Cheng *et al. Big Data Analytics* (2016) 1:3
DOI 10.1186/s41044-016-0003-3

Big Data Analytics

REVIEW Open Access

Survey on data science with population-based algorithms

Shi Cheng[1*], Bin Liu[2*], T. O. Ting[3], Quande Qin[4], Yuhui Shi[3] and Kaizhu Huang[3]

*Correspondence:
cheng@snnu.edu.cn; bins@ieee.org
[1] School of Computer Science, Shaanxi Normal University, Xi'an 710119, China
[2] School of Computer Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
Full list of author information is available at the end of the article

Abstract

This paper discusses the relationship between data science and population-based algorithms, which include swarm intelligence and evolutionary algorithms. We reviewed two categories of literature, which include population-based algorithms solving data analysis problem and utilizing data analysis methods in population-based algorithms. With the exponential increment of data, the data science, or more specifically, the big data analytics has gained increasing attention among researchers. New and more efficient algorithms should be designed to handle this massive data problem. Based on the combination of population-based algorithms and data mining techniques, we understand better the insights of data analytics, and design more efficient algorithms to solve real-world big data analytics problems. Also, the weakness and strength of population-based algorithms could be analyzed via the data analytics along the optimization process, a crucial entity in population-based algorithms.

Keywords: Big data analytics, Data analysis, Data science, Evolutionary algorithms, Swarm intelligence, Population-based algorithm

Background

With the amount of data growing constantly and exponentially, the current data processing tasks are beyond the computing ability of traditional computational models. The data science, or more specifically, the big data analytics, has attracted more and more attention among researchers. The data are easily generated and gathered, while the volume of data is increasing very quickly. It exceeds the computational capacity of current systems to validate, analyze, visualize, store, and extract information. To analyze these massive data, there are several kinds of difficulties, such as the large volume of data, dynamical changes of data, data noise, etc. New and efficient algorithms should be designed to handle massive data analytics problems.

Swarm intelligence and evolutionary algorithms are two sets of search and optimization techniques [1–3]. To search a problem domain, a swarm intelligence algorithm processes a population of individuals. Different from traditional single-point based algorithms such as hill-climbing algorithms, each swarm intelligence algorithm is a population-based algorithm, which consists of a set of points (population of individuals). Each individual represents a potential solution to the problem being optimized. The population of individuals is expected to have high tendency to move towards better and better solution areas along iteration through cooperation and competition among themselves.

In this paper, we present the analysis of the relationship from data science to population-based algorithms, which include swarm intelligence and evolutionary algorithms. Swarm intelligence/evolutionary algorithms could be applied to optimize the data mining problems or to handle data directly. In population-based algorithms, individuals move through a solution space and search for solution(s) for the data mining task. The algorithm could be utilized to optimize the data mining problem, e.g., the parameter tuning. The swarm intelligence algorithm could be directly applied to the data samples, e.g., subset data extraction. With the swarm intelligence, more effective methods can be designed and utilized in the massive data analytics problem.

In population-based algorithms, every solution is spread in the search space. Each solution is also a data point; the distribution of solutions can be utilized to reveal the landscape of a problem. Data analysis techniques have been exploited to design new swarm intelligence/evolutionary algorithms, such as brain storm optimization algorithm [4, 5] and estimation of distribution algorithms [6]. In this paper, the population-based algorithms indicate the evolutionary computation algorithms and swarm intelligence algorithms. There are several existing solutions at the same time, and massive information is generated over iterations. Thus, the big data analytics could be utilized to analyze the process of optimization. For non-population based techniques, such as neural networks, a large number of parameters are tuned in different layers, which may also be analyzed by data analytics techniques.

There are three key challenges in big data problems, the data modeling, computing model, and implementation platform. The mainstream of big data research includes computing model, such as deep learning [7], MapReduce [8], and Platform, such as Hadoop [9], and Apache Spark. The aim of this paper is to provide a comprehensive review of the optimization of population-based algorithms on data science and on the analysis of data science method for population-based algorithms. The remaining of the paper is organized as follows. "Data science" Section reviews the basic concepts of data science methods. "Population-based algorithms" Section reviews the general concepts of swarm intelligence and evolutionary algorithms and in particular, four algorithms, which are particle swarm optimization (PSO), ant colony optimization (ACO), brain storm optimization (BSO), and fireworks algorithm (FWA). "Data science with swarm intelligence and evolutionary algorithms" Section reviews the swarm intelligence and evolutionary algorithms utilized to optimize data science methods and data analysis methods utilized to analyze swarm intelligence algorithms are reviewed. A real-world application on freight prediction and recommendation system is introduced in "Freight prediction and recommendation system" Section. The key challenges and future directions of data science with population-based algorithms are discussed in "Key challenges and future directions" Section, followed by conclusions in "Conclusions" Section.

Data science

Currently, data science or data analytics is a popular topic in computer science and statistics. It concerns with a wide variety of data processing tasks, such as data collection, data management, data analysis, data visualization, and real-world applications.

Definition

The data science is a fusion of computer science and statistics. The statistics is the study of the collection, analysis, interpretation, presentation, and organization of data [10]. From the perspective of statistics research, the data science has the same objectives as the statistics, except that the data science emphasizes more on volume, and the variety of data. The data science is more like a synonym of big data research. From the perspective of statistics, there are two aims in data analyses [10]:

- Prediction: To predict the response/output of future input variables;
- Inference: To deduce the association among response variables and input variables.

From the perspective of computer science research, the data science is more practical. The phrase "data mining" is often used to indicate the data science tasks. The process of converting raw data into useful information, termed as knowledge discovery in databases (KDD). Data mining, which is data analysis process of knowledge discovery, attempts to discover useful information (or patterns) in large data repositories [11].

Tasks

In short, the data science field includes many subfields, such as data classification, data clustering, association analysis, and anomaly detection. The data classification and data clustering are two different kinds of basic problems, essential approaches in data mining research.

Classification

Data classification is a problem that finds the correct category (or categories) for objects (i.e., data) when a set of categories (subject, topics) and a collection of data set are given. Data categorization can be considered as a mapping function denoted as $f : \mathcal{D} \rightarrow \mathcal{C}$, which is from the object space $\mathcal{D}$ onto the set of classes $\mathcal{C}$. The objective of a classifier is to obtain an accurate categorization results or predictions with high confidence.

The massive data processing needs to handle many kinds of unstructured or semi-structured data; in some cases, these data need to be transformed into structured data. Each data record with many attributes or features is transformed as a vector with many dimensions. The dimension of the feature space is equal to the number of different attributes that appear in the data set. Different weights can be assigned to each feature. The methods of assigning weights to the features may vary. The simplest is the binary method in which the feature weight is either one – if the corresponding feature is present in the data – or zero otherwise [12, 13]. Recall and precision are two metrics to evaluate the classification results. Both metrics consider the categorized objects and relevant objects together.

Data clustering

The data clustering analysis is a technique that divides data into several groups (clusters). The goal of clustering is to classify objects being similar (or related) to one another in the same cluster and put objects being distant from each other in different clusters [14].

Clustering is the process of grouping similar objects together. From the perspective of machine learning, the clustering analysis is sometimes termed as unsupervised learning. There are N points in the given input, $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{N}$, the "interesting and/or useful pattern" can be obtained through similarity calculation among points [15].

