Genetic algorithms are mathematical procedures utilizing the process of genetic inheritance. They have been usefully applied to a wide variety of analytic problems. Data mining can combine human understanding with automatic analysis of data to detect patterns or key relationships. Given a large database defined over a number of variables, the goal is to efficiently find the most interesting patterns in the database. Genetic algorithms have been applied to identify interesting patterns in some applications. They usually are used in data mining to improve the performance of other algorithms, one example being decision tree algorithms, another association rules.

Genetic algorithms require certain data structure. They operate on a population with characteristics expressed in categorical form. The analogy with genetics is that the population (genes) consist of characteristics (alleles). One way to implement genetic algorithms is to apply operators (reproduction, crossover, selection) with the feature of mutation to enhance generation of potentially better combinations. The genetic algorithm process is thus:

---

1. Randomly select parents.
2. Reproduce through crossover. Reproduction is the operator choosing which individual entities will survive. In other words, some objective function or selection characteristic is needed to determine survival. Crossover relates to changes in future generations of entities.
3. Select survivors for the next generation through a fitness function.
4. Mutation is the operation by which randomly selected attributes of randomly selected entities in subsequent operations are changed.
5. Iterate until either a given fitness level is attained, or the preset number of iterations is reached.

Genetic algorithm parameters include population size, crossover rate (the probability that individuals will crossover), and the mutation rate (the probability that a certain entity mutates).

Genetic Algorithm Advantages: Genetic algorithms are very easy to develop and to validate, which makes them highly attractive if they apply. The algorithm is parallel, meaning that it can applied to large populations efficiently. The algorithm is also efficient in that if it begins with a poor original solution, it can rapidly progress to good solutions. Use of mutation makes the method capable of identifying global optima even in very nonlinear problem domains. The method does not require knowledge about the distribution of the data.

Genetic Algorithm Disadvantages: Genetic algorithms require mapping data sets to a form where attributes have discrete values for the genetic algorithm to work with. This is usually possible, but can lose a great deal of detail information when dealing with continuous variables. Coding the data into categorical form can unintentionally lead to biases in the data.

There are also limits to the size of data set that can be analyzed with genetic algorithms. For very large data sets, sampling will be necessary, which leads to different results across different runs over the same data set.

Demonstration of Genetic Algorithm

The value of genetic algorithms is in their ability to deal with complex sets of data, where there are an unreasonably large number of variable combinations. One of the commonly applied applications of data mining is to loan applications. We use a loan application dataset with representative observations given in the Appendix. The purpose of the genetic algorithm in this case is simply to identify a set of loan applicant characteristics with the optimal fitness function. The example is very small, so you will know