PROJECT PROPOSAL

Attribute Subset Selection using Distributed Genetic Algorithm

PROJECT ADVISOR:

Ma’am Sadaf Aslam

PROJECT MANAGER:

Sir Fahad Maqbool

GROUP MEMBERS:

Shahzad Bahir	BCSF16E015
Fatima Shahid Hashmi	BCSF16E011
Nageen Asghar	BCSF16E020

SUBMISSION DATE:

24-Sept-2019

Table of Contents

ABSTRACT:

In this era of big data with facilities for advanced real-time data acquisition, the solutions to large-scale optimization problems are strongly desired. An **optimization problem** is the problem of finding the best solution from all feasible solutions. **Evolutionary computation**, the term now used to describe the field of investigation and the focus of research that concerns evolutionary algorithms, offers practical advantages to the researcher facing difficult optimization problems. **Genetic algorithms** are a subset of evolutionary computing that borrow techniques from classic evolution theories to help find solutions to problems within a large search space. **GA** are efficient optimization algorithms that have been successfully applied to solve a multitude of complex problems. **It** is a method for solving both constrained and unconstrained optimization problems that is based on natural selection, the process that drives biological evolution.

In this project, we will compare our Distributed Genetic Algorithm with its Serial version on Benchmark functions like Ackley, Square Sum and Shubert.

LITERATURE VIEW:

Distributed implementation using Apache Spark of a genetic algorithm applied to test data generation [1]

Main Contribution:

- Parallel implementation of a genetic algorithm in Apache Spark.
- The approach in this research was rather generic, so it had potential to be adapted to other types of GA.
- A fitness function was proposed based on some “probabilities” that certain branch conditions occur in some order and use them to guide the tests towards areas not yet explored. This was different from existing “goal-oriented” approaches, which attempted to find test data for a given path in the control-flow graph.

Drawbacks:

The loading time of the fitness evaluation process took significant time due to operation system dependencies and initialization times.

Reading and Writing data to disk also took significant time, especially in the case of distributed computing when the same machine had several physical processes that could execute parallel tasks but usually shared the same disk.

Spark-Based Parallel Genetic Algorithm for Simulating a Solution of Optimal Deployment of an Underwater Sensor Network [2]

Main Contribution:

- The Shubert multi-peak function (SMPF) is used to simulate deployment of underwater sensor network. By calculating the extremums of the Shubert multi-peak function (ESMPF), the simulating optimal deployment sites can be obtained.
- Based on RDD computation model of the Spark framework, a parallel GA for optimizing the deployment of a UWSN (DUWSN) is designed and implemented.
- By the comparison with the GAs based on single-node and Hadoop, it is verified that the proposed GA runs more efficiently while showing a higher accuracy.

Conclusions:

A Spark-based parallel GA is proposed to simulate the optimal deployment of an underwater sensing network. Compared with single-node-based GA and Hadoop-based GA, the Spark-based parallel GA can significantly shorten the computation time of iterative evolution when dealing with large-scale underwater sensing nodes. Thus, a remarkable improvement in the timeliness of solving the optimal deployment of a UWSN is achieved. Encouragingly, the innate natural randomness and distribution of the Spark-based parallel GA entitles it to avoid local optimization effectively, which further improves the accuracy of the proposed method and finally results in optimal deployment of the UWSN

Scalable Distributed Genetic Algorithm Using Apache Spark (S-GA) [3]

S-GA has been tested on several numerical benchmark problems for large-scale continuous optimization containing up to 3000 dimensions, 3000 population size, and one billion generations. S-GA presents a variant of island model and minimizes the materialization and shuffles in RDDs for minimal and efficient network communication. At the same time it maintains the population diversity by broadcasting the best solutions across partitions after specified Migration Interval. They have tested and compared S-GA with the canonical Sequential Genetic Algorithm (SeqGA).

Conclusion:

S-GA has been found to be more scalable and it can scale up to large dimensional optimization problems while yielding comparable results.

RESEARCH QUESTIONS:

Our research can be evaluated based on the following questions:

I. Is the DGA time efficient?
II. Is the DGA scalable?
III. Will the DGA in fact compete with the serial version?

RESEARCH METHODOLOGY:

Scope:

Genetic algorithms (GAs) are a heuristic search and optimisation technique inspired by natural evolution. They have been successfully applied to a wide range of real-world problems of significant complexity.

Project objectives:

To find different ways of:

I. Computation
II. Implementation
III. Strategies
IV. Optimization Method
V. Best Fit Population

Implementation:

Scala in context of spark while results will be evaluated on cluster. **Scala** helps to dig deep into the **Spark's** source code that aids developers to easily access and implement new features of **Spark**.

I. Scala provides code complexity optimization.
II. Scala offers concise notation.

Benefit:

We can find a different way to solve optimization problems.

Technical Details:

Being population-based characteristics of GA, it would solve evolutionary problems.

REFERENCES:

[1] Paduraru, Ciprian & Melemciuc, Marius-Constantin & Stefanescu, Alin. (2017). A distributed implementation using apache spark of a genetic algorithm applied to test data generation. 1857-1863. 10.1145/3067695.3084219.

[2] Liu, P., Ye, S., Wang, C., & Zhu, Z. (2019). Spark-Based Parallel Genetic Algorithm for Simulating a Solution of Optimal Deployment of an Underwater Sensor Network. *Sensors*, *19*(12), 2717.

[3] Maqbool F., Razzaq S., Lehmann J., Jabeen H. (2019) Scalable Distributed Genetic Algorithm Using Apache Spark (S-GA). In: Huang DS., Bevilacqua V., Premaratne P. (eds) Intelligent Computing Theories and Application. ICIC 2019. Lecture Notes in Computer Science, vol 11643. Springer, Cham