

Applications of Genetic Algorithm in Software Engineering, Distributed Computing and Machine Learning

Samriti Sharma
Assistant Professor,
Department of Computer Science and Applications
Guru Nanak Dev University, Amritsar

Abstract

There are different types of computational approaches like deterministic, random and evolutionary. Evolutionary techniques are also known as nature inspired techniques as these types of techniques have stolen the idea from nature. Genetic algorithm (GA) is one of the most commonly used evolutionary techniques which is used to solve different NP-hard computational problems. GA is based upon the principle of human genetic. Past research shows that it has been effectively used to solve the different problems from the domain of Computer Science viz. software cost estimation, task scheduling, clustering, natural language processing, query optimization, image processing etc. In this paper, an effort is made to study the use and role of GA in Software Engineering, Distributed Computing, Query Optimization and Machine Learning. .

Keywords: Evolutionary algorithm, Genetic algorithm, Software Engineering, Database.

1. Introduction

In last few years, a significant progress has been found in software and research industry. Initially, all the NP-Hard problems and complex system were developed using deterministic techniques. However, nowadays, evolutionary techniques are most frequently used to solve the complex problems. The evolutionary Algorithms (EA) are basically based on the concept of Evolution by Darwin. In EA, the individuals are represented by fixed length strings called chromosomes. The basic idea behind all the techniques is same as firstly the population of individuals is randomly selected then the function known as fitness function is used to select the best candidates. The new offspring is generated by the application of Crossover and Mutation operators. In other words, evolutionary computation uses iterative progress, such as growth or development in a population. This population is then selected in a guided random search using parallel processing to achieve the desired end. Such processes are often inspired by biological mechanisms of evolution. As evolution can produce highly optimized processes and networks, it has many applications in computer science. Some of the major evolutionary techniques are given below:

- Genetic algorithm
- Honey bee
- Ant colony optimization
- Particle swarm optimization
- Differential evolution
- Harmony search
- Genetic Programming
- Cultural algorithm

2. Genetic Algorithm:

Based upon complexity, a problem can be classified as P, NP-hard and NP-complete. NP hard problem are complex and consume significant amount of CPU time for their execution. In Computer Science, the problems like task scheduling, query optimization, software cost estimation, and data mining all are NP-hard in nature. It is challenging to solve these problems using traditional techniques.

Genetics is a heuristic approach in which assumptions can be made. The execution of the problem can be faster by using genetics approach. The implementation of GA starts with randomly generated

chromosomes. The Fitness function is used to find the fittest ones. The Crossover and Mutation operators are applied to the better candidates to generate new offspring. The algorithm terminates when either the candidate with sufficient quality is produced or the maximum number of generations have been produced. The genetic algorithms are used for solving the problems like optimization problems, search problems etc. The structure and working of genetic algorithm is based upon some concepts like chromosome, fitness function, selection, crossover and mutation. In addition, size of population and number of generation also plays important role.

3. Role of GA in Software Engineering

Software engineering is one of the dominant research areas. In simple words, software engineering is a systematic approach to develop and maintain the automated system. Some of the important aspects of software engineering are software metric, software testing and software quality assurance. Software metrics helps us to measure the performance of the system. The objective of software testing is to reveal the bugs involved in a software module. Additionally, software quality assurance is used to judge and improve the quality of the software.

Different authors have used different techniques to estimate the cost of the software module. Software costs estimation is one of the challenging tasks in software engineering. Isa Maleki¹ et al. [1] have proposed a hybrid solution for estimating the cost of the software. Authors stated that cost estimation is one of the significant factors of software development. Authors first apply genetic algorithm to generate the initial population. After applying crossover and mutation operations, Ant Colony Optimization is used to train the system to compute software cost estimate and evaluate the result.

Authors [3] have developed a binary genetic solution for the same. Authors tested the performance of their algorithm on NASA software and found it more effective as compared to existing software cost estimation models.

Praveen Ranjan Srivastava¹ and Tai-hoon have used genetic algorithm for improving the testing efficiency of the software. Authors stated that using genetic approach; one is able to effectively find most critical path in the software. In software testing, GA outperforms even exhaustive testing for small scale problems.

