Research Progress of Intelligent Optimization Algorithm in Big Data Background

DUO ZHANG

ABSTRACT

Big data is the inevitable outcome of the rapid development of modern information technology, effective analysis and processing of big data will not only bring great economic value, but will also promote social development. Under the big data environment, the data scale, speed of emergence and its difficulty make optimizing issues very complex. In recent decades, genetic algorithm (GA), particle swarm optimization (PSO), ant colony optimization (ACO), Artificial Fish School Algorithm, Bacteria Foraging Optimization Algorithm (BFOA), artificial neural networks (ANNs) and other multi-population intelligent algorithms appeared. In this paper, several typical intelligent optimization algorithms are introduced, including genetic algorithm, particle swarm optimization algorithm, ant colony algorithm, artificial fish swarm algorithm and bacterial foraging algorithm. The basic principles of five algorithms are described respectively, along with the direction of improvement and feasible applications.

KEYWORDS

Big data, intelligent optimization algorithm, algorithm optimization, genetic algorithm.

INTRODUCTION

Big data is another revolution following the mobile Internet, Internet of things and cloud computing, which is a new opportunity emerged under the contradiction of the rapid development of information technology and the growing demand for information processing capabilities. There are many different definitions and understandings for its concept.

Big data [1], also known as huge amounts of information, refers to large-scale data that cannot be collected, processed or collated in a reasonable amount of time through software. The ability of data collection, storage, analysis and management of conventional database software cannot meet such a large scale of data. The types and formats of data are also beyond the traditional structural data categories. More and more social prediction, decision-making, business and other activities are based on objective analysis of data.

Duo Zhang, College of Management and Economics, Tianjin University, Tianjin 300072, China.
zhangduo1996@163.com
Big data is becoming a boom, continues to impact business, education, health care, transportation, science and technology and other fields. However, rapid growth of massive data is in sharp contrast to the absence of a comprehensive and effective data analysis method, hence there is a phenomenon - "rich data, poor information", where people are eager to discover useful information from the ocean of data in order to improve the utilization of data, because of which data mining technology is driven to appear.

With the deepening of information technology applications, the development of data mining technology suffered a bottleneck. Firstly, the speed and amount of data generation, the scale and scope of the application are constantly expanding, resulting in large-scale data sets. Since the amount of computation are increasing, some of the original data mining methods cannot excavate useful information in a short period of time, the quality of mining information declines or even cannot be executed. Secondly, the feature dimension of data mining object is getting higher and higher, a large number of high-dimensional small sample data with redundant features and noise characteristics appears, and direct data mining may lead to dimensional disaster. Thirdly, in the process of data mining, there will be many rules, but some of them are not useful information, which requires us to use constraints to guide the direction of data mining, yet the existing mining methods are mostly lack of guidance and control of interactions.

As a new force of the optimization methods, some typical intelligent optimization algorithms have been widely used in data analysis and processing. However, the traditional group intelligence model cannot meet the complex problems such as high dimension, strong constraint and multi-objective optimization in the real big data environment. Therefore, it is very important to design a new intelligent optimization algorithm for solving complex optimization problems in big data environment.

THE BASIC PRINCIPLE OF THE ALGORITHM

Genetic Algorithm, GA.

Genetic algorithm (GA) [2,3] is a random global optimization search algorithm developed by imitating the evolutionary process of biological evolution in nature, which was proposed by J Holland in 1975, in his paper “Adaption in Natural and Artificial Systems”, with the characters of widely application, easy to use, and strong robustness. The basic idea of GA comes from Darwin's theory of evolution and Mendel's genetics. It borrows the biological genetic point of view, and achieve the goal of improving the individual adaptability through natural selection, crossover, mutation and other genetic operations. The algorithm has a group of populations consist of individuals. Each individual has been evaluated in the process of population evolution and has obtained its fitness value. Individuals have evolved to higher fitness under the influence of selection, crossover and mutation operator to achieve the goal of seeking the optimal solution [4]. Figure 1 shows the genetic algorithm flow chart.
Particle Swarm Optimization, PSO.

Particle swarm optimization (PSO) [5] is a relatively new optimization technique, which is similar to the GA algorithm in the computational method, yet it is still different from GA that the PSO algorithm does not use the factors utilized in evolutionary computation, such as hybridization and mutation. It was inspired by the social behaviour of the birds and proposed by biologist Frank Heppner's according to his biota model. It uses non-volume massless particles as individuals, and provides simple rules of social behaviour for each particle, and obtains the search of the optimal solution of problems by collaboration of individuals among populations. Since the algorithm convergence is fast, the amount of parameters is low and the implementation is easy, it can effectively solve the complex optimization problem. GA is widely used in function optimization, neural network training, graphic processing, pattern recognition and some engineering fields.

Ant Colony Optimization, ACO.

The ant colony algorithm was inspired by the ant colony foraging biological behaviour. In the process of foraging, the ants can always find the shortest path between a nest and a food source. Biologists have found that ants will encounter a lot of paths that have never been used in the process of finding food sources, and the ants will randomly select one and release a chemical called pheromone along the way. Ants in the short path will finish the round-trip faster, hence the accumulation of pheromone will be relatively higher and does not volatile easily, on the contrary, in the long path where the accumulation of pheromone is less and will gradually volatile with time passing by. When other ants looking for food sources pass through the intersection, they have a great probability to choose the path with higher pheromone concentration. As shown in Figure 2, point A is the nest, D is the food. Through this positive feedback mechanism, the entire ant colony can find the shortest path between the nest and food source after a period of time [6].
Artificial Fish School Algorithm, AFSA.

