Using Data Mining for Life Insurance Network Optimization

Brunela Karamani[1]

[1]**Polytechnic University of Tirana,
Computer Engineering Department,
Tirana, Albania**

Esteriana Haskasa[2]

[2] **"Aleksander Xhuvani" University,
Department of Mathematics,
Elbasan, Albania**

Shkelqim Kuka[3]

[3]**Polytechnic University of Tirana,
Department of Mathematics,
Tirana, Albania**

Abstract

Managing the large amounts of information and efficiently using this information in improved decision making has become increasingly challenging. In this paper we will demonstrate how genetic algorithms integrated in MATLAB help in optimizing a life insurance network. We will analyze the case of study of a life insurance company which wants to enlarge its network. We will explain how we have calculated the fitness function and how we have implemented the data mining optimizing technique of cluster analysis. We will use MATLAB to deliver the results of our study, accompanied them with appropriate explanations. We will conclude arguing the benefits of our study.

Keywords: *optimizing, network, data mining, cluster analysis, genetic algorithm, fitness function, insurance.*

1. Introduction

Nowadays businesses are gathering terabytes of data. Managing the large amounts of information and efficiently using this information in improved decision making has become increasingly challenging. In the last decade business intelligence and especially artificial intelligence are becoming very popular in the economy of the world most developed countries. This is due to significant precise solutions that intelligence methods give to complex economic problems. In the last two decades Genetic Algorithms have increasingly been applied in economics to find an appropriate solution for optimization problems. Current mathematical optimization model do not offer the same precision solution and performance for complex problems as Genetic Algorithms (GA) do. GA finds a solution through evolution that in contrast of the natural life always evolves towards a good solution. They can quickly scan vast solution set. [1]

Our study had as objective the use of an appropriate technical data mining for network optimization of a life insurance company. By analyzing the results of different research that have been done on the relationship GA optimization and analysis of the following reasons, the GA was data mining techniques that we needed for our study.

GA is parallel. The most part of other methods are serial, and explore the solution to a problem in one direction at a time. In cases when the solution discovered by these methods is suboptimal then the work will be abandoned and it should start over. GA explores the solution space in multiple directions at once. In cases when 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.[2]

GA is particularly well-suited to solving problems where space of all potential solutions is truly huge-too vast to search exhaustively in any reasonable amount of time. So they implicitly evaluate many schemas at once.

GA perform well in problems for which the fitness landscape is complex – ones where the fitness function is discontinuous, noisy, changes over time, or has many local optima. [3]

GA has the ability to manipulate many parameters simultaneously. Many real world problems cannot be stated in terms of a single value to be minimized or maximized, but they must be expressed in terms of multiple objectives, usually with tradeoffs involved. GA is very good in giving an optimal solution to such problems.

Finally GA know nothing about the problems they are deployed to solve, but they for sure give an optimize solution. [4][5][6]

Although this kind of technique is widely used in a lot of developed countries in the world, in Albania the situation is different. Mathematical methods are still the most used and preferred. This because of the below factors:

- ⇨ Company should consider the high cost before implementing the needed tools and framework in order to implement this technique. Since the most part of companies in Albania are small or medium ones, they don’t have a powerful economic potential to support the costs of integration of GA in their business.

- Another important factor is the difficulties of finding qualified people that can work with GA and to improve them.

2. Genetic Algorithms model

The method starts with a population of objects, or individuals, each having a known fitness. GA takes in consideration a population of individuals. Then on these individuals are performed genetic operations such as crossover (sexual recombination) and mutation. The genetic operation use Darwinian principle of survival and reproduction. Genetic Algorithm solves a problem through driving the population’s evolution in the direction we want it to go. It does this by using fitness function to give a probability for which individuals to choose from. Some basic concepts of GA are:

- *Chromosome:* is a string composed by a set of bits that represent the solution space.
- *Population:* is the set of chromosomes randomly selected.
- *Fitness:* is the performance evaluation of individuals in the population.

The most frequently employed operators in GA are:

- *Reproduction:* is the process in which individuals copy themselves, according to the probabilities that are proportional to their fitness values.
- *Crossover*: is the operator that produces two new chromosomes by exchanging some bits of a couple of randomly selected chromosomes (parents).
- *Mutation:* operates on a single chromosome with a very small probability.