S. Keshavarz and Reza Javidan have proposed a genetic based solution to control software quality assurance. The objective was to generate optimal test data. Authors have simulated their experiments and prove that genetic approach is best as compared to others.

4. Role of GA in Distributed Computing:

From the last few decades, number of approaches have been developed to expedite the execution process of a job. Initially, the effort was to improve the design of hardware. However, nowadays, emphasizes is given on the software methodology so that the task can be effectively expected. In general, the distributed task can be categorized as distributed application and distributed queries. In distributed application, the focus is to reduce the make span on the job so that the resources of the system can be effectively used. In distributed query, the focus is to reduce the total time or response time of a query. Genetic Algorithm has been extensively used in optimization of distributed tasks. Different authors have used GA to optimize distributed task scheduling. The following part of this section explains the work carried out to optimize makespan in parallel task scheduling and distributed query optimization.

Tzung-Pei Hong, Sheng-Shin Jou, and Pei-Chen Sun[15] have proposed a GA-based approach to solve the scheduling problem with the minimum makespan on identical machines with mold constraints. This GA-based algorithm uses an adjustment operator and they showed that the adjustment operator does increase the performance of the scheduling.

Gurvinder Singh, Kamaljit Kaur, Amit Chhabra[16] stated that the use of multiprocessors have emerged as an effective computing mean for running real-time applications, especially the tasks that a uni-

processor system would not be able to execute. For this purpose, an effective algorithm is required which will optimally determine the time and schedule of the tasks. One can decompose a task into number of sub tasks. Authors represented a cluster as a DAG (Directed Acyclic Graph). In multiprocessor scheduling problem, a given program is to be scheduled in a given multiprocessor system such that the program's execution time should be minimized. They used genetic algorithm approach as it is one of the heuristic approaches which have the high capability to solve the complicated problems like the task scheduling. They developed a new genetic algorithm, named heuristics based genetic algorithm for scheduling static tasks in homogeneous parallel system in which its population size and the number of generations depends on the number of tasks. This algorithm tends to minimize the completion time and increases the throughput of the system. Then they compared the combined approach named as heuristics based genetic algorithm (HGA) based on MET (Minimum execution time)/Min-Min heuristics and b-level or t-level precedence resolution and with a pure genetic algorithm, min-min heuristic, MET heuristic and First Come First Serve (FCFS) approach. They explained that the heuristics based method produces much better results in terms of quality of solutions. Their performance study is based on the best randomly generated schedule of the suggested GA.

Yi-Hsuan Lee, Cheng Chen[21] proposed a modified genetic algorithm to schedule parallel program on multiprocessor system. They also constructed a simulation and evaluation environment to evaluate the execution of task in parallel and find a schedule that minimizes the completion time. They used the heuristic methods to obtain near-optimal solutions. Genetic Algorithms are used to solve this problem. Genetic algorithms are powerful but usually suffer from longer scheduling time. But the proposed algorithm overcomes this drawback. The proposed algorithm integrates the concept of Divide-and-Conquer mechanism to partition the entire problem into subgroups and solve them individually. They showed that the proposed algorithm can not only decrease the scheduling time, but also obtain similar performances as original genetic algorithms, sometimes it is even better. This feature makes the proposed algorithm more scalable and extends its practicability.

Javier Carretero, Fatos Xhafa[24] presented Genetic Algorithms based schedulers for efficiently allocating jobs to resources in a Grid system. They used two encoding schemes and most of GA operators are implemented on them. They proved that the GA-based schedulers are very fast and hence they can be used to dynamically schedule jobs arrived in the grid system by running in batch mode for a short time.

Marin Golub, Suad Kasapovic[26] developed an efficient method based on genetic algorithms to solve the multiprocessor scheduling problem to determine the assignment of tasks to the processors and the execution order of the tasks so that the execution time is minimized. They assumed the fixed number of processors and tasks are represented by a directed acyclic graph (DAG) called task graph.

