Genetic algorithms in forecasting of Internet shops demand

Chodak, Grzegorz

Wrocław University of Technology, Poland

2009

Grzegorz Chodak*

GENETIC ALGORITHMS IN FORECASTING OF INTERNET SHOPS DEMAND

The general aim of this article is to present genetic algorithms as a tool, that can be used in demand forecasting in internet shops. First part of article identities factors, which have to be taken into consideration during analysing demand in internet shops, e.g. dispersion of demand, delivery time influence and different e-marketing factors. Specific form of used demand function is shown in the next section of the article. Then genetic algorithm is defined by its genetic operators acting on bit strings (examples of the operators are: crossover, inversion, and mutation) and its method of credit allocation (fitness evaluation and selection). Next the method of identification of the function parameters using genetic algorithms is shown. The next part of article shows appliance of presented genetic algorithm. The advantages and disadvantages of proposed method are shortly discussed in summary.

1. INTRODUCTION

Demand forecasting in internet shop is very interesting and also difficult matter, because of specific character of electronic commerce. Diversity of customers, digital and virtual way of shopping can cause forecast miscalculation and bigger error. There is a question: if demand forecasting in such a dynamic environment is reasonable. The answer is: yes, even if we know that exactness of forecast is not perfect, because demand forecasting can facilitate lowering stocks level. Proper inventory control is one of the most important problem in internet shop, because it leads to great cost reduction, and the profit increase. Demand forecasting tools can give the manager knowledge useful in inventory control, so it should be used regularly.

---

* Industrial Engineering and Management Institute, Wroclaw University of Technology

The research done by author, shows that only 12% of Polish internet shops tries to forecast the demand. From this 12% only 32% use these tools often or very often. 22% of this 12% say, that tools they use are too complicated; 26% of this 12% claim that forecasts were not accurate [9]. This data give general idea, that tools for demand forecasting are not popular in Polish e-commerce and there is a need to propose a tool that could be more exact and easy to use for managers. Many quantitative methods described in literature (e.g. [2], [3]) requires from managers understanding complicated mathematical equations, therefore AI approach has a chance to get more popular than "traditional" methods.

Applying artificial intelligence techniques, such as genetic algorithms (GA) should improve correctness of demand forecasting. A presentation of demand forecasting method in e-commerce environment is the main objective of this article. Identification of demand function's parameters is based on genetic algorithms approach.

1. DEMAND IN INTERNET SHOP

In demand forecasting area, it is very important to understand the characteristics of demand environment. In this part of article some features of demand in internet shops are discussed.

Description of customers in internet shops differs from the average customer in brick-and-mortal shops. First of all there is no geographical barriers in internet. We can assume that potential client of e-shop is a person connected to internet with no language barrier. When we consider multilanguage user interface, the number of potential customers can be counted in hundreds millions.

No geographical barriers leads to huge diversity of clients - different tastes, interests, hobbies, needs etc. Heterogeneousness of customers usually causes flattening of cumulative distribution, that shows distribution of sold commodities. In internet shops, which sell huge range of assortment, cumulative distribution resemble long tail, what is the proof, that in internet one can sell not only best-sellers. This heterogeneousness of customers diversifies demand and wider range of assortment can be sold. But there is also disadvantage of such a lack of customers concentration – demand forecasting is very difficult, because of demand irregularity. Diversity of customers leads to variable stream of orders, therefore shape of demand function could be very rough.

Marketing activities have a great influence in internet shop. In traditional shops there is very difficult to advertise directly. Marketing in internet shops give the possibility immediate action after viewed advertising. This e-marketing feature causes rapid increase of demand after marketing campaign (for example e-mailing, banner campaign etc.). We can assume that demand in internet shop stronger correlate with marketing

activities than in traditional trade, therefore forecasting demand in e-shop must take into consideration information about all undertaken and future marketing campaigns.