Data Science with Statistics

Statistics is defined as a set of tools such as probability theory, real analysis, measure theory, asymptotic theory, decision theory, Markov chains, martingales, ergotic theory, etc., [16]. In contrast with data mining, the subject of statistics emphasizes much more on mathematical validation. In the field of statistics, the goal of data analysis is to explain data with relevant models, whereas the data mining or machine learning research is more interested in prediction with accuracy and thus does not concern much on the model's interpretability.

In this big data era, there is a big demand to improve traditional statistical data analysis tools, which were used to study observations of instances of particular phenomena. A big data research paradigm claims for analyzing multi-source heterogeneous data in a theoretically sound and consistent way. To face challenges arising from big data analytics problem, we need to integrate advanced techniques in disciplines like metaheuristics based optimization into the statistical toolbox.

As metaheuristics based optimization advocates naturally inspired strategy, whereas statistics stresses mathematical validation, a fusion of them can bring in a lot of new opportunities in proposing new models or algorithms. Some existing trials in fusion of statistics and optimization include the Bayesian simulation scheme mixed with computational intelligence [17], the sequential Monte Carlo optimization methods [18], sequential Monte Carlo simulated annealing [19], sequential Monte Carlo samplers [20], population model-based optimization with sequential Monte Carlo [21], estimation of distribution algorithms [6], etc.

We argue that one of the most promising directions in combining meta-heuristics and statistics is to handle the curse of dimensionality in high-dimension data analysis or optimization problems.

Two Examples of Applications

Massive data are generated in our daily life. With the data analytics techniques and swarm intelligence methods, more effective applications or systems can be designed to solve real-world problems. The intelligent transportation system and wireless sensor networks are two typical examples of massive data analytics applications.

Intelligent transportation system

The traffic problem is arising in many cities now. Many factors affect the traffic and transportation system, e.g., the number of vehicles, weather, accident occurrence, etc., and the traffic information changes in real time. The purpose of intelligent transportation is to build more rapid, safe, and more efficient traffic and transportation system by constructing the intelligent vehicles and road environment [22, 23]. In the intelligent transportation system, there are multiple objectives that ought to be satisfied. For instance, in the case of rapid transportation, environmental pollution, transportation scheduling: there are multi-objective conflicting objectives arising to achieve the predefined goal.

The intelligent transportation problem can be modeled as a large-scale, dynamical, and multi-objective optimization problem. The traditional methods have difficulties to solve this problem. Swarm intelligence has been proven to be an effective method in solving these problems [24] as no gradient information is necessary to find a feasible solution.

Wireless sensor network

Based on the wireless sensor networks, the physical world is turning to be a kind of information system [25]. Different sensors are connected to form a network; information is transferred in this network by communication techniques. The information on the physical surroundings is available easily from sensors in the network. These sensor networks and communication techniques have constructed a new paradigm, which is called the internet of things [25, 26].

The wireless sensor networks have been applied to many real-world problems, such as environmental surveillance, transportation monitoring, engineering surveying, and industrial control, just to name a few [27]. Massive data will be generated from the long-term and/or large-scale wireless sensor network system. The goal of data analysis is to make the fastest possible revelation toward the "useful" information. Data mining techniques are effective ways to handle these data, and to obtain "useful" information [28, 29].

Population-based algorithms

Many real-world applications can be represented as optimization problems where algorithms are required to have the capability to search for the optimum. Most traditional methods can only be applied to continuous and differentiable functions [5]. The meta-heuristic algorithms are proposed to solve the problem, which the traditional methods cannot solve or, at least, be difficult to solve. Recently, kind of meta-heuristic algorithms, termed as swarm intelligence (SI) and evolutionary algorithm (EA), are attracting more and more attentions from researchers.

Swarm intelligence and evolutionary algorithms are collections of population-based searching techniques [1]. To search a problem domain, an SI or EA algorithm processes a population of individuals. Each individual represents a potential solution of the problem being optimized. In EA or SI, an algorithm maintains and successively improves a collection of potential solutions until some stopping condition is met. The solutions are initialized randomly in the search space, and are guided toward the better and better areas through the interaction among solutions.

Swarm intelligence and evolutionary algorithms are two kinds of population-based and nature-inspired algorithm for optimization techniques. The difference is that swarm intelligence is roughly based on mimic of species interaction such as fish school, birds flock, and ant swarm, while evolutionary algorithms is roughly based on mechanisms of evolution such as biological genetics and natural selection. The term "iteration" is used in SI and the "generation" is adopted in EAs to represent individuals at a time. The SI and EAs have the common paradigms in the optimization as follows.

- A population of individuals (potential solutions) is utilized in the search process.
- The "fitness" information, i.e., the "quality" of solutions, instead of function derivatives or other related knowledge are directly used in the search.
- The probabilistic, rather than deterministic, update rules are used to improve the "quality" of solutions.

As a general principle, the expected fitness of a solution returned should improve as the search method is given more computational resources in time and/or space. More desirable, in any single run, the quality of the solution returned by the method over iterations

should improve monotonically – that is, the fitness of the solution at time $t + 1$ should be no worse than the fitness at time t, i.e., $fitness(t + 1) \leq fitness(t)$ for minimization problems. The general procedure of swarm intelligence/evolutionary algorithms is given in Algorithm 1.

Algorithm 1: General procedure of swarm intelligence/evolutionary algorithms

```
Generate random solutions for an optimized problem, repair solutions if solutions
  violate any of the constraints;
Initialize all individuals in the population;
Evaluate all initialized individuals;
while the stopping criteria is not satisfied do
    for all individuals in the population do
        Reproduce individuals to form a new population;
        Evaluate the fitness of each solution;
        Select solutions with better fitness values;
        Update non-dominate solutions in the archive;
```

There exist many swarm intelligence algorithms; among them most common ones are the particle swarm optimization (PSO) algorithm [30], which was originally designed for solving continuous optimization problems, and the ant colony optimization (ACO) algorithm, which was originally designed for discrete optimization problems [31]. The Brain Storm Optimization (BSO) algorithm and Fireworks algorithms (FWA) are two recently proposed swarm intelligence algorithms that are based on the convergence and divergence of solutions. Both BSO and FWA algorithms have the same operators: divergence operator and convergence operator; but two operators are utilized in different sequence.

Particle swarm optimization

Particle Swarm Optimization (PSO), which is one of the swarm intelligence techniques, was invented by Eberhart and Kennedy in 1995 [30, 32]. It is a population-based stochastic algorithm modeled on the social behaviors observed in flocking birds. Each particle, representing a solution, flies through the search space with a velocity that is dynamically adjusted according to its own and its companion's historical behaviors. The particles tend to fly toward better and better search areas over the course of the search process [1, 33].

In the particle swarm optimization problem, a particle not only learns from its own experience, but also learns from its companions. It indicates that a particle's 'moving position' is determined by its own experience and its neighbors' experience [34]. The general procedure of PSO algorithm is given in Algorithm 2.

Ant colony optimization

Ant Colony Optimization (ACO) is another type of swarm intelligence, which takes inspiration from the foraging behavior of some ant species [2, 31]. These ants deposit a chemical called pheromone on the ground, and other ants tend to choose routes with strong pheromone concentration. When an ant finds a short route, the signal

Algorithm 2: Procedure of particle swarm optimization algorithm

```
Initialize velocity and position randomly for each particle;
while the stopping criteria is not satisfied do
    Calculate each particle's fitness value;
    Determine each particle's best position, and the best position of entire swarm;
    for each particle do
        Update particle's velocity;
        Update particle's position;
```

of pheromone can mark some favorable path that should be followed by other members of the colony. Ant colony optimization exploits a similar mechanism for solving optimization problems.

