Applying Multi-objective Optimization for Variable Selection to Analyze User Trust in Electronic Banking

F. Liébana-Cabanillas[1], R. Nogueras[2], F. Muñoz-Leiva[3], I. Rojas[1], and A. Guillén[2]

[1] Former Director of Distribution Channels in Caja Rural Granada, Dpt. of Marketing and Market Research, University of Granada, Spain
[2] Dpt. of Computer Technology and Architecture, University of Granada, Spain
[3] Dpt. of Marketing and Market Research, University of Granada, Spain

Abstract. The potential fraud problems, international economic crisis and the crisis of confidence in markets have affected financial institutions, which have tried to maintain customer trust in many different ways. To maintain the trust level in financial institutions, the implementation of electronic banking for customers has been considered a successful strategy. However, the parameters that define user trust have not been analysed in detail due to the lack of experience and the recent use of e-banking. This paper aims to determine which variables are relevant to user trust by applying machine learning techniques as multi-objective genetic algorithms for the preparation of business strategies to improve confidence and profitability. The algorithms have been tuned following the indications given by experts and their results have been validated by them, setting a level of reliability. There is also a comparison among different fitness functions used in the evolution process that are able to rank the subset of variables encoded by the individuals.

1 Introduction: The Economic Crisis and Trust in the Financial Sector

The behaviour of the financial system against the economic crisis has been different among the countries within the European Union. While many international institutions focused their interest on credit and risk transfer, neglecting customer service, the banking sector continued to have an extensive network of offices through which to distribute financial products and to foster close client relationships. This very competitive environment forced banks to strictly control costs, which has made the financial system one of the world's most efficient [1]. Despite these advantages, the Spanish financial system was also in a precarious position particularly due to its exposure in real estate. In the latter part of the 90's and in the early part of the last decade there was an excess supply of real estate and therefore a large demand for financing. This situation forced financial institutions to go to wholesale markets since domestic markets did not have the resources to cover as much investment as was

J. Casillas et al. (Eds.): Management Intelligent Systems, AISC 171, pp. 143–152.
springerlink.com

being generated. Due to this and the pressing international crisis, the government and the Central Bank had to intervene different economies, among them, the Spanish.

The Spanish financial sector has already started to change as a result of this situation thanks to the Bank Restructuring Fund (FROB[1]), restructuring banks and strengthening the resources of credit institutions. In this complicated situation, the Spanish financial system has had to make technological improvements to reduce costs and optimize investments. Of all the available tools used to achieve these objectives, electronic banking has been the most widely implemented.

From the 90s to the present, electronic banking has become the distribution channel with the greatest potential for financial institutions [2]. Currently, the majority of companies offer their customers access to most of their services through this channel. Therefore, electronic banking has become a crucial service by which to gain customer satisfaction and loyalty and establish closer customer relationships, thereby meeting user expectations [3] and [4].

Thus, the primary alternative channel to the traditional bank branch is electronic banking as it has many advantages for customers including convenience, global access, availability, cost and time-savings, information transparency, choice and comparison, customization, and financial innovation [5] and [6]. However, this service also has some drawbacks, mainly related to trust and security. But trust, together with satisfaction, is considered one of the key elements in building long-term relationships, a fundamental business strategy in the current economic situation [7] and [8].

By the time of offering a client the e-banking services, it would be useful to know in advance if that client could be more attracted or not to the use of it. To do so, data mining and machine learning techniques can be applied to data bases in order to identify which attribute determine the use trust and if there is a model that could predict it. In this paper, these two phases are studied and the results have been validated using an expert committee. Therefore, the rest of the paper is structured as follows: Section 2 will describe the fieldwork and how the data was collected, Section 3 will introduced the algorithms and models that were applied as well as the novelties added to classical algorithms. Then, Section 4 will present the results and in Section 5, Conclusions will be drawn.

2 Perceived Trust in Electronic Banking and Fieldwork

The behaviour of users of electronic banking (Orange Foundation, 2010: 118) is characterized by prudence. To this end, users periodically review on-line bill movements, do security checks for electronic transactions and connections, avoid access from public computers, do not provide personal information by e-mail or phone, and log off bank web sites before closing browsers. However, some groups of customers are still reticent about such services. Regarding electronic commerce in general,

[1] The Bank Restructuring Fund was established by Royal Decree-Law 9 / 2009 of June 26, 2009.

consumers show more concern about the use of banking services strictly speaking, when the amount of money potentially exposed to fraud is significantly larger, than with other types of services or organizations.

