Role and Applications of Genetic Algorithm in Data Mining

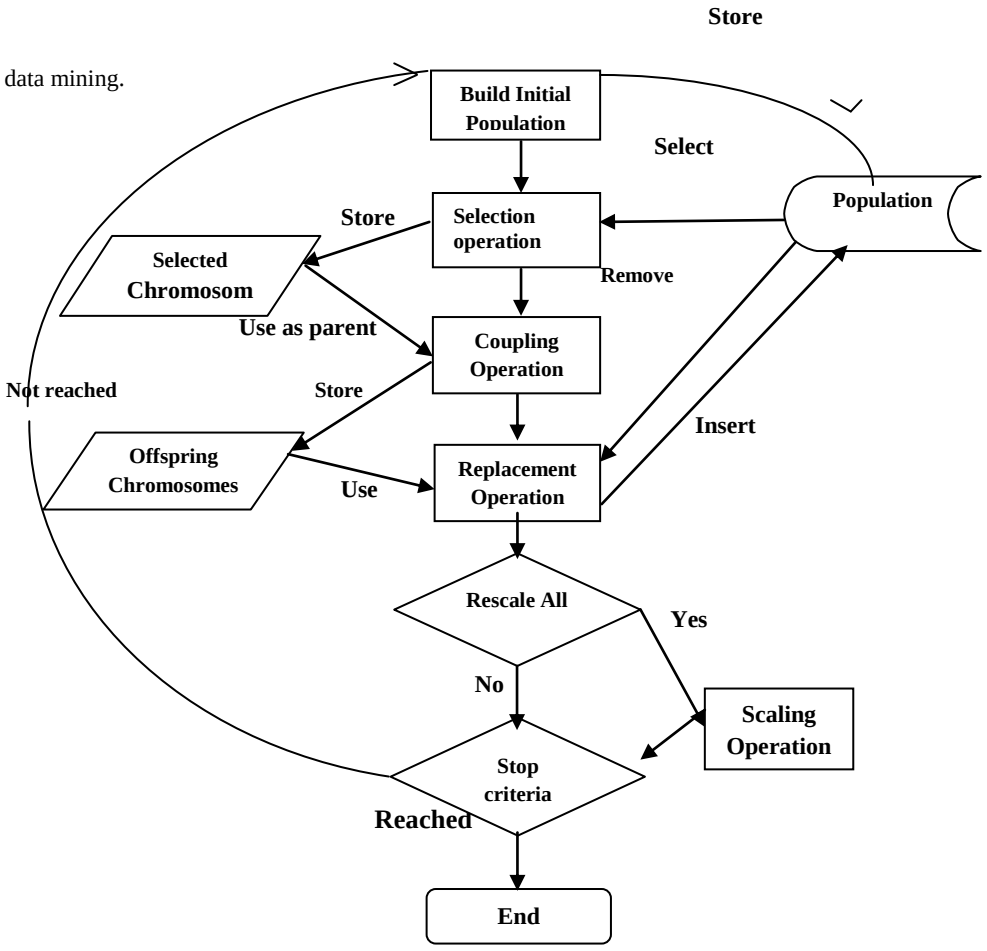
Gunjan Verma
(Assistant Professor)
Meerut Institute of Engineering & Technology
Meerut

Vineeta Verma
(Assistant Professor)
Sardar Vallabhbhai Patel University
of Agriculture & Technology
Meerut

ABSTRACT
Data mining has as goal to extract knowledge from large databases. To extract this knowledge, a database may be considered as a large search space, and a mining algorithm as a search strategy. In general, a search space consists of an enormous number of elements, making an exhaustive search infeasible. Therefore, efficient search strategies are of vital importance. Search strategies based on genetic-based algorithms have been applied successfully in a wide range of applications. In this paper, we discuss the suitability of genetic-based algorithms for data mining. We discuss the various application areas where genetic Algorithm plays evolutionary role with data mining technique and explain them in details.

Keywords
GA, Classifier, data mining.

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



Flowchart of a genetic algorithm

2. INTRODUCTION OF DATA MINING

A field that deals with extracting knowledge from databases, without putting restrictions on the amount or types of data in a database, is data mining [8]. Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

2.1 Data

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:

- operational or transactional data such as, sales, cost, inventory, payroll, and accounting
- nonoperational data, such as industry sales, forecast data, and macro economic data
- meta data - data about the data itself, such as logical database design or data dictionary definitions

2.2 Information

The patterns, associations, or relationships among all this data can provide information. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

2.3 Knowledge

Information can be converted into knowledge about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

2.4 Data Warehouses

Dramatic advances in data capture, processing power, data transmission, and storage capabilities are enabling organizations to integrate their various databases into data warehouses. Data warehousing is defined as a process of centralized data management and retrieval. Data warehousing, like data mining, is a relatively new term although the concept itself has been around for years. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis. Dramatic technological advances are making this vision a reality for many companies. And, equally dramatic advances in data analysis software are allowing users to access this data freely. The data analysis software is what supports data mining.

3. HOW DOES DATA MINING WORK?

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

- **Classes**: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
- **Clusters**: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
- **Associations**: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.
- **Sequential patterns**: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

3.1 Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

4. USE OF GENETIC ALGORITHM IN DATA MINING

In this paper, we discuss the applicability of a genetic-based algorithm to the search process in data mining. Data mining algorithms require a technique that partitions the domain values of an attribute in a limited set of ranges, simply because considering all possible ranges of domain values is infeasible.

