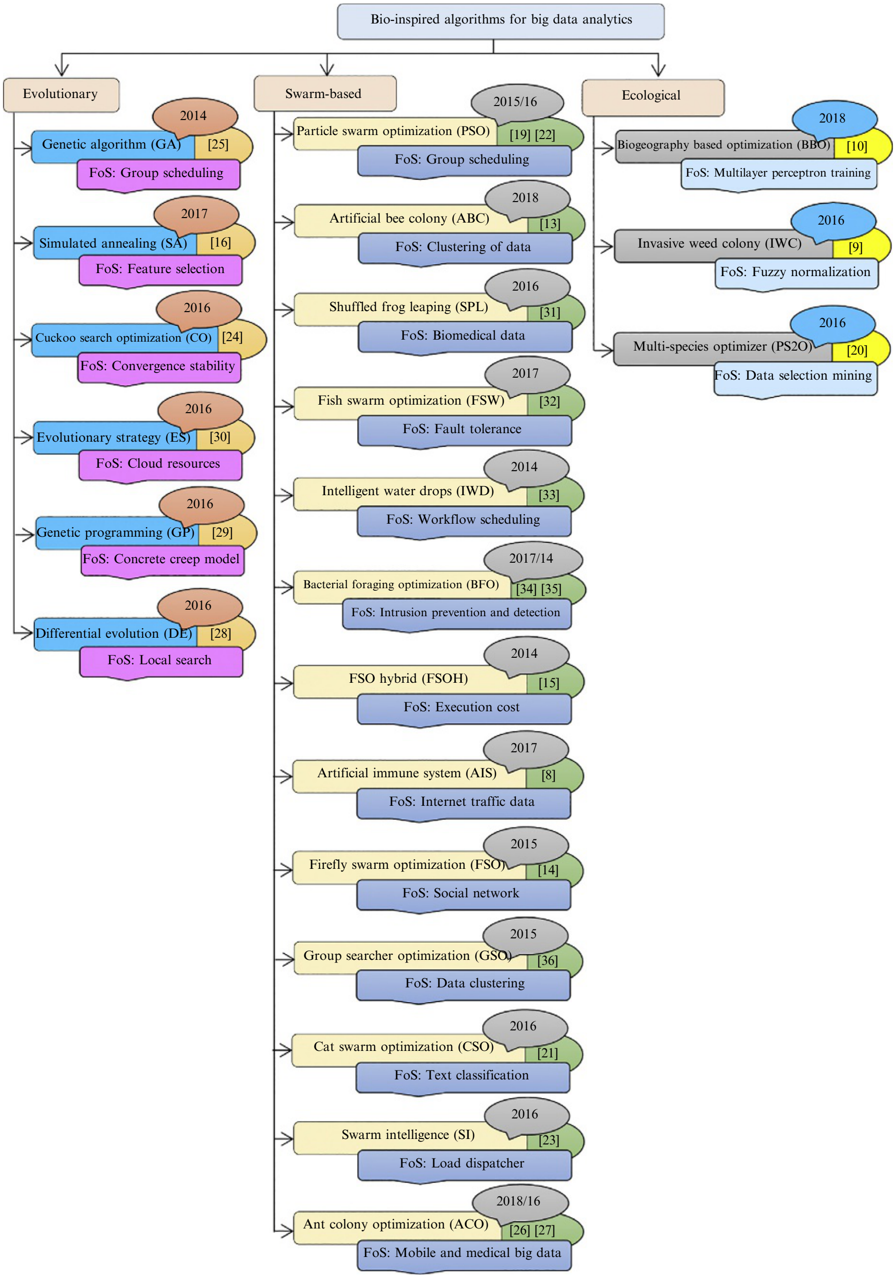
Bio-inspired algorithms for big data analytics
Evolutionary
2014
Genetic algorithm (GA)
[25]
FoS: Group scheduling
2017
Simulated annealing (SA)
[16]
FoS: Feature selection
2016
Cuckoo search optimization (CO)
[24]
FoS: Convergence stability
2016
Evolutionary strategy (ES)
[30]
FoS: Cloud resources
2016
Genetic programming (GP)
[29]
FoS: Concrete creep model
2016
Differential evolution (DE)
[28]
FoS: Local search
Swarm-based
2015/16
Particle swarm optimization (PSO)
[19] [22]
FoS: Group scheduling
2018
Artificial bee colony (ABC)
[13]
FoS: Clustering of data
2016
Shuffled frog leaping (SPL)
[31]
FoS: Biomedical data
2017
Fish swarm optimization (FSW)
[32]
FoS: Fault tolerance
2014
Intelligent water drops (IWD)
[33]
FoS: Workflow scheduling
2017/14
Bacterial foraging optimization (BFO)
[34] [35]
FoS: Intrusion prevention and detection
2014
FSO hybrid (FSOH)
[15]
FoS: Execution cost
2017
Artificial immune system (AIS)
[8]
FoS: Internet traffic data
2015
Firefly swarm optimization (FSO)
[14]
FoS: Social network
2015
Group searcher optimization (GSO)
[36]
FoS: Data clustering
2016
Cat swarm optimization (CSO)
[21]
FoS: Text classification
2016
Swarm intelligence (SI)
[23]
FoS: Load dispatcher
2018/16
Ant colony optimization (ACO)
[26] [27]
FoS: Mobile and medical big data
Ecological
2018
Biogeography based optimization (BBO)
[10]
FoS: Multilayer perceptron training
2016
Invasive weed colony (IWC)
[9]
FoS: Fuzzy normalization
2016
Multi-species optimizer (PS2O)
[20]
FoS: Data selection mining

