Proceedings of the 8th International Conference on Applied Informatics
Eger, Hungary, January 27–30, 2010. Vol. 1. pp. 41–51.

Genetic Algorithm in Default Forecast

József Bozsik

Eötvös Loránd University – Faculty of Informatics
e-mail: bozsik@inf.elte.hu

Abstract

In this article I would like to describe a genetic algorithm, which like the well-known discriminance analysis is able to produce a weighted-sum function, which can be used for financial default forecast given well defined conditions. In this method a linear combination of financial ratios can be used, which can be calculated from the annual Balance Sheet and Profit and Loss Statement of the companies. Using this function a classification rule can be set up just like in case of discriminance analysis. With this classification we are able to give default forecast using the financial ratios of the company. I used real companies and real data (2008) for testing the heuristic method. I compared the results with the results of a now-used economical model (discriminance analysis). The comparison shows the reliability of the method and the influence of each parameter to the reliability of the result.

Keywords: genetic algorithm, default forecast, finance default, discriminance analysis

1. Introduction

I would like to describe a heuristic algorithm for in the economics well-known problem, namely the default-forecast. This is an important issue in the current financial situation, especially if we think about the financial crisis. Default models are used to project defaults in the economics. One of the famous models is the so called discriminance analysis. I would like to develop the new heuristic model according to the base of this model. In order to solve the problem I use the results of the well-known genetic programming. The genetic algorithm can be used in those problems, where the explicit connection between the input variables and the expected result is unknown or it is not cost-effective to determine this connection. In order to use these genetic algorithms in an effective way, we need to measure the "goodness" of the function. For this measurement I used test data, where the expected result is known, so we can determine the well-known fitness-function. With the help of this method we can develop an approximate method, which results

a classification function. Using this function and the appropriate financial ratios the model is able to project the financial default of a company.

2. The reference model

In order to understand and be able to analyse the new developed model, it is necessary to understand the economical model. In this article I would like to introduce the well-known model, the so called discriminance model. This model is the reference model, which is well-known and worldwide-used model. I would like to describe the reference model only that deep which is necessary. After a short overview I would like to introduce the results of the discriminance analysis. This model's result gave me the inspiration to build up a model based on new approach. I would like to describe the problems which occurred during building the new genetic algorithm and the solutions in details with the partial results.

2.1. Discriminance analysis model

"The discriminance analysis with more variables analyses the distribution of more ratio in the same time and sets up a classification rule, which contains more weighted financial ratio (these are the independent variables of the model) and summarize them in only one discriminant value." [1] The most important criteria of choosing the financial ratios which will be used in the model is that the ratios should have low correlation to each other. Otherwise the added value for the classification will be low. It is worth to begin the construction of the model with a significant ratio and in each further step the less correlated but the second significant ratio should be involved. In the discriminance-function which is a linear combination the values of the financial ratios should be substituted which are calculated from the annual report of the individual companies. In order to be able to classify if a company is able to pay or not, we have to compare the values with the discriminance value.
The general formula of the discriminance function is the following (see [2]):

$$Z = w_1X_1 + w_2X_2 + \cdots + w_nX_n \tag{2.1}$$

In the formula used signs are:

- Z – discriminance value
- w_i – discriminance weights
- X_i – independent variables (financial ratios)
- $i = 1, \ldots, n$ where n means the number of financial ratios

The analysis of 2008 as well as the analysis from 1996 showed that the companies which are able to pay and which are not are different in the following financial ratios:

- X_1 – quick liquidity ratio[1]

[1] This ratio shows if the company is able to pay immediately.

- X_2 – cash flow/total debts
- X_3 – current assets[2]/total assets[3]
- X_4 – cash flow/total assets

The order of the financial ratios is reflects the discriminance power of the ratios, it means that the most discriminative ratio is the quick liquidity ratio after that the three other ratios. The discriminance function is made by the involvement of these ratios based on the data of 2008 (see [3]):

$$Z = 1,2387 \cdot X_1 + \ 1,7153 \cdot X_2 + \ 2,9761 \cdot X_3 + \ 0,07158 \cdot X_4 \tag{2.2}$$

The critical Z value is 2,39756, it means if we substitute the values of a company's financial ratios and the result is higher than 2,39756, the company will be classified by the function as solvent (it is able to pay), otherwise insolvent (it is not able to pay).