Distributed Query Optimization: Design of database and query execution plans are two major building blocks of an effective distributed database system. Past research revealed that genetic algorithms are also effectively used in optimizing centralized and distributed queries. Numbers of researcher have used different algorithms viz. Genetic Algorithm, Honey Bee, ACO, PSO etc. and found that by using these techniques one is able to get an optimal query execution plan in seconds as compared to minutes, hours or weeks when deterministic approaches were used. Authors have used different variations of genetic algorithm viz. simple genetic approach, novel genetic approach, stochastic approach, restricted genetic approach and hybrid genetic approach to optimize OLTP and DSS queries. Authors found that hybrid genetic approach gives better results as compared to other variations of genetic algorithm. Navid Khlilzadeh Sourati and Farhad Ramezni have implemented genetic algorithm to reduce the communication or shipping cost of fragments in a distributed database system. Authors found that their approach is beneficial in reducing the shipping costs by effectively allocating the fragments [4][5][6][7][8][15].

Machine Learning

Data Mining is also known as Knowledge data discovery. It is used to reveal or extract some meaningful information from the raw data. It is normally used to handle large or high volume data sets. Data mining is a broad subject that deals with association, prediction, correlation, classification and clustering.

Gunjan Verma and Vineta Verma [8] have discussed that GA is effectively used in pattern recognition, business optimization and stock exchange data mining. Authors [9] used artificial neural network to diagnose one of the killer disease that is lung cancer. Authors found that once the system is trained with using ANN, the performance in terms of accuracy is increased. Behrouz Minaei-Bidgoli, William F. Punch [13] have used genetic algorithm for developing web based educational system. Authors have used GA to predict and classify the students based upon their logged in features. Lean Yu et al. [14] have examined the tendency of stock market using GA based support vector machines. Authors have combined the features of genetic algorithm and support vector machine to explore the stock market policies. GA is used to assist in selection of parameters and to improve the speed of SVM in predicting the exact position of stock market. Authors have empirically test different cases and found their proposed approach as best.

Natural language processing deals with the interaction among computers and natural (human) language, as spoken and written language bodies are being processed for different purposes [11]. Authors stated that evolutionary computing can be effectively used in the different phases of natural language processing viz. machine text summarization, machine translation, part of speech tagging etc. Enrique Alba [12] et al. has tried to solve one of the dominant problems of natural language processing viz. Categorization of Words. Authors have tried and compared different optimization techniques to solve word categorization problem. Authors found the integer coding produces better results as compared to binary coding in GA. Authors found that the accuracy rate of GA is very similar to as given by one of the specific method of word categorization i.e. Viterbi.

Conclusion

In the last few years, the use of genetic algorithm has grown up exponentially. GA has been used in diverse fields or domains viz. Agriculture, Physics, Mechanical Engineering, Chemistry, Astronomy, Computer Science, Medical Science etc. GA assists us to find the solution of NP-hard problems in few seconds. The rate of accuracy of genetic algorithm is good. In this paper, an effort has been made to briefly discuss some of the key research of different authors who have employed genetic algorithm to solve the problem related to Software Engineering, Database and Machine Learning. The review shows that GA has been effectively used in these domains for solving different problems viz. software cost estimation, software testing, quality improvement, query optimization, design optimization of database, data mining, grammar checking etc.

References

1. Isa Maleki, Ali Ghaffari and Mohammad Masdari. 2014. "A New Approach for Software Cost Estimation with Hybrid Genetic Algorithm and Ant Colony Optimization". International Journal of Innovation and Applied Studies. Vol. 5 No. 1.
2. Brajesh Kumar Singh, A. K. Misra. 2012. "Software Effort Estimation by Genetic Algorithm Tuned Parameters of Modified Constructive Cost Model for NASA Software Projects". International Journal of Computer Applications (0975 – 8887) Volume 59– No.9, December 2012.
3. Manik Sharma, Gurvinder Singh, Gurdev Singh, Rajinder Singh. "Analysis of DSS Queries in Distributed Database System Using Exhaustive and Genetic Approach". International Journal of Advanced Computing, Vol.36, Issue.2, April 2013.
4. Manik Sharma, Gurvinder Singh et. Al. 2013. Stochastic Analysis of DSS Queries for a Distributed Database Design. International Journal of Computer Applications (IJCA), Vol. 82, No. 5.
5. Manik Sharma, Gurvinder Singh, Rajinder Singh. "Design and Analysis of Stochastic DSS Query Optimizer in a Distributed Database System". Egyptian Informatics Journal. doi:10.1016/j.eij.2015.10.003
6. Chande Swati V, Sinha Madhvi. 2008. "Genetic algorithm: a versatile optimization tool". BVICAM's Int. J. Inf. Technol. Volume 1, Issue 1.
7. Sevinc Ender, Cosar Ahmat. 2011. "An evolutionary genetic algorithm for optimization of distributed database queries". The Computer Journal. Volume 54, Issue 5.