**FIG. 1.5**

Taxonomy of bio-inspired algorithms for big data analytics.

### 1.3.2 SWARM-BASED ALGORITHMS

Ilango et al. [9] proposed an artificial bee colony (ABC) algorithm-based clustering technique for management of big data, which identifies the best cluster and performs the optimization for different dataset sizes. The ABC algorithm approach minimizes the execution time and improves the accuracy. A MapReduce-based Hadoop environment is used for implementation and results demonstrate that the ABC algorithm delivers a more effective outcome than the differential evolution and particle swarm optimization (PSO) in terms of execution time. Raj and Babu [18] proposed a firefly swarm optimization (FSO) algorithm for big data analytics for establishing novel connections in social networks to calculate the possibility of sustaining a social network. In this technique, a mathematical model is introduced to test the stability of the social network and this reduces the cost of big data management. Wang et al. [19] proposed an FSO algorithm-based hybrid (FSOH) approach for big data optimization to focus on six multiobjective problems. It reduces execution costs but it has high computational time complexity.

Wang et al. [20] proposed a PSO algorithm-based big data optimization approach to improve online dictionary learning and introduced a dictionary-learning model using the atom-updating stage. The PSO algorithm reduces the heavy computational burdens and improves the accuracy. Hossain et al. [21] proposed a parallel clustered PSO algorithm (PCPSO)-based approach for big data-driven service composition. The PCPSO algorithm handles huge amounts of heterogeneous data and process data using parallel processing with MapReduce in the Hadoop platform. Lin et al. [22] proposed a cat swarm optimization (CSO) algorithm-based approach for big data classification to choose characteristics during classification of text for big data analytics. The CSO algorithm uses the term frequency-inverse document occurrence to improve accuracy of feature selection.

Cheng et al. [23] proposed a swarm intelligence (SI) algorithm-based big data analytics approach for the economic load dispatch problem and the SI algorithm handles the high dimensional data, which improves the accuracy of the data processing. Banerjee and Badr [24] proposed the ant colony optimization (ACO) algorithm-based approach for mobile big data using rough set. The ACO algorithm helps to select an optimal feature for resolved decisions, which aids in effectively managing big data from social networks (tweets and posts). Pan [25] proposed the improved ACO algorithm (IACO)-based big data analytical approach for management of medical data such as patient data, operation data etc., which helps doctors retrieve the required data quickly.

Hu et al. [26] proposed a shuffled frog leaping (SFL) approach to perform the selection of the feature for improved high-dimensional biomedical data. For improved high-dimensional biomedical data, the SFL algorithm maximizes the predictive accuracy by exploring the space of probable subsets to obtain the group of characteristics and reduce irrelevant features. Manikandan and Kalpana [27] proposed a fish swarm optimization (FSW) algorithm for feature selection in big data. The FSO algorithm reduces the combinatorial problems by employing the fish swarming behavior and this is effective for diverse applications. Social interactions among big data have been designed using the movement of fish in their search for food. This algorithm provides effective output in terms of fault tolerance and data accuracy. Elsherbiny et al. [28] proposed the intelligent water drops (IWD) algorithm for workflow scheduling to effectively manage big data. The workflows simulation toolkit is used to test the effectiveness of the IWD-based approach and results show that the IWD-based approach is performed effectively in terms of cost and makespan when compared to the FCFS, Round Robin, and PSO algorithm.

Neeba and Koteeswaran [29] proposed a bacterial foraging optimization (BFO) algorithm to classify the informative and affective content from medical weblogs. MAYO clinic data is used as a medical data source to evaluate the accuracy to retrieve the relevant information. Ahmad et al. [30] proposed a BFO algorithm for network-traffic (BFON) to detect and prevent intrusions during the transfer of big data. Further, it controls the intrusions using a resistance mechanism. Schmidt et al. [31] 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, the AIS algorithm used Euclidian distance and the results demonstrate that this technique produces more accurate results when compared to the Naïve Bayes classifier. George and Parthiban [32] proposed the 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 every iteration of the GSO with the improved values from FSO to test the performance of clustering data.