2.2. The results of the discriminance analysis for the test data

During the test I chose randomly 200 companies from the open database of the webpage of the Hungarian Ministry of Justice and Law Enforcement (http://www.e-beszamolo.irm.hu/kereses-Default.aspx).

100 companies were solvent the other 100 insolvent. I applied the data for the introduced discriminance function; the result is shown in the Table 1:

Type of results	The result of the discriminance analysis
Wrong classified companies (pieces)	24
Wrong classified companies which are able to pay (percent)	12%
Wrong classified companies which are not able to pay (pieces)	18
Wrong classified companies which are not able to pay (percent)	9%
Total (pieces)	42
Total (percent)	21%
Classification accuracy (percent)	79%

Table 1: The results of the discriminance analysis based on data of 2008

Our aim is to set up a default forecast model which gives higher accuracy than the model described above.

Before moving on to describe the iterative way for building up the model based on new genetic algorithm, I would like to introduce the genetic algorithms in general. Here I will introduce the used terms and indicators, because they are necessary for the understanding of the new model.

[2]Current assets: inventories, receivables, cash and cash equivalents.
[3]Total assets are those assets which are used within one year in the company.

3. The genetic algorithm model

3.1. Nominations and definitions

I will use the basics of the classical genetic algorithms. During the experiments I built up more models. I would like to analyse and introduce every mile-stone. In order to understand the construction of the model, the most important definitions and nominations have to be introduced.

The model's base is the classical model (see [4]), so first of all let me introduce the abstract algorithm of genetic algorithm, the connected definitions and nominations.

The genetic algorithms operate on the given population's individuals. The population contains of individuals, each individual is determined by the chromosome. The chromosome is built up based on a well defined sample, it is important for the modelling to choose this sample (see [5]).

The other very important issue is the fitness function which is based on the chosen chromosome structure. The fitness function is one of the key parameter of the genetic algorithms, because the accuracy of the result and the convergence depends on the chosen fitness function.

In order to model a problem we have to choose the characteristics for the genetic algorithms. We have to determine the individuals genetic patterns. If all the parameters are determined, the algorithm can be started:

1. Generate the initial population, it means choose the individuals for the first population.

2. Calculate the fitness function for each individual in the population.

3. Select the most capable individuals.

4. Generate new individuals by crossover and mutation from the selected individuals.

5. Check the new individuals. If we find among the new individuals which fits to the fitness value, the algorithm stops running. If there is no individual among the new individuals, which would be acceptable to the fitness value, the algorithm will start again from the 2^{nd} step.

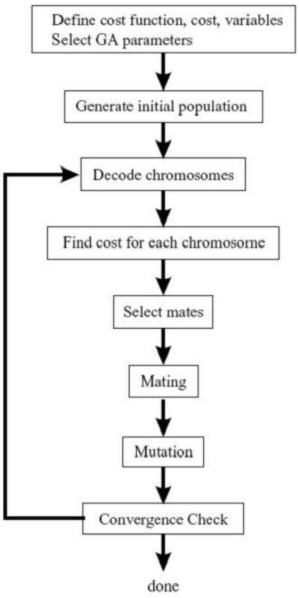

Figure 1: Abstract GA

I would like to introduce the concrete definitions during the introduction of each partial model. In the next paragraph I will show the during the experiment built models and the final model's construction and results.

3.2. The basic model

As a first step I would like to introduce the problem's representation. The gens of chromosomes which represent the individuals in the model are the weights of the financial ratios used for the default forecast; it is shown in the Figure 2. This representation fits well to the introduced economic model, to the discriminance analysis.