Another important marketing tricks which have influence on demand are cross-selling techniques and recommendation systems. Very popular cross-selling form proposes a customer goods which bought clients who bought also the good presented to customer (e.g. "customers who bought this item also bought ..."). When we forecast demand it's very important to take into consideration complementary relation between goods. When the recommendation of goods is given by administrator of the system it's easy to predict the linkage between goods and use it building the demand function. When the recommendation system builds the displayed list of goods dynamically, analyzing the sales and paged viewed by customers, it's much more difficult to build demand function properly.

Some of internet shops give the customers information about the potential time of delivery. Usually when the shop has commodity on stock, time of delivery is 24h, and when commodity is not available in inventory, the delivery time is 3-5 days. Customer who is interested in quick delivery probably will resign knowing that e-shop has not wanted product on stock. It means, that time of delivery should be taken into consideration during analyzing demand function.

The another classical marketing instrument is price. The price policy and promotion of selling depends on price elasticity of demand. In brick and mortal shops customers have limited possibility to compare prices in many shops. It takes time to get information about market prices in traditional trade. When we consider e-commerce, the price comparison is quite easy and quick. There are many price comparison services (e.g. Pricegrabber.com, Ceneo.pl) which give possibility to compare prices in many shops immediately. Also search engines like Google makes it easier to find many shops and check the prices. It causes that in internet shops price elasticity of demand should be higher than in brick and mortal shops [10]. During demand forecasting it is very important to take into consideration price elasticity of demand.

2. DEMAND FUNCTION

Proposed demand function (equation 1) takes into consideration chosen features of e-commerce like: strength of promotion, strength of recommendation, influence of delivery time. It also includes standard features like trend factor, seasonality and price elasticity of demand and vertical offset.

$$D = \frac{E \cdot ISP + F \cdot NLP + C + B \cdot t + A \cdot \sin(\omega \cdot t + \varphi)}{p^e}(1 - G \cdot \frac{TD}{100}) \quad (1)$$

▶ D - demand;

Variables:
- ▸ ISP – is promoted (0 or 1).
- ▸ NLP – number of linked products
- ▸ TD – time of delivery
- ▸ t – time
- ▸ P - price;

Parameters:
- ▸ B – trend factor
- ▸ C – vertical offset
- ▸ E – vertical offset connected to promotion (strength of promotion)
- ▸ F - vertical offset connected to recommendation techniques (strength of recommendation)
- ▸ A – amplitude of fluctuation;
- ▸ ω - frequency (representing periodicity);
- ▸ φ - horizontal offset
- ▸ e – price elasticity of demand
- ▸ G – time of delivery factor

The above function has nine parameters (A, ω, φ, C, B, e, E, F, G) which values ought to be identified using real data. Applying analytical method for evaluation (identification) values of these parameters is very difficult. Therefore it's appropriate to apply efficient approach of GA to identify that parameters.

2. APPLYING GENETIC ALGORITHM TO IDENTIFY THE DEMAND FUNCTION'S PARAMETERS

A genetic algorithm is a search technique, used usually in computing to find approximate solutions to optimization and search problems. GA are categorized as the global search heuristics. GA are a class of evolutionary algorithms (EA) that use techniques inspired by evolutionary biology such as inheritance, mutation, selection, and crossover [7]. Details of GA will not be discussed here, as it can be found in many publications, e.g., [1], [4], [5], [7].

First step in applying GA to our problem of parameters' identification is to define a fitness function (FF). It ought to be a measure of "distance" of model generated data and collected real data [8]. In proposed experiment the following fitness function is used:

$$FF = \sum_t \left| D_t - \left(\frac{E \cdot ISP + F \cdot NLP + C + B \cdot t + A \cdot sin(\omega \cdot t + \varphi)}{p_t^e} \cdot (1 - G \frac{TD}{100}) \right) \right| \quad (2)$$

t – time;

D_t – real sell in t time;

P, ISP, NLP, TD – variables, the same as in 1 equation;

A, B, C, E, F, G, ω, φ, e – the same as in 1 equation – these values of parameters should be identified.