### 1.3.3 ECOLOGICAL ALGORITHMS

Pouya et al. [33] proposed the invasive weed optimization (IWO) algorithm-based big data optimization technique to resolve the multiobjective portfolio optimization task. Further, the uniform design and fuzzy normalization method are used to transform the multiobjective portfolio selection model into a single-objective programming model. The IWO algorithm manages big data more quickly than PSO. Pu et al. [34] proposed a hybrid biogeography-based optimization (BBO) algorithm for multilayer perceptron training under the challenge of analysis and processing of big data. Experimental results show that BBO is effective in providing training to multilayer perceptron and performs better in terms of convergence when compared to the GA and PSO algorithm. Fong et al. [35] proposed the multispecies optimizer (PS2O) algorithm-based approach for data stream mining big data to select features. An incremental classification algorithm is used in the PS2O algorithm to classify the collected data streams pertaining to big data, which enhanced the analytical accuracy within a reasonable processing time.

Fig. 1.6 shows the evolution of bio-inspired algorithms for big data analytics based on existing literature as discussed above.

Fig. 1.7 shows the number of papers published for each category of bio-inspired algorithm per year. This helps to recognize the important types of bio-inspired algorithms [14–23, 35] [11–13, 25–29] that were highlighted from 2014 to 2018.

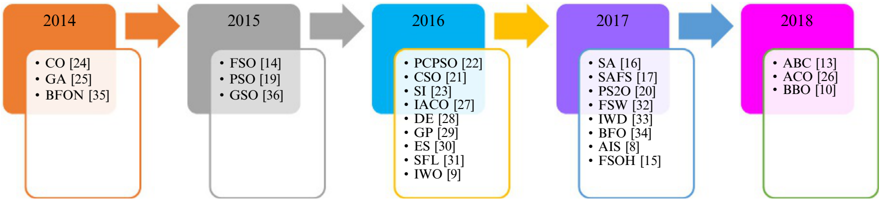

**FIG. 1.6**

Evolution of bio-inspired algorithms for big data analytics.

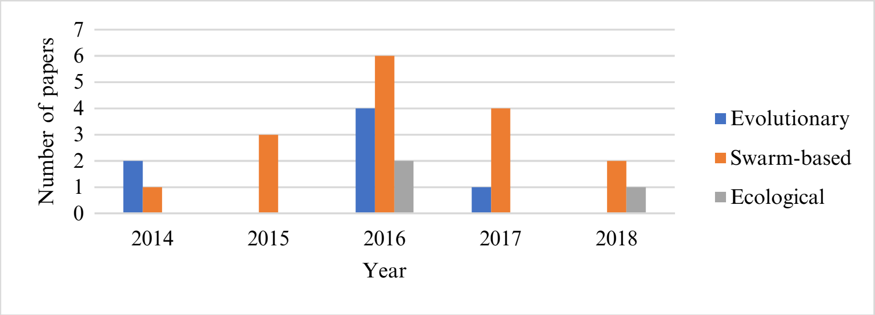

**FIG. 1.7**

Time count of bio-inspired algorithms for big data analytics.

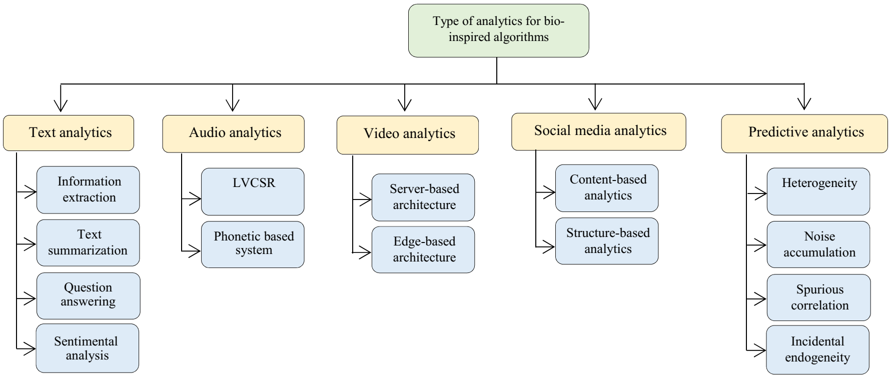

**FIG. 1.8**

Type of analytics for bio-inspired algorithms.

The literature reported that there are five types of analytics for big data management using bio-inspired algorithms: predictive analytics, social media analytics, video analytics, audio analytics, and text analytics as shown in Fig. 1.8.

*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: (1) sentimental analysis, (2) question answering, (3) text summarization, and (4) information extraction. The *information extraction* technique extracts structured data from unstructured data, for example, an extraction of tablet name, type, and expiry date from patient's medical data. The *text summarization* method extracts a concise summary of various documents related to a specific topic. The *question answering* method uses a natural language processing to find answers to the questions. The *sentiment analysis* method examines the viewpoint of people regarding events or products.