In application, GA used encoding scheme, like DNA in nature, in order to represents a probable solution, typically a vector or a simply just binary code. The individuals that have these DNA records will be used as criteria of judging the individuals fitness or how close they are to the solution. A number of strings containing information about how to behave in their environment and some operators that change the strings are the basic element of a GA. After behaving, the strings are evaluated by a fitness function, representing their environment. The strings that are adapted better are them whose get the higher scores. These strings have a better probability to be chosen by the selection operator that determines which strings are allowed to reproduce. The chosen strings then undergo a procedure of crossing/over and mutation and the so built “offspring” forms next period’s generation that undergoes the same operations. [1][2][7][8] The below prototype gives a better overview of the logic used by GA:

```
AGJ(Fitness, Fitness_threshold, p, r, m)
{Fitness:A function that assigns an Evaluation score,given a hipothesis
Fitness_threshold: A threshold specifying the termination criterion.
p: the number of hypotheses to be included the population
r: the fraction of the population to be replaced by Crossover at each step
m: the mutation rate
-    Initialize population:P ← Generate p hypotheses at random
-    Evaluate:For each h in P, compute Fitness(h)
-    While[$max_h$Fitness(h)<Fitness_threshold] do
{  Create a new generation $P_s$:
1.Select: Probabilistically select (1-r)p members of P to dd to $P_s$. The probability Pr($h_i$) of selecting hypothesis $h_i$ from P is given by:
   $\Pr(h_i) = \frac{fitness(h_i)}{\sum_{j=1}^{p} fitness(h_i)}$
 2.Crossover:Probabilisticaly select rp/2 pairs of hypotheses from P, according to Pr($h_i$) given above. For each pair ($h_i$\$h_2$) produce two offspring by applying the Crossover operator. Add all offspring to $P_s$.
3.Mutate:Choose m percent of the members of $P_s$, with uniform propality.For each, invert one randomly selectedbit in its presentation.
4.Update P ←$P_s$
5.Evaluate: for each h in P, compute Fitness(h)
}
Return the hypothesis from P that has the highest fitness
}
```

Figure 1 The prototype of GA

3. Case Study

GA Tool is an important and very often used tool in MATLAB. MATLAB provides high performance language for technical computing. It integrates computation, visualization and programming in an easy-to-use environment where problems and solutions are expressed in a familiar mathematical notation.

In our case of study we have demonstrated how genetic algorithms integrated in MATLAB help in finding the minimum distance between some given objects coordinates. We have taken in consideration a life insurance company that operates in our domestic market. In order to accomplish the clients’ needs, this company intended to expand its network.

We have used GA in MATLAB, to give an optimal solution to this issue. Initially we have defined the physical point (coordinates) where the company could build its new business unit. The logic that we followed to

define the new coordinates is based on the fact that this new unit had to be close to an existing business units of the company, so that the company can then schedule better a distribution program in which it will organize its mean of transports in groups, in order to cover all the areas with a lower cost.
Our first step was collecting all the business unit coordinates of this company that operate in Tirana, and saving these data in an excel database. After we analyzed the possible coordinates for the new business unit at this point, the problem turned in determining the fitness function for GA. Our data were coordinates, so the problem escalated in finding the minimum distance between two coordinates. We have defined the formula which calculates this minimum distance, as fitness function for our GA. We have used existing mathematical conclusions to calculate this formula.
Before implementing and programming MATLAB script for the fitness function, we had to define the optimizer task. In this study is used cluster analysis as optimizer task. The main reason for using this task is justified by the fact that it has as main purpose the classification of objects in clusters so that the two objects in the same cluster are more similar than objects of a different cluster.[9] [10]

The company that we have taken into consideration had M - business units (BU), which based on the above logic cluster analysis, were divided into N groups. The BU characterized by k-variable, where k will represent the dimension of the vector that represents the coordinates. So we initially parted business units into clusters and then realized minimizing the variability of the objects within the group.