In the ant colony optimization problem, a group of artificial ants will build many solutions to an optimization problem at the same time, and the search information is exchanged on their quality (fitness) via a communication scheme. The general procedure of ACO algorithm is given in Algorithm 3.

Algorithm 3: Procedure of ant colony optimization algorithm

Input: Initialize a set of solutions; Initialize pheromone values (τ)

```
while termination condition is not met do
    for each ant i = 1 to N do
        Construct a solution for each ant;
        Apply a local search for each solution (optimal);
        Update the best-so-far solution s_bs;
    Utilized a pheromone updating strategy
```

Output: The best-so-far solution s_{bs}.

Brain storm optimization

The brain storm optimization (BSO) algorithm was proposed in 2011 [4, 5], which is a young and promising algorithm in swarm intelligence. It is based on the collective behavior of human being, that is, the brainstorming process [4, 5]. The solutions in BSO are converging into several clusters. The best solution of the population will be kept if the newly generated solution at the same index is not better. New individuals can be generated based on the mutation of one or two individuals in clusters. The exploitation ability is enhanced when the new individual is close to the best solution so far. While the exploration ability is enhanced when the new individual is randomly generated, or generated by individuals in two clusters. The procedure of BSO algorithm is given in Algorithm 4.

Fireworks algorithm

Fireworks algorithm (FWA) is a swarm intelligence optimization method that mimics the explosion process of fireworks [35–37]. The procedure of fireworks algorithm is given in

Algorithm 4: Procedure of the brain storm optimization algorithm

Initialization: Randomly generate n potential solutions (individuals), and evaluate them;
while *have not found "good enough" solution or not reached the pre-determined maximum number of iterations* **do**
  Clustering: Cluster n individuals into m clusters by a clustering algorithm;
  New individuals' generation: randomly select one or two cluster(s) to generate new individual;
  Selection: The newly generated individual is compared with the existing individual with the same individual index; the better one is kept and recorded as the new individual;
  Evaluate all individuals;

Algorithm 5. There are four operators/strategies in FWA, which are explosion operator, mutation operator, mapping strategy, and selection strategy, respectively.

The most important factor, which affects a swarm intelligence/evolutionary algorithm's performance may be its ability of exploration and exploitation [34]. Exploration means the ability of a search algorithm to explore different areas of the search space to have the high probability to find good promising solutions. Exploitation, on the other hand, means the ability to concentrate the search around a promising region to refine a candidate solution. A good optimization algorithm should optimally balance the two conflicted objectives.

Data science with swarm intelligence and evolutionary algorithms

Data science concerns the extraction of useful information from raw and massive data. It contains several processes on the data, such as collection, management, data analysis, model building, and visualization. Swarm intelligence is a relatively new subfield of computational intelligence which studies the collective intelligence in a group of simple individuals. In the swarm intelligence, the good optimization results could be obtained from the competition and cooperation of individuals. Figure 1 gives a simple illustration on the connection between data analysis and swarm intelligence/evolutionary algorithms.

Algorithm 5: Procedure of fireworks algorithm

Initialize n locations;
while *have not found "good enough" solution or not reached the pre-determined maximum number of iterations* **do**
  Set off fireworks at n locations;
  Generate sparks through **explosion** operator;
  Generate sparks through **mutation** operator;
  Obtain the locations of sparks;
  Mapping the locations into feasible search space;
  Evaluate the quality of the locations;
  Select n locations as new fireworks;

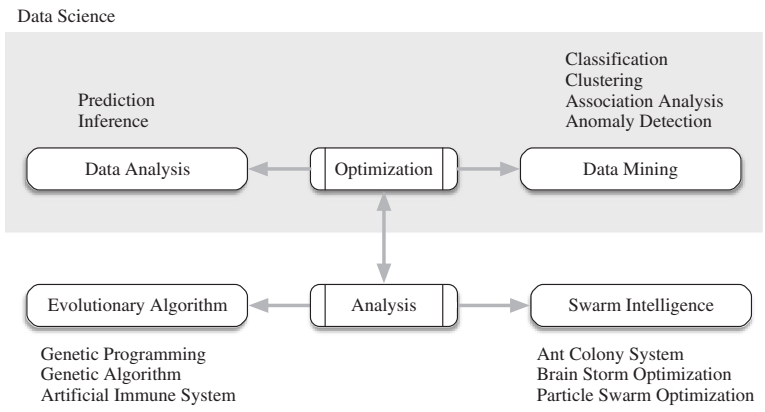

Fig. 1 Relationship between data science with evolutionary algorithm and swarm intelligence. Figure 1 gives a simple illustration on the connection between data analysis and swarm intelligence/evolutionary algorithms. Data science concerns the extraction of useful information from raw and massive data. It contains several processes on the data, such as collection, management, data analysis, model building, and visualization. Swarm intelligence is a relatively new subfield of computational intelligence which studies the collective intelligence of a group of simple individuals. In the swarm intelligence, the good optimization results could be obtained from the competition and cooperation of individuals. The aim of data science with evolutionary algorithms/swarm intelligence is to combine the strengths of evolutionary algorithms/swarm intelligence and data science techniques. It has two meanings: potential applications of evolutionary computation algorithms in data science and data analytics techniques in enhancing evolutionary computation algorithms. With the data analytics during the optimization process, the relationship between the algorithm and problems could be revealed. Evolutionary algorithms/swarm intelligence could be utilized to solve real-world big data analytics problems

Generally, there are two kinds of approaches that apply population-based algorithms as data science techniques [38]. The first category consists of techniques where individuals of a swarm move through a solution space and search for solution(s) for the data mining task. This is a search approach; the swarm intelligence is applied to optimize the data mining problem, e.g., the parameter tuning. In the second category, swarms move data instances that are placed on a low-dimensional feature space to reach a suitable clustering or low-dimensional mapping solution of the data. This is a data organizing approach; the swarm intelligence is directly applied to the data samples, e.g., dimensionality reduction of the data.

The data science could also be utilized in the swarm intelligence/evolutionary algorithms or other optimization techniques. The most common application is to utilize the data analysis methods to improve the optimization techniques. The classification and clustering algorithms could be embedded into optimization algorithms to enhance the population diversity or to accelerate the convergence of solutions. The optimization techniques are also benefited from data visualization methods. For the multi/many-objective optimization problems, several solutions are distributed on Pareto front, which may have three or higher dimensions. A clear illustration on the solutions could help the decision makers have a good understanding on the solved problems.

Population-based algorithms in data science

Many real-world problems could be modeled as optimization problems, which have to find the optimum for one or several objective(s). The opportunities and challenges of evolutionary algorithms solving complex engineering optimization (CEO) problems are introduced in [39]. The optimization problems also happen in data mining tasks or

techniques, such as parameter tuning and minimum subset extraction. Swarm intelligence, especially particle swarm optimization or ant colony optimization algorithms, is utilized in data mining to solve single objective [40] and multi-objective problems [41]. Based on the two characters of particle swarm, the self-cognitive and social learning, the particle swarm has been utilized in data clustering techniques [42], document clustering, variable weighting in clustering high-dimensional data [43], semi-supervised learning based text categorization, and the Web data mining [44]. The potential and possibilities of swarm intelligence algorithms in solving big data analytics problem were summarized in [45, 46]. Table 1 gives a list of applications that population-based algorithms solved problems in data science.

