Comparison among Data Mining Genetic Algorithm, Evolutionary Operational Algorithm and Newton's Method

Satish Kumar Malik

Research Scholar, Singhania University, Rajasthan, INDIA

----------♦----------

1. INTRODUCTION

A genetic algorithm (GA) is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover.

In a genetic algorithm, a population of strings (called chromosomes or the genotype of the genome), which encode candidate solutions (called individuals, creatures, or phenotypes) to an optimization problem, evolves toward better solutions. Traditionally, solutions are represented in binary as strings of 0s and 1s, but other encodings are also possible. The evolution usually starts from a population of randomly generated individuals and happens in generations. In each generation, the fitness of every individual in the population is evaluated, multiple individuals are stochastically selected from the current population (based on their fitness), and modified (recombined and possibly randomly mutated) to form a new population. The new population is then used in the next iteration of the algorithm. Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population. If the algorithm has terminated due to a maximum number of generations, a satisfactory solution may or may not have been reached.

2. REQUIREMENT OF SUCCESSFUL GENETIC ALGORITHM

Genetic algorithms find application in bioinformatics, phylogenetics, computational science, engineering, economics, chemistry, manufacturing, mathematics, physics and other fields.

A typical genetic algorithm requires:

1. a genetic representation of the solution domain,

2. a fitness function to evaluate the solution domain.

A standard representation of the solution is as an array of bits. Arrays of other types and structures can be used in essentially the same way. The main property that makes these genetic representations convenient is that their parts are easily aligned due to their fixed size, which facilitates simple crossover operations. Variable length representations may also be used, but crossover implementation is more complex in this case. Tree-like representations are explored in genetic programming and graph-form representations are explored in evolutionary programming.

3. ROLE OF GENETIC ALGORITHM AND KDD IN RESEARCH WORK

The fitness function is defined over the genetic representation and measures the quality of the represented solution. The fitness function is always problem dependent. For instance, in the knapsack problem one wants to maximize the total value of objects that can be put in a knapsack of some fixed capacity. A representation of a solution might be an array of bits, where each bit represents a different object, and the value of the bit (0 or 1) represents whether or not the object is in the knapsack. Not every such representation is valid, as the size of objects may exceed the capacity of the knapsack. The fitness of the solution is the sum of values of all objects in the knapsack if the representation is valid, or 0 otherwise. In some problems, it is hard or

even impossible to define the fitness expression; in these cases, interactive genetic algorithms are used.

Once we have the genetic representation and the fitness function defined, GA proceeds to initialize a population of solutions randomly, then improve it through repetitive application of mutation, crossover, inversion and selection operators.

Capable and well-organized data mining algorithms are essential and fundamental to helpful, useful, and successful knowledge discovery in databases. We discuss several data mining algorithms including genetic algorithms (GAs). In addition, we propose a modified multivariate Newton's method (NM) approach to data mining of technical data. Several strategies are employed to stabilize Newton's method to pathological function behavior. NM is compared to GAs and to the simplex evolutionary operation algorithm (EVOP). We find that GAs, NM, and EVOP all perform efficiently for well-behaved global optimization functions with NM providing an exponential improvement in convergence rate. For local optimization problems, we find that GAs and EVOP do not provide the desired convergence rate, accuracy, or precision compared to NM for technical data. We find that GAs are favored for their simplicity while NM would be favored for its performance.

Data mining and Knowledge Discovery in Databases have become commercially important techniques and active areas of research in recent years. Business applications of data mining software are commonplace and are commodities in many cases. However, data mining of technical data is still a relatively disorganized discipline compared to business applications of data mining. For example, the application of neural networks trained by genetic algorithms to a business' market basket analysis procedures would not be unusual. The use of informatics, a field that is similar to On-line Analytical Processing (OLAP), in biology and chemistry is increasing, however. There is an increasing need for data mining algorithms with scientific precision.

4. RESEARCH SUMMARY

In this work, we survey the algorithms of data mining and propose several new algorithms for data mining. Specifically, we show how Newton's method, especially local Newton's method, could be applied to data mining applications for technical data – the method may also find uses in specialized business applications as well, *i.e.*, non-marketing applications. We also discuss genetic algorithms (GA), the fixed simplex evolutionary operation (EVOP), and the variable length simplex EVOP. GA and EVOP are evolutionary algorithms. GAs use a stochastic process and EVOPs use a deterministic process.

In the next chapter, a literature survey of data mining is given. In the following chapters, we develop an algorithm based on Newton's method as a data mining algorithm for applications involving technical data. Chapter 3 (Newton's Method) will be a literature survey of Newton's method (NM), explains quadratic convergence, gives the NM convergence criteria, and illustrates convergence criteria with examples from chaos theory. Chapter 4 (Modeling and Newton's Method) will explains how Newton's method fits into modeling theory and describes the local Newton's method, global Newton's method, non-linear regression, and robust non-linear regression.

Chapter 5 (Matrix Algebra) will explains the methods necessary to implement Newton's method for higher dimensional problems. The derivation of NM from the method of maximum likelihood estimation is given. And, the variance-covariance matrix for NM is derived such that statistical analysis of NM results can be obtained. Chapter 6 (Results and Discussion) gives the comparison of using Newton's method, the simplex EVOP methods, and genetic algorithms on some model problems in terms of precision, accuracy, and convergence rate. Chapter 7 (Comparison of Algorithms) will compares these algorithms in terms of computational steps required, the storage space required, and the complexity of the algorithms.

5. TENTATIVE RESEARCH CONTENT

The research will contain these chapter (approximately)

- INTRODUCTION
- DATA MINING LITERATURE SURVEY
- NEWTON'S METHOD
- MODELING AND NEWTON'S METHOD
- MATRIX ALGEBRA
- RESULTS AND DISCUSSION
- COMPARISON OF ALGORITHMS
- CONCLUSION
- REFERENCE LIST

6. REFERENCES

Addison, Paul S. 1997. Fractals and Chaos. Philadelphia: Institute of Physics Publishing.

Baase Sara and van Gelder, Allen. 2000. Computer Algorithms: Introduction to Design and Analysis. 3rd ed. New York: Addison-Wesley.

Comaford, Christine. 1997. Unearthing data mining methods, myths. PC Week 14, no. 1 (January 6): 65.

Dunham, William. 1994. The Mathematical Universe. New York: John Wiley & Sons, Inc.

Elert, Glenn, ed. 2000. The Physics Factbook. http://www.hypertextbook.com/facts/2000.

Harris, J. W. and Stocker, Horst. 1998. Handbook of Mathematics and Computational Science. New York: Springer-Verlag.

Lesk, Michael. 1997. How Much Information is There in the World?

http://www.lesk.com/mlesk/ksg97/ksg.html.

Smith, K. A and Gupta, J. N. D. 2000. Neural Networks in Business: Techniques and Applications for the Operations Researcher. Computers & Operations Research