Every year consumers lose an increasing amount of money through internet fraud. According to internet Fraud Watch (http://www.fraud.org) directed by the National Consumers League, consumers lost approximately 18.82 million dollars through fraud in 2010, significantly higher than the $5.79 million lost in 2004. On the average, the losses per person varied from $293 in 1999 to $2,165.15 in 2007. The Rivest, Shamir and Adleman (RSA, 2010) laboratories identified 281,000 phishing attacks in January 2010 aimed at financial institutions of any size.

In virtual environments trust is affected by the security and privacy problems [9] and [10]. However, it seems that these problems are being overcome because according to the Survey on Equipment and Information and Communication Technology Use in Households (National Institute of Statistics - NIE, 2011) in Spain, 72.4% of e-commerce transactions are made on a regular basis although users still shy away from using electronic banking for the reasons noted above.

2.1 Information Collection

The survey was conducted between September and October 2009. Participation in the survey was voluntary and was presented to the user once the authenticated party signed on to the website of a national saving bank in southern Spain.

The survey sample size was 1,081 completed questionnaires by individual visitors, but the final number of questionnaires used for this research was reduced to 946.

Those questionnaires completed by users with juridical personality were eliminated in order to only analyse the behaviour of individuals, or natural persons (see tables 1 and 2).

Table 1 Technical data. * For the estimation of a ratio, where P = Q = 0.5 and a confidence level of 95%, under the principles of simple random sampling

Population: Internet bank users.
Sample Frame: Users online Banking.
Type of Sampling: Simple random sampling.
Sample Size: 946 valid cases.
Sample Error*: 3,19%
Date of Field work: September and October 2009

The literature shows that the level of consumer trust toward a website depends on a number of factors, including the perceived reputation of the website [11]; site characteristics (web design, information availability, ease when navigating the site, privacy and security especially in those places where you can perform financial

Table 2 Respondent characteristics

Items	Data	Frecuency	Percent	Cumulative	(%) Cumulative
Gender	Male	634	67,02%	634	67,02%
	Female	312	32,98%	946	100,00%
Age	16-25	56	5,92%	56	5,92%
	26-35	324	34,25%	380	40,17%
	36-45	277	29,28%	657	69,45%
	46-65	259	27,38%	916	96,83%
	>65	30	3,17%	946	100%

transactions) [13]; and consumer characteristics [6]. In our research we analysed a total of 33 variables grouped into three clusters, socio-demographic, economic-financial and beliefs (trust).

In order to determine the relevance of each variable (see table 3), two criteria were selected: Delta Test and Mutual Information. The subsection below describes both of them.

Table 3 Variables Analyzed

Var. number	**Type**	**Variable**
1	Socio-demographic	Office
2		Geographic Region
23		Age
24		Gender
25		Mobile Telephone
26		E-mail
27		Zip code
28		Province
3	Economic- Financial	Profitability per Customer in 2010 per entity
4		Profitability per Customer in 2009 per entity
5		Average Liability Balance within Client Account 2010
6		Average Liability Balance within Client Account 2009
7		Average Liability Balance outside of Client Account in 2010
8		Average Liability Balance outside of Client Account in 2009
9		Average Balance of Client Assets in 2010
10		Average Balance of Client Assets in 2009
11		Number of products purchased in 2010
12		Number of products purchased in 2009
13		Linked products per client in 2010
14		Linked products per client in 2009
15		Business Volume per Client in 2010
16		Business Volume per Client in 2009
17		Customer profitability in 2010
18		Customer profitability in 2009
19		Direct Deposit for Paychecks
20		Direct Deposit for Pensions
21		Debit Card
22		Credit Card
28		Months of Experience with Ruralvía
29		Number of Operation on Ruralvia in 2010
30		Number of Operation on Ruralvia in 2009
31		Total Euro Amount of Operations on Ruralvia in 2010
32		Total Euro Amount of Operation on Ruralvia in 2009
	Behavioural	Trust

3 Models and Algorithms

This section formulates the problem tackled formally and will present the techniques applied to solve it. A new selection operator is discussed which improves the optimization results and the computational costs.

In order to identify which elements are the most important variables regarding the customer's trust and in order to design accurate models, a pre-processing of the data should be made by variable selection.