Particle swarm optimization

Since the invention of the PSO algorithm in 1995, it has attracted many attentions in the swarm intelligence research community. Around 650 PSO application papers were analyzed in 2008; all applications were categorized into 26 classes, such as clustering and classification, control, design [47]. Several classes of applications could be regarded as PSO solving problems in data science, for example, approximately 29 papers (4.3 %) are categorized into clustering, classification, and data mining; 2.9 % of the papers on applications were utilized to solve problems with prediction and forecasting; and the applications on modeling and financial may also be related to the data science. PSO variants are also utilized to solve different problems in data science.

Clustering A method, named particle swarm clustering (PSC) algorithm, was proposed to solve data clustering problems [42]. A set of particles in the input data space is generated by the PSC algorithm, and the particles are moved so that they could become the prototypes representing the natural clusters of the original data [42]. In [43], PSO was utilized to solve the variable weighting problem in projected clustering of high-dimensional data.

Association rules mining A PSO variant based on the notion of rough patterns, named rough PSO algorithm (RPSOA), has used rough values defined with upper

Table 1 The applications of population-based algorithms solved problems in data science

Population-based algorithms	Data science
Particle swarm optimization	Clustering [42, 43]
	Association rules mining [48]
	Classification [13, 49, 50]
	Data modeling [51], BioInformatics [52]
Ant colony optimization	Classification [53–56]
Artificial immune system	Classification [57, 58]
	Anomaly detection [59–61]
Genetic algorithms	Classification [62–64]
	Clustering [65–69]
	Web mining [70]
	Association rules mining [71–73]
Genetic programming	Regression [74]
	Classification [75–77]

and lower intervals that represent a range or set of values. RPSOA has been used to solve data mining problems, especially in the automatic mining of numeric association rules [48].

Classification Spam detection could be seen as a special case of classification problems. Based on clonal principle in the natural immune system, the clonal particle swarm optimization (CPSO) algorithm was utilized in solving spam detection problems [49]. By cloning the best individual of successive generations, the CPSO algorithm could enlarge the area near the promising candidate solution and accelerate the evolution of solutions [49]. A PSO algorithm based semi-supervised learning method was engaged to categorize Chinese text in [13]. A prototype generation method based on multiobjective PSO was developed to improve the performance of nearest neighbor classification techniques [50].

For other data mining problems, a PSO algorithm aided orthogonal forward regression was utilized in solving unified data modeling problem [51]; and a binary PSO algorithm proved effective to select the small subset of informative genes from gene expression data [52].

Ant colony optimization

ACO algorithm was originally designed for discrete optimization problems. It has also been applied in different kinds of data mining problems, such as classification rules extraction, decision trees induction.

An ACO algorithm was used to induce decision trees method in [53]. An algorithm for data mining called Ant-Miner (ant-colony-based data miner) was proposed to extract classification rules from data [54]. A new sequential covering strategy for inducing classification rules with ant colony algorithms was introduced [55]. The vast majority of ACO algorithms for inducing classification rules by an ACO-based procedure are proposed [56].

Artificial immune system

An artificial immune system uses principles in the operation of the human immune system and applies them to computationally intelligent systems. Artificial immune system (AIS) has been exploited to solve various data mining problems, especially classification or anomaly detection problems.

Classification A problem-oriented approach by designing an AIS algorithm for data mining, especially for classification, was advocated in [57]. An artificial immune system based multiclass classifier, named artificial immune system with local feature selection (AISLFS), was introduced in [58]. The local feature selection mechanism was embedded to reduce the dimensionality of the problem.

Anomaly detection Artificial immune system has been utilized to solve real-world anomaly detection problems, which are related to information security. A hybrid system based on artificial immune system and the self-organizing map was introduced to solve network intrusion detection problems [59]. An artificial immune system based virus detection system (VDS) was introduced in [60]. An artificial immune system based method for spam filtering was introduced in [61].

Genetic algorithms

Genetic algorithms have been utilized to solve many kinds of data mining problems, such as classification, clustering, Web mining, and association rules mining.

Classification Genetic algorithms have been used to construct a compact fuzzy classification system consisting of a small number of linguistic classification rules [62]. Two objectives are optimized by this fuzzy classification system, one is to maximize the number of correctly classified training patterns and the other is to minimize the number of selected rules. A genetic algorithm has been utilized to solve multi-label classification [63]. The goal of the multi-label classification task is to learn a classifier that predicts multiple class labels to an unlabeled instance based on features of an instance. A real-coded genetic algorithm (RCGA) was proposed to improve the classification performance of a polynomial neural network (PNN) [64]. The mean classification accuracy (CA) is used as the fitness value of each solution for the training dataset.

Clustering Genetic algorithm has been used to optimize the clusters created during unsupervised clustering [65]. The solutions are consisted by hard partitions of the feature space, and the fitness function is a version of the hard c-means optimization function. In [66], genetic algorithms were utilized to search for the optimal, in the least squares sense, hierarchical clustering of a dataset. A multiobjective genetic algorithm based approach was proposed for fuzzy clustering of categorical data [67]. The fuzzy compactness and fuzzy separation of the clusters are optimized at the same time. A multiobjective algorithm, named DYNMOGA (DYNamic MultiObjective Genetic Algorithms) was proposed to solve community discovery problems in dynamic networks [68]. The aims of this proposed algorithm are to maximize cluster accuracy with respect to incoming data of the current time step, and to minimize clustering drift from one time step to the successive one. Two objectives are optimized by DYNMOGA, the first is the maximization of cluster accuracy, i.e., to maximize the snapshot quality of the current time step, and the second is the minimization of clustering drift, *i.e.*, to minimize the temporal cost, which measures the distance between two clusters from one time step to the successive one. A genetic algorithm with spectral-based methodologies, named GANY was proposed to deal with the large data analysis problem [69]. The goal of GANY is generating a method to analyze more data using less resource.

Web mining Genetic algorithms were used for data-driven Web question answering problems, which need to find exact answers to natural languages (NL) questions. Answers are extracted directly from the N-best snippets, which have been identified by a standard Web search engine using NL questions [70].

Association rules mining Association rules mining problem is usually modeled as a multiobjective optimization problem. A literature reviews on multiobjective genetic algorithms and multiobjective genetic programming for rule knowledge discovery in data mining are given in [71]. An automated clustering method based on multiobjective genetic algorithms was proposed to solve fuzzy association rules mining problems [72]. The method is applied to decide on the number of fuzzy sets and for the autonomous mining of both fuzzy sets and fuzzy association rules. The goal of the method is to obtain a large number of item-set in less time by automatically cluster values of a quantitative attribute. A multiobjective genetic algorithm based approaches were utilized for mining optimized fuzzy association rules [73]. Two different forms of criterion are used: the one tries to determine the appropriate fuzzy sets of quantitative attributes in a pre-specified rule, which is also called as certain rule, and the other deals with finding both uncertain rules and their appropriate fuzzy sets.

Genetic programming

The genetic programming (GP) has the similar operators to the genetic algorithm (GA), which includes crossover, mutation and selection. The difference between genetic programming and genetic algorithm is that the GP has a population of tree-shaped individuals, while the GA has a population of string-shaped individuals [1]. GP has been utilized to solve many kinds of data analysis problems.

Regression Symbolic regression based on Pareto Front GP is utilized to generate empirical models for industrial applications [74]. From the results of a small-sized industrial data set, the optimal settings of three parameters: the number of cascades, the number of generations, and the population size are tuned based on Pareto front GP.