The structure of individuals in proposed GA is following: every individual consists of 1 chromosome divided into nine segments. Each segment represents one identified parameter. Chromosome is composed of 90 bits, divided into 10 bits segments, the segments code values of amplitude, trend factor, vertical offset, promotion strength, recommendation strength, delivery time factor, frequentness, horizontal offset and price elasticity factor. Binary code is used and every gene is coded as one bit. An example of individual is presented at Figure 1.

A	B	C	E	F	G	ω	φ	e
1	0	0	1	1	0	0	1	1
0	1	1	0	0	1	1	0	0
0	1	0	0	1	1	0	1	0
0	1	0	0	0	1	0	0	0
1	1	1	0	1	1	1	1	1
1	0	1	1	0	0	1	1	1
0	1	1	1	0	1	1	0	1
1	1	0	0	1	1	0	0	0
0	0	1	1	0	0	1	1	1
1	0	1	1	1	0	1	1	1

Figure 1. GA individual

The size of chromosome (90 bits) is a compromise between exactness of parameters representation and quickness of computing. The greater number of bits for each parameter representation allows more exact evaluation of its value but length of calculation grows rapidly. Of course contemporary supercomputers could calculate GA with much bigger chromosome, but this size can be also calculated on average PC.

It's used two main genetic operators, namely: mutation (inverts one bit in a chromosome) and crossover (exchanges of bits string between two chromosomes).

The next important phase of GA application is proper definition of the credit-allocation scheme. It ought to be based on relative fitness over the entire population of solutions to the problem so that a given individual has a probability of entering into the next generation according to its relative fitness. In proposed experiment modified roulette method is used. Each individual (chromosome) has assigned piece of a roulette wheel, which size is proportional to its fitness. The higher value of fitness function is has, the more probably it will be chosen to the next generation. The whole wheel

represents a sum of fitness of all individuals of the population [5]. Additionally the individual with best fitness is chosen to the next generation (the rule "don't lose the best") [8].

To make GA search more efficient it is important to estimate the scope of the parameters' domain. In proposed experiment the heuristic knowledge about the explored demand is used. For example, to evaluate the range of amplitude (A) one can estimate maximum and minimum amount of sell. Maximum vertical offset (C) ought to be located between zero and maximum of demand function. The range of horizontal offset (φ) ought to be placed in the range (0-2π). The variability range of price elasticity of demand (e) is related to individual good and can be set up by a sell manager (plausible values are 0 to 2). Range of frequency (ω) can also be set by manager who can be able to predict the seasonal demand. Accepted values of the trend parameter are related to dynamics of increase (decrease) of sell. In e-shops we can observe rather big dynamics, so it can set up between -1000 and 1000.

Strength of promotion (E) and strength of recommendation (F) are parameters which are similar to range of vertical offset, so they should be located between zero and maximum of demand function. If we take into consideration negative influence of marketing activities, range of these parameters should be between minus maximum of demand function and maximum of demand function. Influence of delivery time (G) should be in range (0, 100/TD), when TD is mentioned time of delivery.

If one of parameters values got from optimisation experiment is getting closer to the limit of range of domain, the rage should be shifted in such way that value of parameter will be near the centre of new range [8].

If the ranges of parameters are set up accurately, the values of identified parameters will also estimated quicker and more precisely. It is important when we make calculations (simulation) on not very powerful computer (like average PC class machine).

Definition of the ending (breaking) condition is the important for proper GA application. There are two possibilities of breaking the algorithm: when the settled number of generations is achieved or when fitness is less then settled error. Both criteria are used in proposed algorithm.

3. EXPERIMENT DESCRIPTION AND RESULTS

Experiments were done using software written by author. Proposed GA was developed using VBA language connected to Microsoft Excel environment.