The problem of variable selection should be stated as: Given a set of N input/output pairs $(\mathbf{x}_i^j, y_i)$ where $i = 1...N, j = 1...d, \mathbf{x}_i^j \in R^d$ and $y_i \in R$ it is desired to obtain a subset of variables where the cardinality of it and the validation error are minimums chosen from a Pareto.

In order to solve this problem, Genetic Algorithms (GAs) where adapted in such a way that they keep a Pareto of non-dominated solutions defining a new class of GAs: Multi-objective GAs (MOGAs). Among the MOGAs, the NSGA-II (Non-dominated Sorting Genetic Algorithm) [14] [15] which is an updated version of the classical NSGA, i.e., multi-objective optimization algorithm from the field of Evolutionary Computation. The main goal of the NSGA-II algorithm is to find the best individuals of a population of candidate solutions according to the Pareto front by performing a sorting procedure that considers the different objectives to be optimised. NSGA-II has been succesfully applied to a wide variety of problems providing an excellent performance.

3.1 Multi-objective Optimization

Multi-objective optimization is the process of optimizing two or more objectives subject to certain constraints. A multi-objective optimization problem (MOP) has a set of n decision variables (x), a set of k objective functions $(y = f(x))$ and a set of m inequality constraints $(g(x))$ and a p equality constraints $(h(x))$, and objective functions and constraints depends on the n decision variables. More formally:

Optimize $y = f(x) = \{f_1(x), f_2(x), \dots, f_k(x)\}$ subject to $g(x) = \{g_1(x), g_2(x), \dots, g_m(x)\} \geq 0$ and $h(x) = \{h_1(x), h_2(x), \dots, h_p(x)\} = 0$

where $x = \{x_1, x_2, \dots, x_n\} \in X$ and $y = \{y_1, y_2, \dots, y_k\} \in Y$ and the decision vector is x, the decision space is X, the objective vector is y and the objective space is Y. We assume that a solution to this problem can be described in terms of a decision vector $(x_1, x_2, ..., x_n)$ in the decision space X. A function $f : X \rightarrow Y$ evaluates the quality of a specific solution by assigning it an objective vector $(y_1, y_2, ..., y_k)$ in the objective space Y. Therefore the problem consists in finding x with the best value for $f(x)$. The set of all decision vectors which satisfies the $m + p$ constraints is named Feasible Solution Set and denoted as X_f.

An decision vector x_1 is said to dominate another decision vectors x_2 $(x_1 < x_2)$ if no component of x_1 is greater (smaller) than the corresponding component of x_2 and at least one component is smaller (greater). This concept is known as Pareto dominance: $\forall i \in \{1, 2, \dots, k\}, f_i(x_1) \leq f_i(x_2) \wedge \exists j \in \{1, 2, \dots, k\} \mid f_j(x_1) < f_j(x_2)$

The set of all optimal solutions in the decision space X is in general denoted as the Pareto set $X^* \subseteq X$ and it is defined as: $P^* = \{x \in X_f \mid \neg\exists x\prime \in X_f \wedge x\prime \succ x\}$ and its image in objective space as Pareto front $Y^* = f(X^*) \subseteq Y_1$.

3.2 Multi-objective Selection Genetic Algorithm: MSGA

The NSGA-II has a good behaviour although it is quite expensive when measuring the computation time. As data bases become larger, this aspect should be kept in mind when choosing an algorithm. Another element that could be improved of this algorithm when applied to Variable Selection (VS) is to reduce the size of the Pareto, allowing to exploit more convenient solutions that include too many variables.

In order to avoid the cost of the non-dominated sort but keeping the MO aspect, a new selection operator within the GA has been defined. Selecting the binary tournament selection, one of the parents is selected considering the quality of the subset of variables and the other one is chosen considering the number of variables.

As in VS the solutions with a high number of variables are not desired (even if they provide the best optimization criterion), another operator has been introduced into the algorithm. On each generation, the individuals that have more than α variables are discarded. The α value should be selected manually considering the expert's opinion.

The results provided by a classical GA with these two elements ends up in better results and smaller computation times.

3.3 Delta Test

This method [19] is able to perform an estimation of the noise between input/output pairs, therefore, it is a good indicator of how precise a model can approximate a data set without overfitting these data. The application to the variable selection problem is quite straight forward: the solution is to find the subset of variables that provides the smallest value of the Delta Test (DT) [20].