Classification Performance bias may occur due to the unbalanced data sets, which may lead classifiers have good accuracy on the majority class, but very poor accuracy on the minority class. A multi-objective genetic programming (MOGP) is proposed, with accurate and diversified classifiers, performed satisfactory on both minority and majority classes [75, 76]. New fitness functions in GP for binary classification with unbalanced data are introduced in [77].

More literature about the evolutionary algorithms for clustering or data mining problems could be found in [78–80].

Data analysis in population-based algorithms

The data mining techniques could be applied to design or analyze swarm intelligence algorithms. Massive information exists during the search process. For swarm intelligence/evolutionary algorithms, there are several individuals existing at the same time, and each individual has a corresponding fitness value. New individuals are generated as the iteration increases. There is also a massive volume of information on the "origin" of an individual, such as that an individual was created by applying which strategy and parameters to which former individual(s). The data-driven evolutionary computation/swarm intelligence is a new approach to analyze and guide the search in evolutionary algorithms/swarm intelligence. These strategies could be divided into off-line methods and online methods. An off-line method is based on the analysis of previous storage search history, such as history based topological speciation for multimodal optimization [81] or maintaining and processing submodels (MAPS) based estimation of distribution algorithm on multimodal problems [82]. In comparison, for an online method, the parameters could be adaptively changed during the different search states.

The data modeling methods could be applied to inspire new swarm intelligence algorithms. In the brain storm optimization algorithm [5], every solution is spread in the search space. The distribution of solutions can be utilized to reveal the landscape of a problem. From the clustering analysis of solutions, the search results can be obtained. In the estimation of distribution algorithms [6], the space of potential solutions is explored by building and sampling explicit probabilistic models of promising candidate solutions.

Data clustering method could be utilized to improve the performance of swarm intelligence/evolutionary algorithms. A cluster and gradient-based artificial immune system is proposed to apply in optimization scenarios [83], and a clustering-based adaptive crossover and mutation probabilities for genetic algorithms are proposed in [84]. The number of clusters was analyzed in BSO algorithm solving different kinds of problems [85]. Based on clusters analysis, the search status could be obtained.

Freight prediction and recommendation system

International trade plays an important role for a country's economy. The parts/goods are transported among countries every day. A simple description of the business model for freight transportation by ships is shown in Fig. 2. The parts and goods are transported from a factory to another country through the freight forwarders and the carrier company. From a cost perspective, shipping companies pay a price to the carriers or freight forwarders, then the carriers or freight forwarders add a markup for profit and might hire other forwarders. It should be noticed that there may be more than one freight forwarders for one goods.

The freight price from a port to the same destination changes at every schedule. Taking the route of Ningbo to Kobe (Japan) as an example, there are five ships for every week. The prices may change considerably at each schedule. Table 2 gives an example of prices in a week, and the visual description of prices in one week is shown in Fig. 3a. There is no significant correlation among five ships. However, the changing pattern of the prices can be found from the previous prices for the same schedule. Table 3 gives an example of price changes for ten-time shipments for the same schedule, e.g., shipments on Monday. Figure 3b shows the curve of changing prices in the last ten times. From Fig. 3, the price of a specific shipment is more related to the previous prices on the same schedule. However, this situation is only for this special short distance case. For other freight forwarding courses, the prices in one week and the previous prices at the same schedule may have different influence on future price.

There are several primary carrier companies in China, and the number of freight forwarders is more than thirty thousand. The freight price changes at each week. It is difficult to predict the future freight market price and give different forwarders a recommended price. Currently, with the assistance of big data analytics techniques, the large amount of historical shipping and trading data could be analyzed to help companies make predictions.

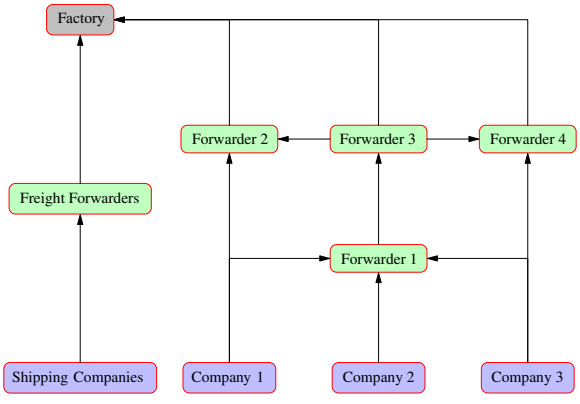

Fig. 2 A simple flowchart of international logistics process. Figure 2 shows a simple description of the business model for freight transportation by ships. International trade plays an important role for a country's economy. The parts/goods are transported among countries every day. The parts and goods are transported from a factory to another country through the freight forwarders and the carrier company. From a cost perspective, shipping companies pay a price to the carriers or freight forwarders, then the carriers or freight forwarders add a markup for profit and might hire other forwarders. It should be noticed that there may be more than one freight forwarders for one goods

Table 2 The price changes for a week

Date	Mon	Tue	Wed	Fri	Sun
Price	60	0	40	0	50

The freight shipment system has to handle a large volume of multi-dimensional data, dynamically changing data, and multiple objectives. There are thousands of freight forwarders on the current system. A visitor from factories or freight companies can search the price of forwarding courses. The search keywords contain source port, destination port, the schedule of ships, the name of shipping company, etc. The search could be one

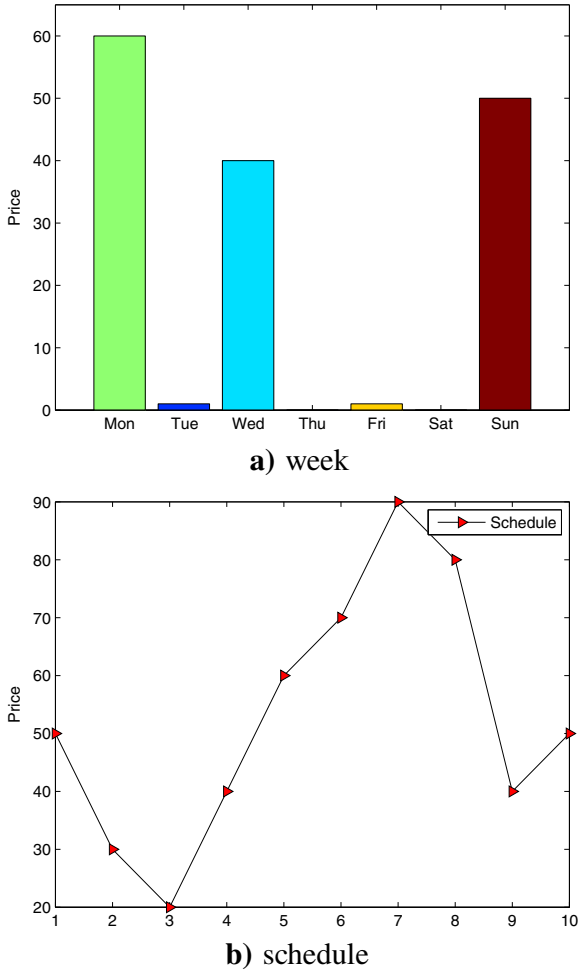

Fig. 3 Freight changes in a week or ten times for a ship in the same schedule. Figure 3 shows the freight changes in a week or ten times for a ship in the same schedule. The freight price from a port to the same destination changes at every schedule. Taking the route of Ningbo to Kobe (Japan) as an example, there are five ships for every week. The prices may change considerably at each schedule. Figure 3 (**a**) gives the visual description of prices in one week. There is no significant correlation among five ships. However, the changing pattern of the prices can be found from the previous prices for the same schedule. Figure 3 (**b**) shows the curve of changing prices in the last ten times. From Figure 3, the price of a specific shipment is more related to the previous prices on the same schedule. However, this situation is only for this special short distance case. For other freight forwarding courses, the prices in one week and the previous prices at the same schedule may have different influence on the future price

Table 3 The price changes for the same schedule at ten times

Run	1	2	3	4	5	6	7	8	9	10
Price	50	30	20	40	60	70	90	80	40	50

keyword or the combination of keywords. The visitors' identification, login time, logout time are also recorded in the system.