After test experiments the following parameters of GA were set up:

generation_size = 60
crossing_over_probability = 0,4
mutation_probability = 0,1

every_feature_size = 10

Table 1 Given input data

Month	Sell	Price [PLN]	Is promoted	Linked products	Delivery time
0	83	560	1	6	1
1	89	575	1	6	1
2	49	575	0	2	1
3	41	575	0	2	1
4	25	575	0	2	1
5	29	575	0	2	1
6	75	575	1	5	1
7	87	578	1	5	1
8	84	595	1	5	1
9	88	595	1	5	1
10	95	595	1	7	1
11	78	595	1	3	1
12	75	610	1	3	1
13	81	610	1	3	1
14	89	610	1	3	1
15	22	612	0	1	1
16	32	610	0	2	1

For data from Table 1 GA identified parameters of demand function given by equation (1). The demand function which was chosen to the experiment is very irregular, to present, that even in such a rough shape of demand function, the proposed GA can identify parameters with relatively small error.

The result of identification experiment: 10 best individuals and fitness function values are presented in Table 2.

Table 2 Results of simulation (identification) experiment

Amplitude A	Frequency ω	Elasticity e	Horiz. offset φ	Vert. offset C	Trend factor B	Prom. strength E	Recomm. strength F	Deliv. time influence G	Fitness FF
17693,1	**1,118**	**1,226**	**5,703**	**39296,8**	**2052,8**	**69208,5**	**21115,2**	**6,16**	**836,7**
14076,2	1,112	1,226	5,703	39296,8	2052,8	69208,5	21115,2	6,16	896,5

18963,8	1,021	1,258	0,283	20430,9	6158,4	51906,6	35191,3	18,38	1796,3
65591,4	1,016	1,257	0,604	93255,2	9970,7	58944,7	85728,4	60,61	1970,2
13392,0	2,980	1,276	4,514	80743,1	11241,4	75171,3	73118,5	53,67	2310,3
18866,1	2,980	1,294	4,526	80743,1	11143,7	75171,3	73118,5	50,15	2572,5
70088,0	2,898	1,294	4,083	79179,1	11143,7	75171,3	73216,3	50,15	3434,9
97,8	2,246	1,415	0,179	42424,8	25806,5	96187,7	75953,3	10,56	3732,3
19061,6	1,710	1,023	3,529	11731,1	977,5	27469,0	38026,0	77,91	4340,3
64907,1	1,047	1,394	0,062	51026,9	39491,7	49267,4	71945,5	18,67	6918,7

The best individual (i.e. the bolded one in Table 2) is used to calculate the overall error (defined by equation (3)).

$$E = \sum_{t}\left|Dr_t - Dp_t\right| \qquad (3)$$

E – overall error;
Dr_t – given demand;
Dp_t – demand counted using identified values of parameters (eq. (2));
t – time.

The sum of error equals to 102 and the average error equals to 6,4 – it means that the error equals 9.1% of average monthly sell.
Graphic representation (see Figure 2) also shows that the fitting of the given demand and identified demand is adequate. Of course the proposed approach ought to be tested on different types of given demand functions.
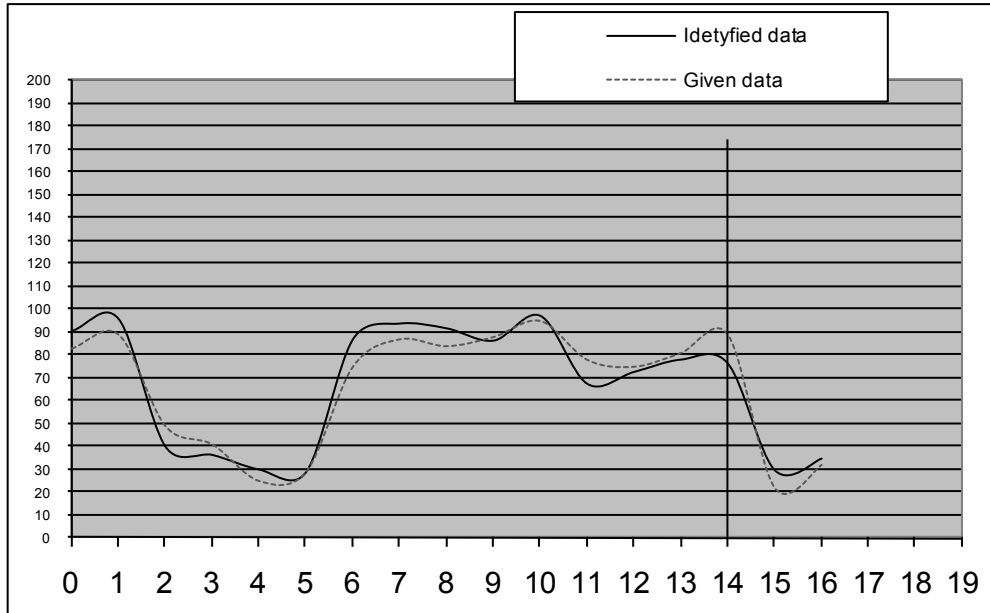
Figure 1 Given and identified demand curve