*Audio analytics* or speech analytics is a process of extraction of structured data from unstructured audio data and examples of an audio analytics are healthcare or call center data. Audio analytics has 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 deals with phonemes or sounds and performs phonetic indexing and searching.

*Video analytics* visualize, examine, and extract meaningful information from video streams such as CCTV footage, live streaming of sport matches etc. Video analytics can be performed at end devices (edge) or centralized systems (server).

*Social media analytics* examines the unstructured or structured data of social media websites (a platform that enables an exchange of information among users) such as Facebook, Twitter etc. There are two kinds of social media analytics: (1) content-based (data posted by users) or (2) structure-based (synthesizing the structural attributes). *Predictive analytics* is a method that uses historical and current data to predict future outcomes, which can be done based on: heterogeneity (data 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).

Fig. 1.9 shows the different parameters that are considered in different bio-inspired algorithms for big data analytics.

There are four 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 a different set of categories. The *prediction* technique is used to find out the unknown values. *Clustering* is an

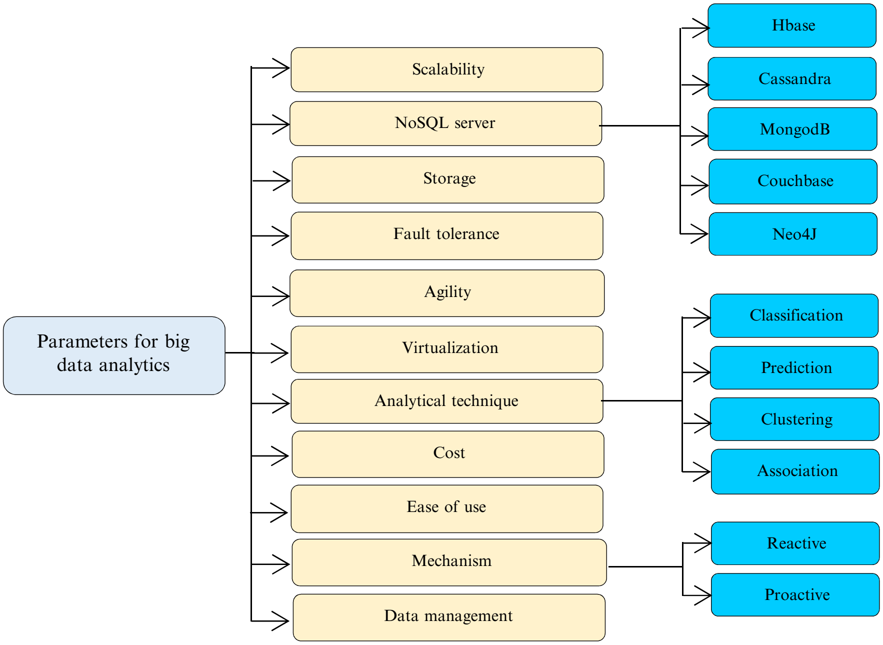

**FIG. 1.9**

Parameters of different bio-inspired algorithms for big data analytics.

unsupervised technique, which clusters the data based on related attributes. The *association* technique is used to establish a relationship among different datasets. There are five types of *NoSQL* database management systems (DBMS) that are used in existing techniques: Hbase, Cassandra, MongodB, Couchbase, and Neo4J. The two different types of *mechanism* for making decisions in bio-inspired algorithms for big data analytics are: proactive (working on forward-looking decisions, which requires forecasting or text mining) and reactive (decisions based on the requirement for data analytics).

*Scalability* refers to the mechanism of a computing system to scale-up or scale-down its nodes based on the amount of transfer of data for analytics. Big data analytics techniques use a large amount of storage space to store the information to perform the different types of analytics to extract the required information. *Fault tolerance* of a system is the ability to process the user data within the required time frame. The type of data that is requiring analytics is continually changing, so there is a need for *agility*-based big data analytical models to process user data in a required format. *Virtualization* is a technique that is required for cloud-based systems to create virtual machines for processing user data in a distributed manner. *Execution cost* is the amount of efforts that are required to perform big data analytics. *Ease of use* is defined as the mechanism that explains how easily the system can be used to perform big data analytics. *Data management* is discussed in Section 1.2. Table 1.1 shows the comparisons of bio-inspired algorithms for big data analytics based on different parameters.

### 1.3.4 DISCUSSIONS

Table 1.1 shows the comparisons of bio-inspired algorithms for big data analytics based on different parameters, which helps the reader to choose the most appropriate bio-inspired algorithm. In the current scenario, cloud computing has emerged as the fifth utility of computing and has captured 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 minimal possible network latency and response time. Therefore, fog or edge devices can also perform big data analytics at the edge device instead of at a decentralized database or server.