Based on the analysis among the browsing data, search data and trading data, the hot course, potential quantity of goods for different courses, the future market price (based on regions, on ports, on historical data of a single company) could be predicted. The future market price for different courses could be predicted. From the predicted price, the potential quantity of goods, and the capacity of forwarders, the system could give a recommendation price to obtain the largest benefits for different forwarders.

The data analytics and optimization are combined together in this case. This real-world application could be modeled as a dynamic, large scale optimization problem. A more detailed example of a commodity routing system in the Ningbo port is introduced in [86]. Based on the data analytics, the optimization model could help the forwarders to obtain the largest profits.

Key challenges and future directions

Key challenges

The big data analytics is a new research area of information processing, however, the problems of big data analytics have been studied in other research fields for decades under a different title. A rough association between big data analytics and evolutionary computation algorithms can be established and shown in Table 4.

There are five accepted properties of big data, which are volume, variety, velocity, veracity, and value. These complexities are a collection of different research problems that existed for decades. Corresponding to the population-based algorithms, the volume and the variety mean large-scale and high dimensional data; the velocity means data is rapidly changing, like an optimization problem in dynamic environment; the veracity means data is inconsistent and/or incomplete, like an optimization problem with noise or approximation; and the value is the objective of the big data analytics, like the fitness or objective function in an optimization problem.

The key challenges of population-based meta-heuristics algorithms solving big data analytics problems could be divided into four elements: handling a large amount of data, handling high dimensional data, handling dynamical data, and multiobjective

Table 4 A rough association between big data analytics and population-based algorithms

Big data analytics	Population-based algorithms
Volume	Large scale/high dimension
Variety	
Velocity	Dynamic environment
Veracity	Noise/uncertain/surrogates
Value	Fitness/objective

optimization. Most real world big data problems can be modeled as a large scale, dynamical, and multiobjective problems.

Future directions

The future direction is combining the strengths of population-based algorithms and big data analytics to design new algorithms on the optimization or data analytics.

Population-based algorithms for big data problems

The big data is created in many areas in our everyday life. The big data analytics problem not only occurs in Internet data mining, but also in complex engineering or design problems [39]. The big data problem could be analyzed from the perspective of computational intelligence and meta-heuristic global optimization [87]. A real-world application could be modeled as a multiobjective, dynamic, large scale optimization problem. It is recognized that the population-based algorithms are good ways to handle this kind of problems. Based on the utilization of swarm intelligence algorithms, the real-world system will be more efficient and effective [45, 86].

Big data analytics for population-based algorithms

A population of individuals in Population-based algorithms is utilized to evolve the optimized functions or goals by cooperative and competitive interaction among individuals. Massive information exists during the search process, such as the distribution of individuals and the fitness of each solution. To improve the search efficiency or to recognize the search state, the data generated in the optimization process should be analyzed.

Conclusions

In swarm intelligence and evolutionary algorithms, a population of individuals is utilized to evolve the optimized functions or goals by cooperative and competitive interaction among individuals. Massive information exists during the search process, such as the distribution of individuals and the fitness of each solution. To improve the search efficiency or to recognize the search state, the data generated in the optimization process should be analyzed.

With the amount of data growing constantly and exponentially, the data processing tasks have been beyond the computing ability of traditional computational models. To handle these massive data, i.e., deal with the big data analytics problem, more effective and efficient methods should be designed. There is no complex mathematical model in swarm intelligence/evolutionary algorithms. The algorithm is updated based on few iterative rules and the evaluation of solution samples. The massive data analytics may be benefited from these properties because massive data are difficult or impossible represented by mathematical models.

This paper has reviewed the connection between data science and swarm intelligence/evolutionary algorithms. The potential combination of data science and swarm intelligence/evolutionary algorithm in optimization and data analytics was also analyzed. Data science involves prediction or inference on a large amount of data. Swarm intelligence studies the collective behaviors in a group of individuals. With the combination of data science, swarm intelligence and evolutionary algorithms, more rapid and effective methods can be designed to solve optimization and data analytics problem.

Abbreviations
ACO, ant colony optimization; AIS, artificial immune system; BSO, brain storm optimization; EA, evolutionary algorithm; EC, evolutionary computation; EDA, estimation of distribution algorithms; FWA, fireworks algorithms; GA, genetic algorithm; GP, genetic programming; PSO, particle swarm optimization; SI, swarm intelligence

Acknowledgements
This work is partially supported by Natural Science Foundation of China under grant no. 71402103, 60975080, 61273367, 61571238, 61302158 and 61473236, the Natural Science Foundation of Jiangsu Province (BK20141005), the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (14KJB520025), Foundation for Distinguished Young Talents in Higher Education of Guangdong, China, under grant 2012WYM_0116 and the MOE Youth Foundation Project of Humanities and Social Sciences at Universities in China under grant 13YJC630123, and Ningbo Science & Technology Bureau (Science and Technology Project No.2012B10055).

Authors' contributions
SC and BL planned and drafted the content of the paper. Both authors, together with TOT and YS, wrote the paper. Other authors participated in critical revision of the manuscript and approved the final submission.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]School of Computer Science, Shaanxi Normal University, Xi'an 710119, China. [2]School of Computer Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing 210023, China. [3]Department of Electrical & Electronic Engineering, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China. [4]Department of Management Science, Shenzhen University, Shenzhen 518060, China.