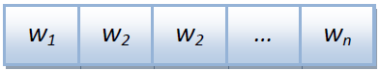

Figure 2: Representation of individual

3.2.1. The fitness function

In this article I mentioned the fitness function as an important parameter. In order to construct the fitness function I used the basic knowledge of economics. The fitness function represents the classification accuracy of a given individual. The classification accuracy is not determined for the total sample, but for the smaller, the so called test sample.

Let be $P \subseteq (\mathbb{R}^n \times \{0,1\})$:

$$P = \{\ p \in (\mathbb{R}^n \times \{0,1\}) \ \mid\ \Delta\,(p_1,\ldots,p_n) = p_{n+1}\ ,\ n \in \mathbb{N}\}$$

where Δ:$\mathbb{R}^n \longrightarrow \{0,1\}$ is:

$$\Delta\,(x_1,\ldots,x_n) = \begin{cases} 0, & if\ x\ insolvent \\ 1, & if\ x\ solvent \end{cases}$$

From P (pattern-set) is the test-set established: $T \subset P$, the multiplicity of T must be much smaller than the multiplicity of P, it means:

$$T \subset P:\ |T| \ll |P|$$

Let be the multiplicity of T and P by:

$$m_p := |P|$$

$$m_t := |T|$$

with these indicators the fitness function can be defined as:

$$f : \mathbb{R}^n \longrightarrow \mathbb{N}_0^+$$

$$\forall q \in Q\ : f\,(q) = \sum_{t \in T} \Phi(t_{n+1}, \varphi(w_1,\ldots,w_n))$$

Q indicates the set of the individuals which are in the actual population which is used in the actual step of the algorithm's iteration. The individuals are the elements in the population.

$$\Phi : \{0,1\}^2 \longrightarrow \{0,1\}$$

$$\Phi\,(\mathrm{x},\mathrm{y}) = \begin{cases} 0, & \mathrm{x}+\mathrm{y}=1 \\ 1, & x+y \neq 1 \end{cases}$$

$$\varphi:\ \mathbb{R}^n \longrightarrow \{0,1\},\quad \xi \in (0,1)$$

$$\varphi\,(w_1,\ldots,w_n) = \begin{cases} 0, & \sum_{i=1}^n w_i x_i < \xi \\ 1, & \sum_{i=1}^n w_i x_i \geq \xi \end{cases}$$

During the experiments the system was not sensitive for $\xi \in (0,1)$. This happens because on the $(0,1)$ interval only one borderline is between the solvent and insolvent companies. Naturally the change of the borderline causes the changes of the weights, but it has no effect for the classification accuracy of the system. In the experiments the $\xi = 0,5$ value is used.

To make work the algorithm the method must be defined, which will be used for the crossovers and the mutation. For the crossovers I tried to do some sophistications, I would like to describe them later in details.

The mutation is random in the meaning that the mutating gen is chosen by chance, but the mutation itself will work after the rules defined before. The experiments verified 20% mutation rate as the most effective.

It is true in general for all models like the mutation rate that the initial population must be initialized by random data.

3.2.2. Change of population

In the classical model the traditional model is used for choosing the initial population. By this step it is made sure that in every population the same number of individual is chosen. In the transitional population first the individuals are crossovered in pairs, than each individual is mutated by the introduced 20% mutation rate (Figure 3).

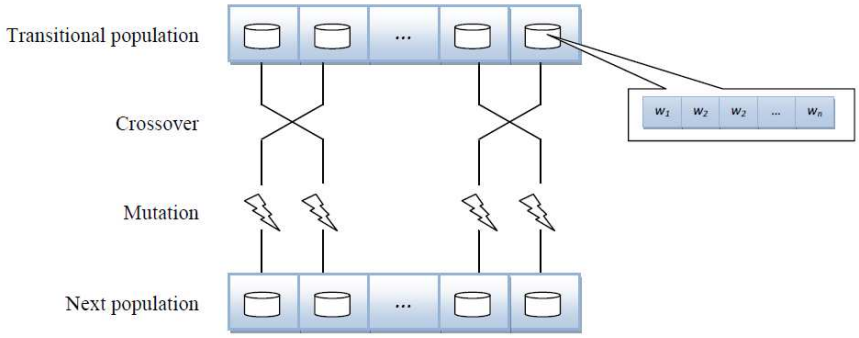

Figure 3: Crossover

For all sophisticated model this basic model is used. The difference between the models is in the crossover.

