Data Mining using Genetic Algorithm

Media Analytics using TAM Database and Real-Time Stock Market Values

Sameer Save, Sayali Jojan, Smit Shah

Mrs. Madhavi Gangurde

Department of Information Technology, Vidyavardhini's College of Engineering and Technology, Vasai Road, Thane 401202, Maharashtra, India.
E-Mail ID: sameer_save@hotmail.com.

ABSTRACT:
Data-Mining is often associated with the KDD process, which involves sub-processes from data cleaning and integration to pattern evaluation and Knowledge representation along with Data Mining. Many algorithms have been developed in the past just to suit the needs of an organization for mining patterns in their search space. Also the more generalized and common algorithms like APriori have been extensively used. Such algorithms have given satisfactory results when applied in a static environment and a finite search space.
In this paper we make an effort to implement mining using the Genetic Algorithm's Evolutionary approach.

KEYWORDS: Data mining, KDD, GENETIC ALGORITHM.

LITERATURE SURVEY:
Adaptive systems, in computing, comprise numerous techniques that, when applied to a problem, autonomously adjust the relative importance of input parameters to solve that problem. In other words, the goal of all adaptive systems is machine learning through identifying the patterns and relationships between the input parameters and the desired outputs. Although these systems can be 'tuned' to apply to specific problem domains, they do not have any one specific goal in their conclusion. Rather, they adopt a 'best fit' strategy to problem solving. This 'best fit' strategy results in systems that are well suited to non-linear and NP complete problems. While the field of adaptive systems research is often referred to as artificial intelligence, there are two very distinct approaches to the problem of machine learning, of which artificial intelligence is one. The two main streams of adaptive systems are artificial intelligence and evolutionary systems.

Genetic Algorithm based on Evolutionary Technique:
Based on the concept of biological evolution, the evolutionary approach has many advantages over the traditional approach of the Genetic Algorithm.
The evolutionary methodology allows dynamic allocation of selection parameters. Also the select operation is carried out in a particular way in evolutionary approach, wherein a certain portion of the population is carried forward into the next generation. Mutation of genetic algorithm is carried out in a subsidiary location. Genetic algorithm which is based on evolution Strategy, can not only conquer getting in state of local convergence, but also accelerate search speed.

Forecasting using GA:
Genetic Algorithms have had varied application domains over the years. Few of them being finance, media, etc.
Buying advertisement spots on a carrier (channel) is expensive. A crucial strategic mistake can lead to a huge financial disaster. Hence it is important to invest in the right carriers (channels). Not to mention the time lost in the mean time. An organization which wishes to invest for spots on a carrier needs to consider the factors having the potential to influence the sales of their product. For example: A product targeted for the age group 15-21 say PSP or XBOX should be advertised on the carriers viewed by the age group. It would be a waste of time and money to advertise the same on a 24 hr NEWS carrier.
In this paper we make an attempt to implement the same in media analytics. As per the given example, we are considering a search space comprising of cosmetics,

electronics, 2 wheelers and 4 wheelers. Using the current available scores for the brands and their respective spots, we shall try to analyze the best spot for a particular brand.
The result of this shall provide us with a few favorable spots to host the advertisement according to the budget and the target group.

Plotting against the REAL-TIME stock-markets:
The database used for study is the TAM (Total Addressable Market) database. Once the analysis based on the 'historical data' from the TAM database is complete, we plot the results across the real-time market values. This guarantees that the results are consistent with the current market trends. Plotting against real-time markets removes the obsolete / irrelevant results.

.

INTRODUCTION:

Television media is the most common approach most of the firms opt for reaching customers. It is important that they select the correct spot for airing their advertisement so as to reach the maximum number of people.
In this project we have made attempt to provide consultancy to firms interested to advertise their products on television media. Working with the historical data from the TAM (Total Addressable Market) database and the market then, we try to predict the market for the present day / investment date.

PROBLEM STATEMENT:

To provide consultancy to firms for investment in television media for their product based on their budget and related target group.