8. Gunjan verma and Vineta Verma. 2012. "Role and Applications of Genetic Algorithm in Data Mining". International Journal of Computer Applications (0975 – 888) Volume 48– No.17, June 2012
9. N. Ganesan, .K. Venkatesh et al. 2010. "Application of Neural Networks in Diagnosing Cancer Disease Using Demographic Data". International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 26.
10. Lars Bungum, Bjorn Gamback. 2010. "Evolutionary Algorithms in Natural Language Processing". Norwegian Artificial Intelligence Symposium, Gjøvik, 22 November 2010.
11. Enrique Alba a, Gabriel Luque and Lourdes Araujo. 2006. "Natural language tagging with genetic algorithms". Information Processing Letters 100 (2006) 173–182.
12. Praveen Ranjan Srivastava1 and Tai-hoon Kim2. 2009. "Application of Genetic Algorithm in Software Testing". International Journal of Software Engineering and Its Applications Vol. 3, No.4, October 2009.
13. Behrouz Minaei-Bidgoli , William F. Punch. 2003. "Using Genetic Algorithms for Data Mining Optimization in an Educational Web-based System". Genetic and Evolutionary Computation — GECCO 2003
14. Lean Yu1,2, Shouyang Wang1,2,3, and Kin Keung Lai3. 2005. "Mining Stock Market Tendency Using GA-Based Support Vector Machines". LNCS 3828, pp. 336 – 345, 2005.
15. Navid Khlilzadeh Sourati1, Farhad Ramezani. 2015. "Reducing Transfer Costs Of Fragments Allocation In Replicated Distributed Database Using Genetic Algorithms". Advances in Science and Technology Research Journal Volume 9, No. 25, March 2015, pages 1–6 DOI: 10.12913/22998624/1917
16. S. Keshavarz and Reza Javidan. 2011. "Software Quality Control Based on Genetic Algorithm". International Journal of Computer Theory and Engineering, Vol. 3, No. 4, August 2011
17. Tzung-Pei Hong, Sheng-Shin Jou, and Pei-Chen Sun, "Finding the Nearly Optimal Makespan on Identical Machines with Mold Constraints Based on Genetic Algorithms", Proceedings of the 8th WSEAS International Conference on Applied Computer and Applied Computational Science, ISSN: 1790-5117, ISBN: 978-960-474-075-8.
18. Gurvinder Singh, Kamaljit Kaur, Amit Chhabra, "Heuristics Based Genetic Algorithm for Scheduling Static Tasks in Homogeneous Parallel System", International Journal of Computer Science and Security (IJCSS), Volume (4): Issue (2)
19. Yi-Hsuan Lee, Cheng Chen, "A Modified Genetic Algorithm for Task Scheduling in Multiprocessor Systems", Proc. Of 6th International Conference Systems and Applications, 1999.
20. Javier Carretero, Fatos Xhafa, "Genetic Algorithm Based Schedulers for Grid Computing Systems", International Journal of Innovative Computing, Information and Control, ICIC International, Volume 3, Number 6, December 2007.
21. Sharma, Manik, Gurvinder Singh, Rajinder Singh, and Gurdev Singh. "Analysis of DSS Queries using Entropy based Restricted Genetic Algorithm." Appl. Math 9, no. 5 (2015): 2599-2609.
22. Manik Sharma, Gurdev Singh and Harsimran Kaur. "A Study of BNP Parallel Task Scheduling Algorithms Metric's For Distributed Database System". International Journal of Distributed and Parallel Systems (IJDPS) Vol.3, No.1, January 2012
23. Marin Golub, Suad Kasapovic, "Scheduling Multiprocessor Tasks with Genetic Algorithms", OACTA Press, from proceedings (351) Applied Informatics, 2002.
24. Manik Sharma. "Role and Working of Genetic Algorithm in Computer Science". International Journal of Computer Applications and Information Technology (IJCAIT). Volume 2, Issue 1.