## 1.4 FUTURE RESEARCH DIRECTIONS AND OPEN CHALLENGES

Bio-inspired algorithm-based big data analytics has several challenges that need to be addressed, such as resource management, usability, data processing, elasticity, resilience, heterogeneity in interconnected clouds, sustainability, energy efficiency, data security, privacy protection, edge computing, and networking.

### 1.4.1 RESOURCE SCHEDULING AND USABILITY

Cloud resource management is the ability of a computing system to schedule available resources to process user data over the Internet. The cloud uses virtual resources for big data analytics to process user data quickly and cheaply. The virtualization technology provides effective management of cloud resources using bio-inspired algorithms to improve user satisfaction and resource utilization. There is a

**Table 1.1 Comparison of Bio-Inspired Algorithms for Big Data Analytics**

| Technique | Scalability | Storage | Fault Tolerance | Agility | Virtualization | Cost | Ease of Use | Type of Analytics | No SQL DBMS | Mechanism | Type of Data | Dimension of Data Management | Data Mining Technique |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ABC [9] | ✖ | √ | ✖ | √ | ✖ | ✖ | √ | Audio | Cassandra | Reactive | Audio | Volume, variety, velocity | Prediction |
| FSO [18] | √ | ✖ | ✖ | ✖ | √ | ✖ | ✖ | Social | MongodB | Reactive | Social | Volume, variability | Classification |
| FSOH [19] | ✖ | √ | √ | ✖ | ✖ | √ | ✖ | Video | Hbase | Proactive | Video | Variety, velocity, veracity | Clustering |
| SA [14] | ✖ | ✖ | √ | √ | √ | ✖ | ✖ | Text | Neo4J | Proactive | Text | Volume, velocity | Clustering |
| SAFS [15] | √ | ✖ | ✖ | ✖ | √ | ✖ | ✖ | Predictive | Hbase | Reactive | Operational | Volume, velocity | Prediction |
| FSOSAH [16] | ✖ | √ | ✖ | ✖ | ✖ | √ | ✖ | Social | Cassandra | Proactive | Social | Variability, veracity, variability | Classification |
| PSO [20] | √ | ✖ | ✖ | √ | ✖ | ✖ | √ | Audio | Neo4J | Proactive | Audio | Variability, velocity | Prediction |
| PCPSO [21] | ✖ | ✖ | ✖ | ✖ | ✖ | ✖ | ✖ | Predictive | Cassandra | Proactive | Cloud service | Variety, veracity | Association |
| PS2O [35] | √ | ✖ | ✖ | ✖ | ✖ | ✖ | ✖ | Predictive | Hbase | Reactive | M2M Data | Volume, velocity, variability | Clustering |
| CSO [22] | ✖ | ✖ | √ | ✖ | √ | ✖ | √ | Text | Cassandra | Proactive | Text | Volume, velocity, variability | Prediction |
| SI [23] | √ | ✖ | √ | ✖ | √ | ✖ | ✖ | Video | MongodB | Proactive | Video | Variety, veracity | Prediction |
| CO [17] | √ | ✖ | ✖ | √ | ✖ | √ | ✖ | Predictive | Cassandra | Proactive | Operational | Variability, variety, variability | Association |
| GA [10] | ✖ | √ | ✖ | ✖ | ✖ | √ | √ | Predictive | Couchbase | Reactive | M2M Data | Veracity, variability | Prediction |

*Continued*

**Table 1.1 Comparison of Bio-Inspired Algorithms for Big Data Analytics—cont'd**

| Technique | Scalability | Storage | Fault Tolerance | Agility | Virtualization | Cost | Ease of Use | Type of Analytics | No SQL DBMS | Mechanism | Type of Data | Dimension of Data Management | Data Mining Technique |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ACO [24] | ✖ | √ | ✖ | √ | ✖ | ✖ | √ | Predictive | MongodB | Proactive | Transactional | Volume, variability, velocity | Clustering |
| IACO [25] | √ | ✖ | ✖ | ✖ | √ | √ | ✖ | Social | Cassandra | Proactive | Social | Veracity, variability, velocity | Association |
| DE [12] | ✖ | √ | √ | ✖ | ✖ | ✖ | ✖ | Video | Hbase | Reactive | Video | Volume, velocity, variety | Clustering |
| GP [11] | ✖ | ✖ | √ | √ | ✖ | ✖ | ✖ | Audio | Neo4J | Reactive | Audio | Volume, veracity, velocity | Association |
| ES [13] | ✖ | √ | ✖ | √ | ✖ | ✖ | √ | Predictive | Cassandra | Proactive | Cloud service | Velocity, variability | Classification |
| SFL [26] | √ | ✖ | ✖ | √ | ✖ | ✖ | ✖ | Predictive | MongodB | Proactive | Operational | Velocity, veracity, variability | Classification |
| FSW [27] | √ | ✖ | √ | ✖ | ✖ | ✖ | √ | Audio | Couchbase | Proactive | Audio | Variety, veracity | Prediction |
| IWD [28] | ✖ | ✖ | √ | ✖ | ✖ | ✖ | ✖ | Text | Hbase | Reactive | Text | Volume, velocity, variety | Clustering |
| BFO [29] | √ | ✖ | √ | ✖ | ✖ | ✖ | √ | Predictive | Cassandra | Reactive | Transactional | Volume, variability, velocity | Clustering |
| BFON [30] | ✖ | √ | ✖ | ✖ | √ | √ | ✖ | Predictive | Hbase | Proactive | Operational | Velocity, veracity, variability | Prediction |
| AIS [31] | ✖ | ✖ | √ | √ | ✖ | ✖ | √ | Predictive | Cassandra | Proactive | Transactional | Volume, velocity | Association |
| IWC [33] | √ | ✖ | ✖ | ✖ | ✖ | ✖ | ✖ | Text | Couchbase | Proactive | Text | Volume, variability, velocity | Classification |
| BBO [34] | ✖ | √ | ✖ | ✖ | √ | ✖ | √ | Predictive | Cassandra | Reactive | Cloud service | Volume, variety, veracity, variability | Association |
| GSO [32] | ✖ | √ | ✖ | ✖ | ✖ | √ | ✖ | Predictive | MongodB | Reactive | Transactional | Volume, velocity, veracity | Clustering |