The DT for a set of input vectors $X = \{\mathbf{x}_k\}$ and their output $Y = \{y_k\}$ with $k = 1...n$ is defined as:

$$\delta_{n,k} = \frac{1}{2n} \sum_{i=1}^{N} (y_i - y_{nn[i,k]})^2 \tag{1}$$

where $nn[i,k]$ is the index of the k-th nearest neighbour to x_i usually according to the Euclidean distance. Since $\delta_{n,1} \approx \sigma_e^2$, where σ_e^2 is the variance of the noise in the output, $\delta_{n,1}$ can be used as an estimation of the minimum mean squared error that can be obtained by a model without overfitting.

The main drawback of this methodology is its lack of robustness when the number of input samples is not large because the convergence to σ_e^2 is achieved when increasing n.

3.4 Mutual Information

The concept of Mutual Information (MI), also known as cross-entropy has been used already to solve the problem of selecting or identifying the most relevant variables from a set of input-output pairs, showing a good performance. Let $X = \{\mathbf{x}_k\}$ and $Y = \{y_k\}$ for $k = 1...n$, then the MI between X and Y can be understood as the amount of information that the subset of variables X provide over the output variable Y and its formulation is $I(X,Y) = H(Y) - H(Y|X)$ where $H(Y)$ is the entropy of Y and $H(Y|X)$ is the conditional entropy that measures the uncertainty of Y given a known X. Thus, we can obtain a numerical value that measures the relevance of X.

For the case where the variables are continuous, following Shannon formulation, the entropy can be defined as:

$$H(Y) = -\int \mu_Y(y) \log \mu_Y(y) dy, \tag{2}$$

where $\mu_Y(y)$ is the marginal density function. This function can be defined as the joint between the probability density functions of X and Y ($\mu_{X,Y}$), this is: $\mu_Y(y) = \int \mu_{X,Y}(x,y)dx$.

Therefore, once it is known the value of $H(Y)$, to obtain the MI value it is necessary to compute $H(Y|X)$, that, for the continuous case and reformulating it using the properties of the entropy, is defined as:

$$I(X,Y) = \int \mu_{X,Y}(x,y) \log \frac{\mu_{X,Y}(x,y)}{\mu_X(x)\mu_Y(y)} dxdy. \tag{3}$$

Then, to obtain the MI value it is only needed to estimate the joint probability density function (PDF) between X and Y. This value can be obtained using methods based on the k-NN [21] or in Parzen Window [22].

4 Experimental Results

Several experiments where carried out in order to find the best subset of variables that characterise client's trust. The classical NSGA-II was applied using the three criteria described in the previous section as well as the MSGA considering two values of α. The parameter setting for the GAs, that were using binary encoding, was:

- Population size: 50,100 and 150.
- Crossover: two-points ; Probability: 0.85
- Mutation: single gene level operator ; Probability: 0.1
- Selection: binary tournament for NSGA-II and the MO one for MSGA
- Stop criterion: no modification of the Pareto front for several iterations

4.1 Algorithms Comparison

Table 4 shows the results obtained after executing the three algorithms commented in the previous section using the three different criteria. The values represent the mean value of the Delta Test and the Mutual Information computed using k-NN and Parzen window provided by the best individual (using only one criterion) in the population after several runs.

Table 4 Results obtained for the three GAs implemented. The value represents the result obtained for each criterion and in the next column, the subset of variables that provide that result. In bold are the best results (DT lower is better, MI higher is better).

NSGA-II		
DT	3.1e-2 (5e-3)	5,10,12,14,15,19,21,22,23,25,29,30
MI-Parzen	1.06e-2 (1e-3)	11,20,27
MI-*k*NN	1.06e-2 (1e-3)	5,11,20,27
MSGA		
DT	3.27e-2 (1e-3)	2,4,5,11,14,19,21,22,23,25,27,28,31
MI-Parzen	**1.77e-2** (1e-3)	27
MI-*k*NN	**1.46e-2** (6e-3)	2,5,6,7,14,19,20,24,25,30
MSGA- α = 10		
DT	**2.92e-2** (6e-3)	5,10,12,14,15,19,21,22,23,25,29,30
MI-Parzen	1.12e-2 (1e-3)	4,27
MI-*k*NN	1.35e-2 (1e-3)	2,4,6,7,8,9,17,19,20,22,26

4.2 Expert's Validation

In order to validate the quality of the results obtained, a committee of five experts, with background in different financial entities, was consulted. The validation process consisted in four stages: personal interview, method and data evaluation, results evaluation, and feedback. The experts background was at least 10 years of experience and they have been working in comercial tasks in the last three. The ages were in the interval [34,46].