A prediction for next two months was made using identified values of parameters. The result of this prediction is presented in Table 3. Comparing the prediction with given sell gives the sum of errors in next 2 months equal to 10,8, i.e., 20% of given sell.

Table 3. Demand forecast on next moths

Month	Given sell	Demand forecast	Error
15	22	29,9	7,9
16	32	34,9	2,9

SUMMARY

As it was mentioned demand in internet shops can be very irregular and hard to predict correctly. It depends on marketing activities, which have a great influence on it, therefore method of demand forecasting should take this activities into consideration. Identification of demand function parameters using genetic algorithms can be efficient, when the shape of demand curve is irregular. In such a situation AI methods are probably more effective than others (compare with [2], [6]). In the presented article demand function is seasonal and has linear trend, and depends on price, marketing activities and

delivery time, but propose method is more general – there is no limitation to adjust the shape of the demand function to specific requirements, e.g. dependence on the number of position in Google search list. One of the greatest advantage of proposed forecasting method is flexibility – number of identified parameters of demand function can be increased or decreased, and it will only cause change of calculation time. Also demand function can be modified freely and algorithm will work in the same way, but forecast error is unpredictable.

The biggest disadvantage of proposed method is lack of certainty that satisfied solution will be found. Another disadvantage is unpredictable time of calculation. Sometimes identification of function parameters give satisfying results after a few generations and sometimes it takes hundreds of generations.

Results of experiments based on irregular sets of data suggests good applicability of this method, but it has to be said that AG does not work well for all data.

REFERENCES

[1] GOLDBERG D., *Algorytmy genetyczne i ich zastosowania*, Wydawnictwo Naukowo-Techniczne, Warszawa 1998
[2] NOWAK E. (ed.), *Prognozowanie gospodarcze – metody, modele, zastosowania, przykłady*, Warszawa 1998
[3] DITTMANN P., *Metody prognozowania sprzedaży w przedsiębiorstwie*, Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław 1998
[4] HOLLAND H. J. *Algorytmy genetyczne*, Świat Nauki, 9/92
[5] RUTKOWSKA D., Piliński M., Rutkowski L., *Sieci neuronowe, algorytmy genetyczne i systemy rozmyte*, Wydawnictwo Naukowe PWN, Warszawa 1997
[6] SARIUSZ-WOLSKI Z., *Ilościowe metody zarządzania logistycznego w przedsiębiorstwie*, Toruńska Szkoła Zarządzania, Warszawa 1997
[7] http://en.wikipedia.org/wiki/Genetic_algorithm

[8] CHODAK G., KWAŚNICKI W., *Genetic algorithms in seasonal demand forecasting,* Information Systems Architecture and Technology '2000, Wrocław University of Technology, Wrocław 2000
[9] CHODAK G., ROPUSZYŃSKA-SURMA E., *Prognozowanie popytu w sklepie internetowym. Wyniki badań*, Gospodarka Materiałowa i Logistyka, 8/2008
[10] ELLISON, G., ELLISON, S.F., *Search, Obfuscation, and Price Elasticities on the Internet*, MIT, http://web.mit.edu/gellison/www/search.pdf, 2001