So initially, we started programming and implementing the formula of the fitness function in MATLAB, by setting: *i= {1,2….M}* and *j={1,2….N}*,which based on the technique of the clusters.
For these variables defined weight:

$$\boldsymbol{w_{ij}} = \begin{cases} 1 & \text{If the object i is part of the cluster j th} \\ 0 & \text{Else} \end{cases}$$

Matrix of weights and took the following form:

$$\boldsymbol{w_{ij}} \in \{0;1\} \quad \text{and} \quad \sum_{\boldsymbol{j}=1}^{\boldsymbol{M}} \boldsymbol{w_{ij}} = 1 \qquad (1)$$

After some arithmetic operations are performed, including the use of distance formula for finding the minimum distance between an object and a centroid we found below who represented the GA and fitness function in our study:

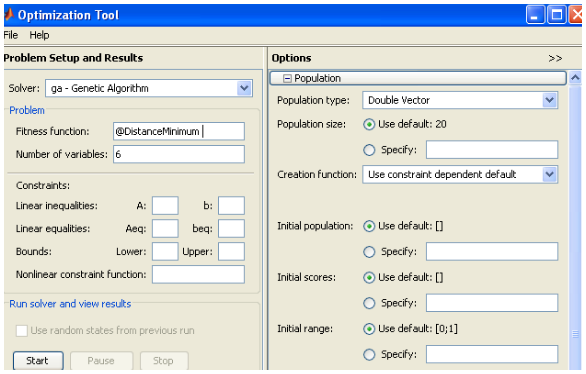
$$\boldsymbol{f}_{\min} = \sum_{\boldsymbol{i}=1}^{\boldsymbol{N}} \min_{\boldsymbol{j}\in(1,2,..\boldsymbol{M})} \left(\sqrt{\sum_{\boldsymbol{l}=1}^{\boldsymbol{k}} \left(\boldsymbol{x_{il}} - \boldsymbol{c_{jl}} \right)^2} \right) \qquad (2)$$

The above fitness function implemented in the MATLAB scripts [11], of which were given as GA input incorporated in MATLAB, gave us the minimum distance as a result of two points. Below we are going to shows the main script that was used to make the implementation of fitness function in MATLAB:

```
num=input(Numri i cluster-ave (grupeve):');
num=3*num;
PopSize=input('Përmasa e Popullatës:');
FitnessFcn = @Distances;
numberOfVariables = num;
LOCATION=(xlsread('Distanca',Vendodhja))
my_plot = @(Options,state,flag)
Draw3(Options,state,flag,LOCATION,num);
Options                                          =
gaoptimset('PlotFcns',my_plot,'PopInitRange',[0;
1],'PërmasaPopullatës',Po
pSize);
[x,fval]                                         =
ga(FitnessFcn,numberOfVariables,Options);
assign=zeros(1,size(LOCATION,1));
for i=1:size(LOCATION,1)
distances=zeros(num/3,1);
for j=1:(size(x,2)/3)
distances(j)=sqrt((LOCATION(i,1)-
x(j))^2+(LOCATION(i,2)-
x(size(x,2)/3+j))^2+(LOCATION(i,3)-
x(2*size(x,2)/3+j))^2);
end
[min_distance,assign(i)]=min(distances);
```

Figure.2 Main script

To present the results of our study, we used another script in MATLAB that realized a suitable interface. Then in the last phase we used fitness function @DistanceMinimum as input to the GA tool, as in the figure below:



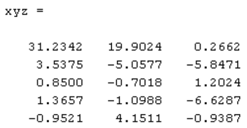
Solver: ga - Genetic Algorithm
Fitness function: @DistanceMinimum
Number of variables: 6
Linear inequalities: A: b:
Linear equalities: Aeq: beq:
Bounds: Lower: Upper:
Nonlinear constraint function:
□ Use random states from previous run

⊟ Population
Population type: Double Vector
Population size: ● Use default: 20 ○ Specify:
Creation function: Use constraint dependent default
Initial population: ● Use default: [] ○ Specify:
Initial scores: ● Use default: [] ○ Specify:
Initial range: ● Use default: [0;1] ○ Specify:

Figure.3 Fitness function in MATLAB

Once GA executes the fitness function built on pooled data from us and that were imported into MATLAB from excel file, provides the result as follows:
Number of clusters: 5