need to optimize provisioning of cloud resources in existing bio-inspired algorithms for big data analytics. To solve this challenge, a quality of service (QoS)-aware bio-inspired algorithm-based resource management approach is required for the efficient management of big data to optimize the QoS parameters.

### 1.4.2 DATA PROCESSING AND ELASTICITY

There is a challenge of data synchronization in bio-inspired algorithms due to data processing that is taking place geographically, which increases overprovisioning and underprovisioning of cloud resources. There is a need to identify the overloaded resources using rapid elasticity, which can handle the data received from different IoT devices. To improve the recoverability of data, there is a need for a data backup technique for big data analytics, which can provide the service during server downtime.

### 1.4.3 RESILIENCE AND HETEROGENEITY IN INTERCONNECTED CLOUDS

The cloud providers such as Microsoft, Amazon, Facebook, and Google are delivering reliable and efficient cloud service by utilizing various cloud resources such as disk drives, storage devices, network cards, and processors for big data analytics. The complexity of computing systems is increasing with an increasing size of cloud data centers (CDCs), which increases the resource failures during big data analytics. The resource failure can be premature termination of execution, data corruption, and service level agreement (SLA) violation. There is a need to find out more information about the failures to make the system more reliable. There is a need for replication of cloud services to analyze the big data in an efficient and reliable manner.

### 1.4.4 SUSTAINABILITY AND ENERGY-EFFICIENCY

To reduce energy consumption, there is a need to migrate user data to more reliable servers for efficient execution of cloud resources. Moreover, introducing the concept of resource consolidation can increase the sustainability and energy efficiency of a cloud service by consolidating the multiple independent instances of IoT applications.

### 1.4.5 DATA SECURITY AND PRIVACY PROTECTION

To improve the reliability of distributed cloud services, there is a need to integrate security protocols in the process of big data analytics. Further, there is a need to incorporate authentication modules at different levels of data management.

### 1.4.6 IoT-BASED EDGE COMPUTING AND NETWORKING

There are a large number of edge devices participating in the IoT-based Fog environment to improve the computation and reduce the latency and response time, which can further increase the energy consumption. Fog devices are not able to offer resource capacity in spite of additional computation and storage power. There is a need to process the user data at an edge device instead of at the server, which can reduce execution time and cost.

## 1.5 EMERGING RESEARCH AREAS IN BIO-INSPIRED ALGORITHM-BASED BIG DATA ANALYTICS

In addition to future research directions, there are various hotspot research areas in bio-inspired algorithm-based big data analytics that need to be addressed in the future such as containers, serverless computing, blockchain, software-defined clouds, bitcoin, deep learning, and quantum computing. In this section, we discuss hotspot research areas in the context of bio-inspired algorithm- based big data analytics.

### 1.5.1 CONTAINER AS A SERVICE (CaaS)

Docker is a container-based virtualization technology that can be used for bio-inspired algorithm-based big data analytics in multiple clouds using a lightweight web server that is, HUE (Hadoop user experience) web interface. The HUE-based Docker container provides a robust and lightweight container as a service (CaaS) data processing facility using a virtual multicloud environment.

### 1.5.2 SERVERLESS COMPUTING AS A SERVICE (SCaaS)

Serverless computing can be used for bio-inspired algorithm-based big data analytics without managing the cloud infrastructure and it is effective in processing user data without configuration of the network and resource provisioning. Serverless computing as a service (SCaaS) has 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 analyze the data quickly.