Received: 28 December 2015 Accepted: 7 April 2016
Published online: 01 July 2016

References
1. Kennedy J, Eberhart R, Shi Y. Swarm Intelligence. San Francisco: Morgan Kaufmann Publisher; 2001.
2. Dorigo M, Stützle T. Ant Colony Optimization. Cambridge: MIT Press; 2004.
3. Eberhart R, Shi Y. Computational Intelligence: Concepts to Implementations. San Francisco: Morgan Kaufmann Publisher; 2007.
4. Shi Y. Brain storm optimization algorithm In: Tan Y, Shi Y, Chai Y, Wang G, editors. Advances in Swarm Intelligence. Lecture Notes in Computer Science, vol. 6728. Berlin Heidelberg: Springer; 2011. p. 303–9.
5. Shi Y. An optimization algorithm based on brainstorming process. Int J Swarm Intell Res (IJSIR). 2011;2(4):35–62.
6. Pelikan M, Goldberg DE, Lobo FG. A survey of optimization by building and using probabilistic models. Comput Optim Appl. 2002;21(1):5–20.
7. LeCun Y, Bengio Y, Hinton G. Deep learning. Nautre. 2016;521:436–44.
8. Dean J, Ghemawat S. Mapreduce: Simplified data processing on large clusters. In: Proceedings of 6th Symposium on Operating Systems Design and Implementation (OSDI 2004); 2004. p. 137–49.
9. White T. Hadoop: The Definitive Guide 4th edn. Sebastopol: O'Reilly Media, Inc; 2015.
10. Donoho DL. 50 years of data science. Technical report, Stanford University. 2015.
11. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Mag. 1996;17(3): 37–54.
12. Cervantes A, Galván IM, Isasi P. AMPSO: A New Particle Swarm Method for Nearest Neighborhood Classification. IEEE Trans Syst Man Cybern B Cybern. 2009;39(5):1082–91.
13. Cheng S, Shi Y, Qin Q. Particle swarm optimization based semi-supervised learning on Chinese text categorization. In: Proceedings of 2012 IEEE Congress on Evolutionary Computation (CEC 2012). Brisbane, Australia: IEEE; 2012. p. 3131–198.
14. Tan PN, Steinbach M, Kumar V. Introduction to Data Mining. Boston: Addison Wesley; 2005.
15. Murphy KP. Machine Learning: A Probabilistic Perspective. Adaptive computation and machine learning series. Cambridge: The MIT Press; 2012.
16. Friedman JH. Data mining and statistics: What's the connection? In: Proceedings of the 29th Symposium on the Interface Between Computer Science and Statistics; 1997. p. 1–7.
17. Liu B, Ji C. A general algorithm scheme mixing computational intelligence with Bayesian simulation. In: Proceedings of the 2013 Sixth International Conference on Advanced Computational Intelligence; 2013. p. 1–6.
18. Liu B. Posterior exploration based sequential Monte Carlo for global optimization. Technical report, Nanjing University of Posts and Telecommunications. 2015.
19. Zhou E, Chen X. Sequential monte carlo simulated annealing. J Glob Optim. 2013;55(1):101–24.
20. Del Moral P, Doucet A, Jasra A. Sequential monte carlo samplers. J R Stat Soc Ser B Stat Methodol. 2006;68(3):411–36.
21. Chen X, Zhou E. Population model-based optimization with sequential monte carlo. In: Proceedings of the 2013 Winter Simulation Conference: Simulation: Making Decisions in a Complex World. Washington: IEEE; 2013. p. 1004–15.
22. Kohata N, Sato M, Yamaguchi T, Baba T, Hashimoto H. Evolutionary computation for intelligent agents based on chaotic retrieval and soft DNA In: McKay B, Yao X, Newton CS, Kim J-H, Furuhashi T, editors. Simulated Evolution and Learning. Lecture Notes in Computer Science, vol. 1585. Berlin Heidelberg: Springer; 1999. p. 251–9.
23. Teodorović D. Transport modeling by multi-agent systems: A swarm intelligence approach. Transp Plan Technol. 2003;26(4):289–312.

24. Li X, Yao X. Cooperatively coevolving particle swarms for large scale optimization. IEEE Trans Evol Comput. 2012;16(2):210–24.
25. Chui M, Löffler M, Roberts R. The internet of things. McKinsey Q. 2010;2:1–9.
26. Atzori L, Iera A, Morabito G. The internet of things: A survey. Comput Netw. 2010;54(15):2787–805.
27. Liu Y, Zhou G, Zhao J, Dai G, Li XY, Gu M, Ma H, Mo L, He Y, Wang J, Li M, Liu K, Dong W, Xi W. Long-term large-scale sensing in the forest: recent advances and future directions of greenorbs. Front Comput Sci China. 2010;4(3):334–8.
28. Kulkarni RV, Venayagamoorthy GK. Particle swarm optimization in wireless-sensor networks: A brief survey. IEEE Trans Syst Man Cybern Part C Appl Rev. 2011;41(2):262–7.
29. Kulkarni RV, Förster A, Venayagamoorthy GK. Computational intelligence in wireless sensor networks: A survey. IEEE Commun Surv Tutor. 2011;13(1):68–96.
30. Kennedy J, Eberhart R. Particle swarm optimization. In: Proceedings of IEEE International Conference on Neural Networks (ICNN 1995); 1995. p. 1942–1948.
31. Dorigo M, Maniezzo V, Colorni A. Ant system: optimization by a colony of cooperating agents. IEEE IEEE Trans Syst Man Cybern B Cybern. 1996;26(1):29–41.
32. Eberhart R, Kennedy J. A new optimizer using particle swarm theory. In: Proceedings of the Sixth International Symposium on Micro Machine and Human Science; 1995. p. 39–43.
33. Eberhart R, Shi Y. Particle swarm optimization: Developments, applications and resources. In: Proceedings of the 2001 Congress on Evolutionary Computation (CEC2001); 2001. p. 81–6.
34. Cheng S, Shi Y, Qin Q. Population diversity of particle swarm optimizer solving single and multi-objective problems. Int J Swarm Intell Res (IJSIR). 2012;3(4):23–60.
35. Tan Y, Zhu Y. Fireworks algorithm for optimization In: Tan Y, Shi Y, Tan KC, editors. Advances in Swarm Intelligence. Lecture Notes in Computer Science, vol. 6145. Berlin Heidelberg: Springer; 2010. p. 355–64.
36. Tan Y. Fireworks Algorithm: A Novel Swarm Intelligence Optimization Method. Berlin Heidelberg: Springer; 2015.
37. Cheng S, Qin Q, Chen J, Shi Y, Zhang Q. Analytics on fireworks algorithm solving problems with shifts in the decision space and objective space. Int J Swarm Intell Res (IJSIR). 2015;6(2):52–86.
38. Martens D, Baesens B, Fawcett T. Editorial survey: swarm intelligence for data mining. Mach Learn. 2011;82(1):1–42.
39. Chai T, Jin Y, Sendhoff B. Evolutionary complex engineering optimization: Opportunities and challenges. IEEE Comput Intell Mag. 2013;8(3):12–15.
40. In: Abraham A, Grosan C, Ramos V, editors. Swarm Intelligence in Data Mining. Studies in Computational Intelligence, vol. 34. Berlin Heidelberg: Springer; 2006.
41. In: Coello Coello CA, Dehuri S, Ghosh S, editors. Swarm Intelligence for Multi-objective Problems in Data Mining. Studies in Computational Intelligence, vol. 242. Berlin Heidelberg: Springer; 2009.
42. Cohen SCM, de Castro LN. Data clustering with particle swarms. In: Proceedings of the 2006 IEEE Congress on Evolutionary Computation (CEC 2006); 2006. p. 1792–8.
43. Lu Y, Wang S, Li S, Zhou C. Particle swarm optimizer for variable weighting in clustering high-dimensional data. Mach Learn. 2011;82(1):43–70.
44. Pal SK, Talwar V, Mitra P. Web mining in soft computing framework: Relevance, state of the art and future directions. IEEE Trans Neural Netw. 2002;13(5):1163–77.
45. Cheng S, Shi Y, Qin Q, Bai R. Swarm intelligence in big data analytics In: Yin H, Tang K, Gao Y, Klawonn F, Lee M, Weise T, Li B, Yao X, editors. Intelligent Data Engineering and Automated Learning - IDEAL 2013. Lecture Notes in Computer Science, vol. 8206. Berlin Heidelberg: Springer; 2013. p. 417–26.
46. Cheng S, Zhang Q, Qin Q. Big data analytics with swarm intelligence. Ind Manag Data Syst. 2015.
47. Poli R. Analysis of the publications on the applications of particle swarm optimisation. J Artif Evol Appl. 2008;2008:1–10.
48. Alatas B, Akin E. Rough particle swarm optimization and its applications in data mining. Soft Comput. 2008;12:1205–18.
49. Tan Y. Particle swarm optimization algorithms inspired by immunity-clonal mechanism and their applications to spam detection. Int J Swarm Intell Res (IJSIR). 2010;1(1):64–86.
50. Hu W, Tan Y. Prototype generation using multiobjective particle swarm optimization for nearest neighbor classification. IEEE Trans Cybern. 2015.
51. Chen S, Hong X, Harris CJ. Particle swarm optimization aided orthogonal forward regression for unified data modeling. IEEE Trans Evol Comput. 2010;14(4):477–99.
52. Mohamad MS, Omatu S, Deris S, Yoshioka M. A modified binary particle swarm optimization for selecting the small subset of informative genes from gene expression data. IEEE Trans Inf Technol Biomed. 2011;15(6):813–22.
53. Otero FEB, Freitas AA, Johnson CG. Inducing decision trees with an ant colony optimization algorithm. Appl Soft Comput. 2012;12(11):3615–26.
54. Parpinelli RS, Lopes HS, Freitas AA. Data mining with an ant colony optimization algorithm. IEEE Trans Evol Comput. 2002;6(4):321–32.
55. Otero FEB, Freitas AA, Johnson CG. A new sequential covering strategy for inducing classification rules with ant colony algorithms. IEEE Trans Evol Comput. 2013;17(1):64–76.
56. Otero FEB, Freitas AA. Improving the interpretability of classification rules discovered by an ant colony algorithm. In: Proceedings of the 15th Annual Conference on Genetic and Evolutionary Computation (GECCO 2013); 2013. p. 73–80.
57. Freitas AA, Timmis J. Revisiting the foundations of artificial immune systems for data mining. IEEE Trans Evol Comput. 2007;11(4):521–40.
58. Dudek G. An artificial immune system for classification with local feature selection. IEEE Trans Evol Comput. 2012;16(6):847–60.
59. Powers ST, He J. A hybrid artificial immune system and self organising map for network intrusion detection. Inf Sci. 2008;178(15):3024–42.