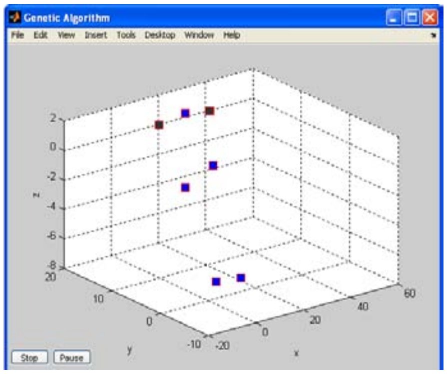
```
Population size:     50
Optimization terminated: average change in the fitness value less than
option .TolFun.
assign = 1   1   1   1   1   1   1   1   1   1   1   1   1
fval = 132.6080
```



```
xyz =

    31.2342   19.9024    0.2662
     3.5375   -5.0577   -5.8471
     0.8500   -0.7018    1.2024
     1.3657   -1.0988   -6.6287
    -0.9521    4.1511   -0.9387
```

GA were used to process a population of size 50, which means 50 business units of the insurance company in Tirana, which were divided into five groups or clusters. The result that achieved was worth 132.6080 minimum distance and it reached a minimum distance to the cluster containing xyz coordinates defined as in the figure above, the coordinates that represent potential new unit that the company wanted to open. Graph 1 below present the dynamical three-dimensional.



Graph 1

As a result of this study we analyzed that the time it took to determine the outcome is 3 minutes, while the economic cost could say that was negligible for consideration.

We found that for the future, the GA is expected to become an essential technique in order to resolve problems of optimization.

4. Conclusions

In this paper we argued why GA are important and why aspire to become future main techniques and the most used in the economy. We gave an overview of the current situation in Albania, by comparing it with the situation in developed countries of the world. We also introduced the main logic function prototype for genetic algorithm. Then we presented presentation of our concrete study. Our goal was to show how genetic algorithms implemented in MATLAB with cluster analysis can be used to find the minimum distance between two coordinates. We used MATLAB as framework for the execution of the GA and display the results in a format suitable GUI. We tried to show why this optimal solution that we are offering is important for insurance company and other similar companies that operate in the Albanian market.

5. References

[1] Goldberg, D. E. Book "Genetic Algorithms in Search, Optimization, and Machine Learning. Reading, Massachusetts": Addison-Wesley, 1989.

[2] Mitchell, M. Book "An Introduction to Genetic Algorithms. Cambridge" MIT Press, 1996.

[3] A.H. Wright, "Genetic algorithms for real parameter optimization, in Foundations of Genetic Algorithms, J.E. Rawlins (Ed.), Morgan Kaufmann, pp. 205-218, 1991.

[4] "Practical genetic algorithms", by Randy L.Haupt, (2004), John Wiley & Sons Inc.

[5] K.F.Man, K.S and Tang, S.Kwong (2001) "Genetic Algorithms: Concepts and Designs"

[6] Riechmann, Thomas. "Learning in Economics: Analysis and Application of Genetic Algorithms". Heidelberg: Physica-Verlag, 2001

[7] K. Deb et al., "A Fast Elitist Non-Dominated Sorting Genetic Algorithm for Multi-Objective Optimization: NSGA-II," Proceedings of the Parallel Problem Solving from Nature. Springer Lecture Notes in Computer Science No. 1917, Paris, France, 2000.

[8] Stefano Monti, Pablo Tamayo, Jill Mesirov, and Todd Golub. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning, 52(1-2):91–118, 2003.

[9] L.O.Hall, B.Ozyurt, and J.C. Bezdek. Clustering with a genetically optimized approach.In Transations on Evolutionary Computation, volume 3, pages 103–112, Department ofComputer Science and Engineering, University of South Florida, Department of ComputerScience and Engineering, University of West Florida, 1999. IEEE

[10] Anand Kumar, Dr. N. N. Jani. "A Novel Genetic Algorithm approach for Network Design with Robust Fitness Function", IJCTE journal, Vol. 2, No. 3, June, 2010

[11] http://www.mathworks.it/gads/genetic-algorithm.html

Brunela Karamani is a pedagogue in Polytechnic University, in Computer Engineering Department. She has 12 years teaching experience in computer science. In 2010 she has finished the Master Thesis and now is PhD student. Her research areas of interest are Data mining and Statistical solution.

Esteriana Haskasa is a pedagogue at the University of Elbasan "Aleksander Xhuvani". She has 4 years of programming experience, and 2 years of teaching experience. Her research areas of interest are Artificial Intelligence and Robotics.

Shkelqim Kuka is a Ass.Professor in Polytechnic University, in Department of Mathematics. He has 25 years teaching experience in algebra and applied mathematics. His research areas of interest are applied mathematics and machine learning.