### 1.5.3 BLOCKCHAIN AS A SERVICE (BaaS)

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

### 1.5.4 SOFTWARE-DEFINED CLOUD AS A SERVICE (SCaaS)

A huge amount of data originates from different IoT devices and it is necessary to transfer data from source to destination without any data 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 a cloud datacenter in a reliable and efficient manner by making intelligent infrastructure decisions. Further, SCaaS offers other advantages for bio-inspired algorithm-based big data analytics such as failure recovery, optimization 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 algorithm-based big data analytics to process the user data with high accuracy and efficiency in a real time manner using hybrid learning and training mechanisms. Deep learning as a service (DLaaS) uses a hierarchical learning process to get high-level, complex abstractions as representations of data for analysis and learning of huge chunks of unsupervised data.

### 1.5.6 BITCOIN AS A SERVICE (BIaaS)

Cryptocurrencies are a very popular technology used to provide secure and reliable service for a 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 algorithm-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 algorithm-based big data analytics to solve complex problems by handling massive digital datasets in an efficient and quick manner. Quantum computing as a service (QCaaS) allows for quick detection, analysis, integration, and diagnosis from large scattered datasets. Further, QCaaS can search extensive, unsorted datasets to quickly uncover patterns.

## 1.6 SUMMARY AND CONCLUSIONS

This chapter presents a review of bio-inspired algorithms for big data analytics. The comparison of bio-inspired algorithms has been presented based on taxonomy, focus of study, and identified demerits. Bio-inspired algorithms are categorized into three different categories and we investigated the existing literature on big data analytics towards finding the open issues. Further, promising research directions are proposed for future research.

## GLOSSARY

**Big data** this is the set of huge datasets, which contains different types 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. 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 the application in a flexible manner, which is running on cloud datacenters. The 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 a processor, networking, and storage and enables virtualization-based computing.

## ACKNOWLEDGMENTS

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. This research work is supported by Discovery Project of Australian Research Council (ARC), Grant/Award Number: DP160102414.

## REFERENCES

[1] S. Khan, X. Liu, K.A. Shakil, M. Alam, A survey on scholarly data: from big data perspective, Inf. Process. Manag. 53 (4) (2017) 923–944.

[2] A. Gandomi, M. Haider, Beyond the hype: big data concepts, methods, and analytics, Int. J. Inf. Manag. 35 (2) (2015) 137–144.

[3] S. Singh, I. Chana, A survey on resource scheduling in cloud computing: issues and challenges, J. Grid Comput. 14 (2) (2016) 217–264.

[4] S.S. Gill, R. Buyya, A taxonomy and future directions for sustainable cloud computing: 360 degree view, ACM Comput. Surv. 51 (6) (2019) 1–37.

[5] C.P. Chen, C.Y. Zhang, Data-intensive applications, challenges, techniques and technologies: a survey on big data, Inf. Sci. 275 (2014) 314–347.

[6] J. Wang, Y. Wu, N. Yen, S. Guo, Z. Cheng, Big data analytics for emergency communication networks: a survey, IEEE Commun. Surv. Tutorials 18 (3) (2016) 1758–1778.

[7] S.S. Gill, I. Chana, M. Singh, R. Buyya, RADAR: self-configuring and self-healing in resource management for enhancing quality of cloud services, in: Concurrency and Computation: Practice and Experience (CCPE), vol. 31, No. 1, Wiley Press, New York, 2019, pp. 1–29, ISSN: 1532-0626.

[8] I. Singh, K.V. Singh, S. Singh, Big data analytics based recommender system for value added services (VAS), in: Proceedings of Sixth International Conference on Soft Computing for Problem Solving, Springer, Singapore, 2017, pp. 142–150.

[9] S.S. Ilango, S. Vimal, M. Kaliappan, P. Subbulakshmi, Optimization using artificial bee colony based clustering approach for big data, Clust. Comput. (2018) 1–9, https://doi.org/10.1007/s10586-017-1571-3.

[10] R. Kune, P.K. Konugurthi, A. Agarwal, R.R. Chillarige, R. Buyya, Genetic algorithm based data-aware group scheduling for big data clouds, in: Big Data Computing (BDC), 2014 IEEE/ACM International Symposium, IEEE, 2014, pp. 96–104.

[11] A.H. Gandomi, S. Sajedi, B. Kiani, Q. Huang, Genetic programming for experimental big data mining: a case study on concrete creep formulation, Autom. Constr. 70 (2016) 89–97.

[12] S. Elsayed, R. Sarker, Differential evolution framework for big data optimization, Memetic Comput. 8 (1) (2016) 17–33.

[13] A.H. Kashan, M. Keshmiry, J.H. Dahooie, A. Abbasi-Pooya, A simple yet effective grouping evolutionary strategy (GES) algorithm for scheduling parallel machines, Neural Comput. & Applic. 30 (6) (2018) 1925–1938.

[14] M.M. Mafarja, S. Mirjalili, Hybrid whale optimization algorithm with simulated annealing for feature selection, Neurocomputing 260 (2017) 302–312.