60. Chao R, Tan Y. A virus detection system based on artificial immune system. In: Proceedings of 2009 International Conference on Computational Intelligence and Security (CIS 2009); 2009. p. 6–10.
61. Tan Y, Mi G, Zhu Y, Deng C. Artificial immune system based methods for spam filtering. In: Proceedings of 2013 IEEE International Symposium on Circuits and Systems (ISCAS 2013); 2013. p. 2484–8.
62. Ishibuchi H, Murata T, Türkşen IB. Single-objective and two-objective genetic algorithms for selecting linguistic rules for pattern classification problems. Fuzzy Sets Syst. 1997;89(2):135–50.
63. Gonçalves EC, Plastino A, Freitas AA. Simpler is better: a novel genetic algorithm to induce compact multi-label chain classifiers. In: Proceedings of Annual Conference on Genetic and Evolutionary Computation (GECCO 2015); 2015. p. 559–66.
64. Lin CT, Prasad M, Saxena A. An improved polynomial neural network classifier using real-coded genetic algorithm. IEEE Trans Syst Man Cybern Syst. 2015;45(11):1389–401.
65. Bezdek JC, Boggavarapu S, Hall LO, Bensaid A. Genetic algorithm guided clustering. In: Proceedings of the First IEEE Conference on Evolutionary Computation (CEC 1994); 1994. p. 34–9.
66. Lozano JA, Larrañaga P. Applying genetic algorithms to search for the best hierarchical clustering of a dataset. Pattern Recogn Lett. 1999;20(9):911–8.
67. Mukhopadhyay A, Maulik U, Bandyopadhyay S. Multiobjective genetic algorithm-based fuzzy clustering of categorical attributes. IEEE Trans Evol Comput. 2009;13(5):991–1005.
68. Folino F, Pizzuti C. An evolutionary multiobjective approach for community discovery in dynamic networks. IEEE Trans Knowl Data Eng. 2014;26(8):1838–1852.
69. Menéndez HD, Camacho D. GANY: A genetic spectral-based clustering algorithm for large data analysis. In: Proceedings of the 2015 IEEE Congress on Evolutionary Computation (CEC 2015); 2015. p. 640–7.
70. Figueroa AG, Neumann G. Genetic algorithms for data-driven web question answering. Evol Comput. 2008;16(1): 89–125.
71. Srinivasan S, Ramakrishnan S. Evolutionary multi objective optimization for rule mining: a review. Artif Intell Rev. 2011;36(3):205–48.
72. Alhajj R, Kaya M. Multi-objective genetic algorithms based automated clustering for fuzzy association rules mining. J Intell Inf Syst. 2008;31(3):243–64.
73. Kaya M. Multi-objective genetic algorithm based approaches for mining optimized fuzzy association rules. Soft Comput. 2006;10(7):578–86.
74. Castillo F, Kordon A, Smits G, Christenson B, Dickerson D. Pareto front genetic programming parameter selection based on design of experiments and industrial data. In: Proceedings of Annual Conference on Genetic and Evolutionary Computation (GECCO 2006); 2006. p. 1613–20.
75. Bhowan U, Johnston M, Zhang M, Yao X. Evolving diverse ensembles using genetic programming for classification with unbalanced data. IEEE Trans Evol Comput. 2013;17(3):368–86.
76. Bhowan U, Johnston M, Zhang M, Yao X. Reusing genetic programming for ensemble selection in classification of unbalanced data. IEEE Trans Evol Comput. 2014;18(6):893–908.
77. Bhowan U, Johnston M, Zhang M. Developing new fitness functions in genetic programming for classification with unbalanced data. IEEE IEEE Trans Syst Man Cybern B Cybern. 2012;42(2):406–21.
78. Hruschka ER, Campello RJGB, Freitas AA, de Carvalho ACPLF. A survey of evolutionary algorithms for clustering. IEEE Trans Syst Man Cybern Part C Appl Rev. 2009;39(2):133–55.
79. Mukhopadhyay A, Maulik U, Bandyopadhyay S, Coello Coello CA. A survey of multiobjective evolutionary algorithms for data mining: Part I. IEEE Trans Evol Comput. 2014;18(1):4–19.
80. Mukhopadhyay A, Maulik U, Bandyopadhyay S, Coello Coello CA. Survey of multiobjective evolutionary algorithms for data mining: Part II. IEEE Trans Evol Comput. 2014;18(1):20–35.
81. Li L, Tang K. History-based topological speciation for multimodal optimization. IEEE Trans Evol Comput. 2015;19(1): 136–50.
82. Yang P, Tang K, Lu X. Improving estimation of distribution algorithm on multimodal problems by detecting promising areas. IEEE Trans Cybern. 2015;45(8):1438–49.
83. Honório LdM, da Silva AML, Barbosa DA. A cluster and gradient-based artificial immune system applied in optimization scenarios. IEEE Trans Evol Comput. 2012;16(3):301–18.
84. Zhang J, Chung HS-H, Lo WL. Clustering-based adaptive crossover and mutation probabilities for genetic algorithms. IEEE Trans Evol Comput. 2007;11(3):326–35.
85. Cheng S, Shi Y, Qin Q, Gao S. Solution clustering analysis in brain storm optimization algorithm. In: Proceedings of The 2013 IEEE Symposium on Swarm Intelligence (SIS 2013). Singapore: IEEE; 2013. p. 111–8.
86. Cheng S, Zhang Q, Qin Q. Big data analytic with swarm intelligence. Ind Manag Data Syst. 2016.
87. Zhou ZH, Chawla NV, Jin Y, Williams GJ. Big data opportunities and challenges: Discussions from data analytics perspectives. IEEE Comput Intell Mag. 2014;9(4):62–74.