APPROACH:

We plan to approach the above mentioned problem statement using Genetic Algorithm's Evolutionary Approach. This process mimics the evolution process for the considered data set.

The process consists of the following steps:
Population Creation: Generates the population which takes part in the process.

Selection: Individuals are selected on basis of their fitness value which is calculated using the fitness function.

Crossover**:** In this step the selected individuals for mating undergo crossover, wherein a part of each individuals' string is swapped.

Mutation: Changing of certain bits in the string of the individual.

Election Tournament: Selection of the new individuals formed for populating the next generation.

Elitism: Property of preservation of individuals with unique characteristics.

Convergence: Condition which marks satisfaction of factors.

The final result is delivered once the convergence criterion is satisfied.
The output of the process is best 3 investment options based on the historical data. Since the market conditions change drastically on a frequency of hours, it is necessary to compare the results with the current market conditions, thus ensuring that the generated results are valid.

Fitness:

Fitness is probably the main concept in Darwinian evolution; it is the 'direction' the genetic algorithm takes in its pursuit of improvement. Fitness refers to an individual's ability to compete within an environment for available resources. Goldberg describes the fitness function as "some measure of profit, utility, or goodness that we want to maximize**"**.
In the genetic algorithm, this competition is based on the chromosome's performance within the problem domain. Some nominal scale is determined that is suitable to the task like 'time to failure', or 'time to balance'. After a chromosome is applied to the problem, it is awarded a fitness value to reflect its performance. In this way, when the entire population has been tested, the relative ability of each chromosome can be identified.

Tournament selection approach can be likened to the natural process of individuals competing with each other in order to mate. The chromosomes in the population are selected for competition, usually randomly, with the victor increasing its expected incidence in the mating pool.
Tournament selection should not be confused with tournament fitness. Initially the entire population is in the tournament. Two members are selected at random to compete against each other with only the winner of the competition progressing to the next level of the tournament". When the tournament is over (i.e. only one individual is left), the relative fitness of each member of the population is awarded according to the level of the tournament it has reached.
The actual tournament most commonly used in tournament selection is the same as that described above for tournament fitness. This technique selects unique individuals based on their relative fitness. The final victor of the competition will always be the fittest

chromosome in the population. The more successful the chromosome is in competition, the more often that chromosome is expected to appear in the mating pool.

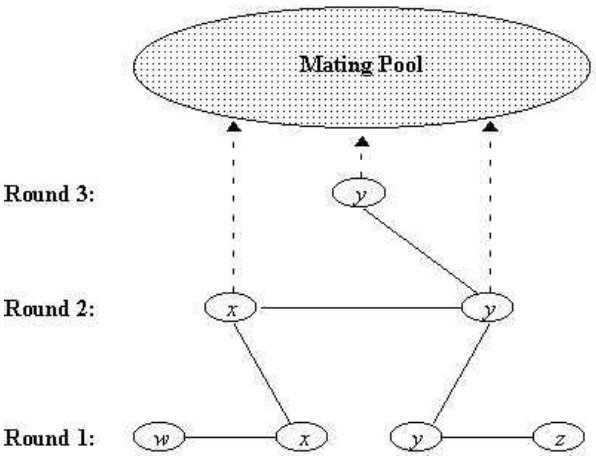

Figure 1: Tournament Competition between Chromosomes

A situation that can be expected in tournament selection is the pairing of two chromosomes that display relatively equal fitness. Competition between well matched, fit individuals in the later stages of the tournament is beneficial to the selection process. In the early stages of the tournament this 'equal' pairing can be detrimental. The most extreme case would occur if the two fittest chromosomes were paired in the first round of the tournament if the two chromosomes are identical, the resulting competition will avoid convergence around this chromosome. If however, the two chromosomes represent distinctly different solutions, their early pairing will result in the less fit competitor (the loser) being awarded no representation in the mating pool: the extinction of a species.