[15] A. Barbu, Y. She, L. Ding, G. Gramajo, Feature selection with annealing for computer vision and big data learning, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2) (2017) 272–286.

[16] A. Tayal, S.P. Singh, Integrating big data analytic and hybrid firefly-chaotic simulated annealing approach for facility layout problem, Ann. Oper. Res. 270 (1–2) (2018) 489–514.

[17] I.B. Saida, K. Nadjet, B. Omar, A new algorithm for data clustering based on cuckoo search optimization, in: Genetic and Evolutionary Computing, Springer, Cham, 2014, pp. 55–64.

[18] E.D. Raj, L.D. Babu, A firefly swarm approach for establishing new connections in social networks based on big data analytics, Int. J. Commun. Netw. Distrib. Syst. 15 (2-3) (2015) 130–148.

[19] H. Wang, W. Wang, L. Cui, H. Sun, J. Zhao, Y. Wang, Y. Xue, A hybrid multi-objective firefly algorithm for big data optimization, Appl. Soft Comput. 69 (2018) 806–815.

[20] L. Wang, H. Geng, P. Liu, K. Lu, J. Kolodziej, R. Ranjan, A.Y. Zomaya, Particle swarm optimization based dictionary learning for remote sensing big data, Knowl.-Based Syst. 79 (2015) 43–50.

[21] M.S. Hossain, M. Moniruzzaman, G. Muhammad, A. Ghoneim, A. Alamri, Big data-driven service composition using parallel clustered particle swarm optimization in mobile environment, IEEE Trans. Serv. Comput. 9 (5) (2016) 806–817.

[22] K.C. Lin, K.Y. Zhang, Y.H. Huang, J.C. Hung, N. Yen, Feature selection based on an improved cat swarm optimization algorithm for big data classification, J. Supercomput. 72 (8) (2016) 3210–3221.

[23] S. Cheng, Q. Zhang, Q. Qin, Big data analytics with swarm intelligence, Ind. Manag. Data Syst. 116 (4) (2016) 646–666.

[24] S. Banerjee, Y. Badr, Evaluating decision analytics from mobile big data using rough set based ant colony, in: Mobile Big Data, Springer, Cham, 2018, pp. 217–231.

[25] X. Pan, Application of improved ant colony algorithm in intelligent medical system: from the perspective of big data, Chem. Eng. 51 (2016) 523–528.

[26] B. Hu, Y. Dai, Y. Su, P. Moore, X. Zhang, C. Mao, J. Chen, L. Xu, Feature selection for optimized high-dimensional biomedical data using the improved shuffled frog leaping algorithm, IEEE/ACM Trans. Comput. Biol. Bioinform. 15 (2016) 1765–1773.

[27] R.P.S. Manikandan, A.M. Kalpana, Feature selection using fish swarm optimization in big data. Clust. Comput. (2017) 1–13, https://doi.org/10.1007/s10586-017-1182-z.

[28] S. Elsherbiny, E. Eldaydamony, M. Alrahmawy, A.E. Reyad, An extended intelligent water drops algorithm for workflow scheduling in cloud computing environment, Egypt Inform. J. 19 (1) (2018) 33–55.

[29] E.A. Neeba, S. Koteeswaran, Bacterial foraging information swarm optimizer for detecting affective and informative content in medical blogs, Clust. Comput. (2017) 1–14, https://doi.org/10.1007/s10586-017-1169-9.

[30] K. Ahmad, G. Kumar, A. Wahid, M.M. Kirmani, Intrusion detection and prevention on flow of Big Data using bacterial foraging, in: Handbook of Research on Securing Cloud-Based Databases With Biometric Applications, IGI Global, 2014, p. 386.

[31] B. Schmidt, A. Al-Fuqaha, A. Gupta, D. Kountanis, Optimizing an artificial immune system algorithm in support of flow-based internet traffic classification, Appl. Soft Comput. 54 (2017) 1–22.

[32] G. George, L. Parthiban, 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, IEEE, 2015, pp. 125–130.

[33] A.R. Pouya, M. Solimanpur, M.J. Rezaee, Solving multi-objective portfolio optimization problem using invasive weed optimization, Swarm Evol. Comput. 28 (2016) 42–57.

[34] X. Pu, S. Chen, X. Yu, L. Zhang, Developing a novel hybrid biogeography-based optimization algorithm for multilayer perceptron training under big data challenge, Sci. Program. 2018 (2018) 1–7.

[35] S. Fong, R. Wong, A.V. Vasilakos, Accelerated PSO swarm search feature selection for data stream mining big data, IEEE Trans. Serv. Comput. 9 (1) (2016) 33–45.

## FURTHER READING

S.S. Gill, I. Chana, R. Buyya, IoT based agriculture as a cloud and big data service: the beginning of digital India, JOEUC 29 (4) (2017) 1–23.