3.3. Basic crossover

The crossover (re-combination) basically creates one or two individuals from two parent-individuals. During the experiments after every crossover two child-individuals are created in order to leave the same level of multiplicity of individuals in the population.

I tried to construct more crossover models, but none of them gave better results than the others. The tried models were: one-point, two-points, N-points and uniform crossover. From these models the uniform crossover model seems to be the best.

The main point of the uniform crossovers is that the gens of the children chromosomes (individuals of the previous population) are created from the parent-gens in the same population, by 50-50% chance from the one or the other parent (see [6]). This is shown in the following figure (Figure 4).

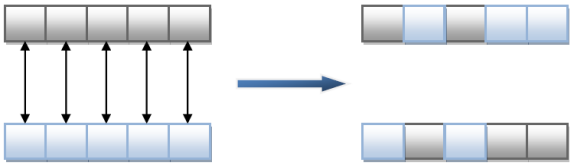

Figure 4: Simple crossover

After using the chosen uniform crossover and sophistications the configuration which gave the best results showed the following results (Table 2).

Name	The results of the basic model
Wrong classified solvent company (pieces)	40
Wrong classified solvent company (percent)	20%
Wrong classified insolvent company (pieces)	38
Wrong classified insolvent company (percent)	19%
Total (pieces)	78
Total (percent)	39%
Classification accuracy (percent)	61%

Table 2: Result of simple crossover

This 61% classification accuracy is worse than the other accuracy given by the classical model's result. This suggests more sophistication. I changed the operation of the crossover, which indicated a new model.

3.4. Sophisticated crossover

There is an interesting outcome after the experiments with the crossovers. By observing the uniform crossover and the inherited characteristics between the populations, it became clear, that those individuals whose gens partially did not change after some generations, have better characteristics until one untouched gen-section is restructured.

This observation leads me to the model where the crossover must be on only one gen-section to achieve the given accuracy. The given accuracy must be set up based on experience, it is in practice the accuracy which can be measured if the convergence decreases dramatically. The method works as following:

1. In the initial population the gens must be declared, which can be crossovered until achieving the given accuracy.

2. In the next step taking the last population with the most fit individuals, the earlier chosen gens must be "frozen", which means that these genes will not be restructured during the process, they don't take part in the crossover again.

It is obvious that the frozen gens can be replaced by the index function to the first n gene place in the chromosome. With this method is another method equivalent by a well chosen index function in which we choose the first n genes in the chromosome and use them on the described way.

The main steps of the method are following:

1. *step:*
   The genes must be fixed which should be frozen, and using the classical uniform crossover the algorithm should run until achieving the threshold. Than the algorithm stops and the output population will be fixed (see Figure 5).

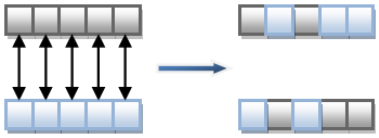

Figure 5: Uniform crossover

2. *step:*
   Each gen set of each individuals fixed in the first step should be extended by relevant extra gens. The new gens are filled randomly by gens. This is how we can get the total gen set for the solution (see Figure 6).

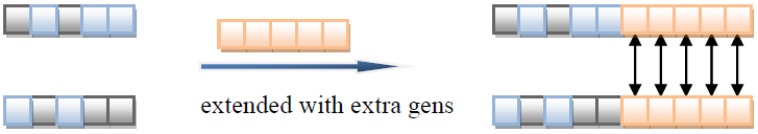

Figure 6: Freezing and extending step

3. *step:*
   The genetic algorithm should continue on the full genset, so that the frozen gens are not crossovered. The algorithm runs until achieving the given accuracy.