5. APPLICATION OF GENETIC ALGORITHM

(a) Genetic Algorithms is an effective tool to use in data mining and pattern recognition.

There are two different methods to applying
GA in pattern recognition:

1 Use GA as a classifier directly in computation.

2. Use a GA to optimize the results i.e. as an optimizer to arrange the parameters in other classifiers. Most applications of GAs in pattern recognition optimize some parameters in the classification process [4].

GAs has been applied to find an optimal set of feature weights that improve classification accuracy. First, a traditional feature extraction method such as Principal Component Analysis (PCA) is applied, and then a classifier such as **k-NN** (**Nearest Neighbor Algorithm**) is used to calculate the fitness function for GA [10], [6]. Combination of classifiers is another area that GAs have been used to optimize. GA is also used in selecting the prototypes in the case-based classification

According to us second method of genetic algorithm to optimize the result from the dataset is more effective to compute the accurate values of observations of data by applying data mining techniques.

(b) Genetic Algorithm has a wide scope in business.

There are large amount of data that has to be filtered to process the results for optimizing the business profits by using various data mining techniques. There are many domains in business to which they can be applied:

I. Optimization

Give a business problem with certain variables and a well defined definition of profit, a genetic algorithm can be used to automatically determine the optimal value for the variables that optimize the profit [1].

II. Prediction

Genetic algorithms have been used as meta level operators that are used to help optimize other data mining algorithms. For instance, they have been used to find the optimal association rules in market-analysis

III. Simulation

Sometimes a specific business problem is not well defined in terms what the profit is or whether one solution is better than the other. The business person instead just has large number of entities that they would like to simulate via simple interaction rules overtime.

(c) Genetic Algorithm in stock exchange data mining.

Stock market and other finance fields, Genetic Algorithm has been applied in many problems [12]. There have been a number of attempts to use GA for acquiring technical trading rules.

One application is how to find the best combination values of each parameter. We know that in a trading rule there are many parameters, when we try to find the most profit, we must test the parameter combination one by one, which is called greedy algorithm which costs a lot of running time and memory**.**

6. CONCLUSION

The first and most important point is that genetic algorithms are intrinsically parallel. Most other algorithms are serial and can only explore the solution space to a problem in one direction at a time since GAs have multiple offspring, they can explore the solution space in multiple directions at once. If one path turns out to be a dead end, they can easily eliminate it and continue work on more promising avenues, giving them a greater chance each run of finding the optimal solution. Genetic algorithms provide a comprehensive search methodology for machine learning and optimization. It has been shown to be efficient and powerful through many data mining applications that use optimization and classification. GAs can rapidly locate good solutions, in data mining even for difficult search spaces. GAs are used in various fields of Data mining to get the optimized solutions for the better performance of the data that are required in decision making and process the accurate result. There is also a greater scope of GA in data mining in future application to stimulate the data mining concepts. Genetic algorithms are widely applicable to classification by means of inductive learning. GAs also provides a practical method for optimization of data preparation and data transformation steps. Hence GA can be used in a real analysis system to get the better solution.

7. ACKNOWLEDGMENTS

Our thanks to the experts and authors of referenced journals who have contributed towards development of the paper and help us for making the concepts clear.

8. REFERENCES

[1] Forrest, Stephanie. "Genetic algorithms: principles of natural selection applied to computation." *Science*, vol.261, p.872-878 (1993).

[2] Bandyopadhyay, S., and Muthy, C.A. "Pattern Classification Using Genetic Algorithms".

[3] Jain, A. K.; Zongker, D. "Feature Selection: Evaluation, Application, and Small Sample Performance", IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 19, No. 2, February (1997).

[4] Pattern Recognition Letters, (1995).Vol. 16, pp. 801-808.

[5] Erick Cantu-Paz, "Feature Subset Selection, Class Separability, and Genetic Algorithms", Center for Applied Scientic Computing Lawrence Livermore National Laboratory Livermore, CA, (1994).

[6] Siedlecki, W., Sklansky J., A note on genetic algorithms for large-scale feature selection, Pattern Recognition Letters, Vol. 10, Page 335-347, (1989).

[7] Agrawal, R., and R. Srikant. 1994. Fast algorithms for mining association rules. In Proceedings of the 20th international conference on very large databases held in Santiago, Chile, September 12-15, 1994, 487-99. San Francisco, CA: Morgan Kaufmann.

[8] M. M itchell (1997) an introduction to genetic Algorithm MIT press MA.

[9] Agrawal, R., and R. Srikant. 1994. Fast algorithms for mining association rules. In Proceedings of the 20th international conference on very large databases held in Santiago, Chile, September 12-15, 1994, 487-99. San Francisco, CA: Morgan Kaufmann.

[10] Pei, M., Punch, W.F., and Goodman, E.D. "Feature Extraction Using Genetic Algorithms", Proceeding of

International Symposium on Intelligent Data Engineering and Learning'98 (IDEAL'98), Hong Kong, Oct. (1998).

[11] Wright, A.W. "Genetic Algorithms for real-parameter optimization", In Rawlings, R.E. (ed) Foundations of Genetic Algorithms, 1990, pp 205 -220, Morgan Kaufman.

[12] Chen, S. H. Genetic Algorithms and Genetic Programming in Computational Finance. Boston, MA: Kluwer. 2002.