Fitness Proportionate Reproduction:
Reproduction in genetic programming is asexual, thus imitating the process of budding in biology. Through reproduction, an identical copy of the individual selected is carried over into the next generation: survival of the fittest. Fitness proportionate reproduction is the asexual reproduction of chromosomes selected stochastically from the population. "The operation of fitness proportionate reproduction for the genetic programming paradigm is the basic engine of Darwinian reproduction and survival of the fittest". In other words, the selection of individual and chromosome is based on a probability that is relative to that chromosomes relative fitness within its population.

Crossover Techniques:
Many crossover techniques exist for organisms which use different data structures to store themselves.
A single crossover point on both parents' organism strings is selected. All data beyond that point in either organism string is swapped between the two parent organisms. The resulting organisms are the children:

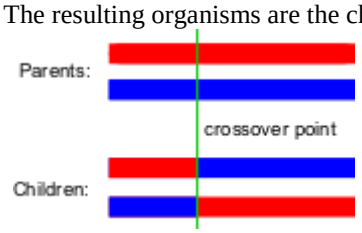

Figure 2: Single point crossover

Two-point crossover calls for two points to be selected on the parent organism strings. Everything between the two points is swapped between the parent organisms, rendering two child organisms:

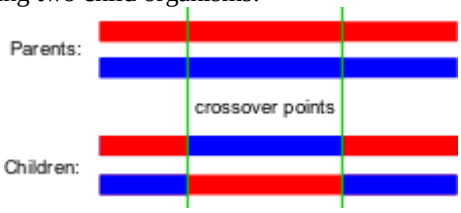

Figure 3: Two point crossover

Mutation:
Mutation in genetic programming is essentially the same as that occurring in the genetic algorithm (figure 3). It is a random alteration of a gene, or genes, in a chromosome. The primary difference between mutation in genetic algorithms and in genetic programming is the mutant gene's type. In the genetic algorithm, mutation involves selecting a random point, or points, in the chromosome and substituting the value found at that point with an appropriate value that has been randomly generated.

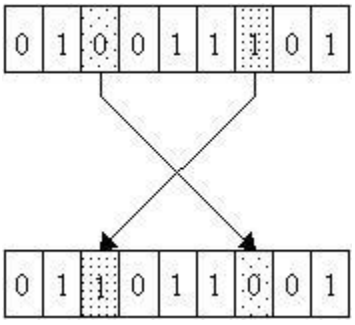

Figure 4: Mutation in Binary Strings

In genetic programming, any atom can be replaced by any other atom. Any terminal can be replaced by either another terminal or a sub function (i.e., a sub-tree of atoms and terminals). Unlike mutation in the genetic algorithm, mutation in genetic programming must consider the genotype, the range of appropriate values for that genotype, and the number of arguments the new function receives if the function is to remain valid.

CONCLUSION:

The proposed application is expected to run on system that is registered to access the TAM database files. The application would act as a Decision Support System for investors, thus giving them a clear idea and statistical justification for their investment.

The application processes the historical data and cross-references it with its current values thus giving predicting the trend and removing the irrelevant results with help of current market values.

ACKNOWLDGEMENT:

The authors of this paper would like to acknowledge, the help of guide Mrs, Madhavi Gangurde, and every person of the Information Technology Department, for their assistance.

Peter Armand Menon, General Manager-IT, PCI for his guidance, Kedar Nachne Associate Director –Media, Spatial Access for his assistance understanding the advertisement media industry.

REFERENCES:

[1]. http://www.stumptown.com/diss/chapter2.html
[2]. Adaptation in Natural and Artificial Systems - John Holland
[3]. http://www.cse.unr.edu/~banerjee/selection.htm
[4]. [IEEE Paper from Journal of Economic Dynamics and Control] Genetic Algorithm Learning and COBWEB model.
[5]. The Applications of Genetic Algorithms in Stock Market Data Mining Optimization [Li Lin, Longbing Cao, Jiaqi Wang, Chengqi Zhang][Faculty of Information Technology, University of Technology, Sydney, NSW 2007, Australia]
[6]. Application of Genetic Algorithm in Data Mining [Tan Jun-shan, He Wei, Qing Yan]