The experience with this method shows that the most efficient for the default forecast model is freezing the first 4 gens. The model based on this idea gave the following results (Table 3):

Name	Result of the sophisticated model
Wrong classified solvent company (pieces)	30
Wrong classified solvent company (percent)	15%
Wrong classified insolvent company (pieces)	31
Wrong classified insolvent company (percent)	15,5%
Total (pieces)	61
Total (percent)	39,5%
Classification accuracy (percent)	68,5%

Table 3: Result of first sophisticated model

3.5. Final model

I constructed the final model by using the before introduced sophisticated model. In this model I used another sophistication, which means running the same method again. In the model I used 200 company data. The population contained 1000 individual and for the first freezing the algorithm stopped by achieving 60% accuracy,

for the second freezing 70% accuracy. The second sophistication used in the final model is shown in the next figure:

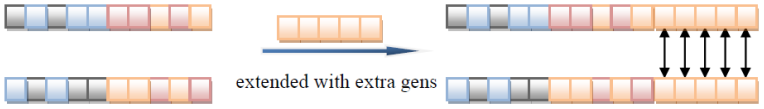

Figure 7: Second freezing and extending step

Finally I managed to reach a model with two freezing and sophistications, which classification accuracy could achieve 84,5% accuracy. The result is shown in the next table (Table 4).

Name	The result of the final model
Wrong classified solvent company (pieces)	14
Wrong classified solvent company (percent)	7%
Wrong classified insolvent company (pieces)	17
Wrong classified insolvent company (percent)	8,5%
Total (pieces)	31
Total (percent)	15,5%
Classification accuracy (percent)	84,5%

Table 4: Result of final sophisticated model

4. Summary

The result of the experiment proves that the genetic algorithms can be used in the economic default forecast models. The artificial intelligence can be used with the same efficiency than the traditional methods. It is important that after some modifications according to the problem became the 84,5% classification accuracy achievable.

The model based on genetic algorithms which is built up on 200 company-data is able to achieve better classification accuracy using the two-steps freezing method than the classical discriminance analysis model. It should not be forgotten that the model's ability was trained and tested on a relative small sample. It can be aim of further experiments to build a sophisticated model on bigger sample, or to examine the classification accuracy by changing the parameterization. Furthermore it can be important to examine new models with other structures.

The sophisticated genetic algorithm model gave better result compared to the discriminance analysis, but we have to consider the important fact, that the discriminance function was made based on 500 randomly chosen company data. This is important remark, because this model is trained and tested only on 200 company data. Despite the fact it was able to achieve very good results compared to the classical method.

References

[1] KOVÁCS ERZSÉBET: Pénzügyi adatok statisztikai elemzése, Egyetemi Tankönyv, *Budapesti Corvinus Egyetem Pénzügyi és Számviteli Intézet*, (2006)

[2] VIRÁG MIKLÓS, HAJDU OTTÓ: Pénzügyi mutatószámokon alapuló csődmodell-számítások. *Bankszemle,* XV. évf. 5. sz. (1996), 42–53.

[3] KOVÁCS GÉZA: Diszkriminanciaanalízis 2008. évi adatok alapján, *Online Study:* `http://kozgazdasag.uw.hu/kovacsg/diszkrim2007/diszkrim2007.html` (2009.09.10.)

[4] VÁRKONYINÉ KÓCZY ANNAMÁRIA: Genetikus algoritmusok, *Typotex Kiadó,* Budapest, (2002)

[5] WOLFGANG BANZHAF, PETER NORDIN, ROBERT KELLER, FRANK FRANCONE: Genetic Programming – An Introduction, Morgan Kaufmann, San Francisco, CA. (1998)

[6] DAVID E. GOLDBERG: Optimization and Machine Learning, *Addison-Wesley Longman Publishing Co., Inc.,* Boston, MA, (1989)