The artificial fish swarm optimization algorithm is an intelligent optimization algorithm proposed by Dr. Li Xiaolei in his thesis in 2002[7]. It starts from constructing the simple underlying behaviour of animals, and finds the global optimal solution in the group through the partial optimization of the individual. The biological principle of the algorithm is that in waters, fish can always find nutritious places and stay in clusters, and there is no unified coordinator in the process, but is only achieved by the adaptation of each individual. Artificial fish swarm algorithm uses a bottom-up design idea, which starts from the behaviours of the fish from the bottom. The entire algorithm does not have centralized control, or the experimental knowledge of the problem, neither does the objective function need to be continuous or derivable. Therefore, it is a highly adaptable intelligent optimization algorithm. Same as other group intelligence algorithms, the artificial fish swarm algorithm also possesses parallelism, self-organization and robustness.

Bacteria Foraging Optimization Algorithm, BFOA.

The bacterial foraging algorithm is an intelligent random search algorithm proposed by Passino in 2002. Its biological basis is the intelligent behaviour of Escherichia coli in the human intestine during the foraging process. The foraging behaviour of bacteria has four typical patterns, which are tropism behaviour, aggregation behaviour, replication behaviour and migration behaviour [8].

IMPROVING DIRECTIONS OF ALGORITHM

Coding.

Coding is the mapping of the solution space from phenotype to genotype. A reasonable coding not only enables the individual to express the information well, but is also conducive to the transfer of information between individuals. There are various coding approaches proposed contraposing different issues: (1) Binary coding, which is the most commonly used encoding method of genetic algorithm. It has the advantages of simple coding, easy cross-mutation operation, and good handling of discrete problems, yet it is not accurate when dealing with continuous optimization problems. (2) Gray code coding [9], Gray code coding is a deformation of binary code, where
there is only one different code bit in two consecutive integral number corresponding to the encoding value while others are the same. (3) Real number coding, which is conducive to genetic algorithm for searching in a large space. It can improve the efficiency of genetic algorithms and computational accuracy, but also conducive to the combination of genetic algorithms with other algorithms.

**Selection of important parameters.**

For the genetic algorithm, it has three basic genetic operators: selection operator, crossover operator and mutation operator. The choice of these parameters will directly affect the performance and searching speed of the algorithm. Therefore, choosing the appropriate genetic operator is the key to the global optimal solution. For the PSO algorithm, due to the fixed parameters, the accuracy of optimizing problems of some functions is poor. Later Shi et al. proposed an improved algorithm of inertia factor \( W \) linear decrement, which greatly improved the performance of standard PSO algorithm. Another example is the BFO algorithm, the swimming step size \( C \) is an important parameter which directly affects the performance of the algorithm. Therefore, the improvement of the BFO algorithm can greatly improve the convergence and the stability of the BFO algorithm.

**Improvement based on biological behavior.**

Researchers have mapped social problems into natural biological evolution and have achieved great results. In recent years, they have been working towards improving the performance of algorithms and have been working on the research and behaviour simulation of biological survival mechanisms. For instance, immune evolution of the bacterial foraging algorithm, parasitic behaviour based on the dual population particle swarm algorithm, particle swarm optimization algorithm enlightened by the geese migration.

**Integrated optimization algorithm.**

In the study of optimization algorithm, it is found that the single algorithm often does not show high convergence speed or convergence precision. Therefore, people try to integrate traditional optimization algorithms, or combine traditional optimization algorithms with other technical methods to improve the overall performance of the algorithm. Such as chaotic particle swarm optimization, integrated genetic algorithm of genetic algorithm and simulated annealing, cultural flora algorithm based on multi-objective optimization.

**OPTIMIZATION AND APPLICATION OF ALGORITHM IN BIG DATA ERA**

Large scale of data not only requires higher storage capacity of the data, but also challenges the ability to deal with data. Many scholars realized that traditional data mining algorithms did not achieve better execution efficiency when dealing with large-scale data sets, so they began to study the mining algorithms for large-scale data sets and achieved some results. Dr. Shen Yan from Jiangsu University [10] proposed a PC-SEM algorithm that incorporates semi-supervised thinking, that is, a large-scale
dataset, EM Probability Clustering Algorithm, based on partial constraint information. The algorithm makes the clustering effect and efficiency of large-scale dataset to be improved. Gregory Buehrer et al [11] optimized the original FP-GROWTH algorithm contraposing the large-scale dataset optimization, making the mining algorithm conscious of the input and output. The efficiency of the algorithm is enhanced to meet the needs of large-scale data set mining.

In order to solve the "failure" of data mining technology in high dimensional space, the related scholars have made efforts to improve the data mining algorithm. Chen et al. [12] incorporated the rough set method into the data mining algorithm, and proposed a high-dimensional outlier data mining algorithm based on rough clustering. The algorithm uses the rough set to select the optimal subspace, and conducts data mining of outlier in the subspace. The simulation results show that this algorithm has good performance. Zhang Tianyou [13] proposed a detection algorithm (SLDF) for spatial local outliers, which takes into account the attribute weight of the object, utilizes the space partial deviation factor to measure the degree of outlier of the spatial point. The detection algorithm is used to detect the remaining outliers.

CONCLUSION

It is effective to apply the intelligent optimization algorithm to the data analysis in big data environment. On the one hand, intelligent optimization algorithm utilizes the global search capabilities can optimize the partial optimal data mining models in the process of data processing, in order to improve the data analysis and processing capabilities. On the other hand, intelligent optimization algorithm can continuously improve itself at the same time while optimizing, which can achieve better practical results. In this thesis, the basic principles of five kinds of classical intelligent optimization algorithms are introduced, followed by the introduction of the improvement of the algorithm from the aspects of coding, parameter selection and operator improvement, improvement based on biological behaviour and integrated optimization algorithm. The last part demonstrates the application of optimization algorithm in big data background.

REFERENCES