The interview had the aim of explaining the experts how the algorithms worked and the criteria they use to determine if a subset of variables is good or not. Afterwards, the results obtained in the experiments were given as input to the experts that had to evaluate in a Likert scale (1-7).

The results of the expert's opinions is shown in table 5. As this table shows, the best algorithm is the MSGA providing satisfactory results both using DT and MI using k-NN. This last criterion was the most valuable for the experts.

Table 5 Expert's punctuation

Algorithm	Method	Expert 1	Expert 1	Expert 1	Expert 1	Expert 1	Mean
NSGA-II	DT	3	4	4	5	4	4
	MI-Parzen	3	4	4	4	5	4
	MI-*k*NN	5	5	6	5	5	**5.2**
MSGA	DT	5	4	4	5	5	**4.6**
	MI-Parzen	1	1	1	1	1	1
	MI-*k*NN	6	4	4	5	4	**4.6**
MSGA α = 10	DT	5	5	6	5	6	**5.4**
	MI-Parzen	3	2	3	2	2	2.4
	MI-*k*NN	4	5	6	6	5	5.2

5 Conclusions and Implications for Management

As the studies show, e-banking seems to make a difference for the costumers to select a bank or to make them keep their savings in it. This paper aims to identify which characteristics define the customers' trust in order to improve the most common operations and to provide more information about e-baking to certain costumers.

Several multi-objective algorithms were evaluated, including a new modifciation that allows the algorithm to obtain accurate results and a reasonable number of variables. Furthermore, three criteria commonly used to perform variable selection were compared and all the results were evaluated and validated by experts in the field.

The results obtained by the algorithms and the opinion of the experts coincide in that the best multi-objective algorithm is the proposed MSGA and the best criteria to perform variable selection is mutual information computed using the *k*-NN algorithm.

References

1. Álvarez, J.M.: La banca española ante la actual crisis financiera. Estabilidad Financiera 15, 23–38 (2008)
2. Karjaluoto, H., Mattila, M., Pento, T.: Factors underlying attitude formation toward online banking in Finland. International Journal of Bank Marketing 20(6), 261–272 (2002)
3. Hsu, S.H.: Developing an index for online customer satisfaction: Adaptation of American Customer Satisfaction Index. Expert Systems with Applications 34, 3033–3042 (2008)
4. Berrocal, M.: Fidelización y Venta Cruzada. Informe Caja Castilla La Mancha (2009)
5. Delgado, J., Nieto, M.J.: Incorporación de la tecnología de la información a la actividad bancaria en España: La banca por Internet. Estabilidad financiera, Banco de España 3, 85–105 (2002)
6. Muñoz-Leiva, F.: La adopción de una innovación basada en la Web. Tesis Doctoral. Departamento de Comercialización e Investigación de Mercados, Universidad de Granada (2008)
7. Lam, S.Y., Shankar, V., Murthy, M.K.: Customer Value, Satisfaction, Loyalty, and Switching Costs: An Illustration from a Business-to-Business Service Context. Journal of the Academy of Marketing Science 32(3), 293–311 (2004)
8. García, N., Sanzo, M.J., Trespalacios, J.A.: Can a good organizational climate compensate for a lack of top management commitment to new product development? Journal of Business Research 61, 118–131 (2008)
9. Ha, H.Y.: Factors Influencing Consumer Perceptions of Brand Trust Online. Journal of Product and Brand Management 13(5), 329–342 (2004)
10. Laroche, M., Yang, Z., Mcdougall, G.H.G., Bergeron, J.: Internet Versus Bricks-and-Mortar Retailers: An Investigation Into Tangibility and Its Consequences. Journal of Retailing 81(4), 251–267 (2005)
11. Muñoz-Leiva, F., Luque-Martínez, T., Sanchez-Fernandez, J.: How to improve trust toward electronic banking. Online Information Review 34(6), 907–934 (2010)
12. Flavián, C., Guinalíu, M.: Un análisis de la influencia de la confianza y del riesgo percibido sobre la lealtad a un sitio web: el caso de la distribución de servicios gratuitos. Revista Europea de Dirección y Economía de la Empresa 16(1), 159–178 (2007)

13. Flavián, C., Guinalíu, M., Gurrea, R.:: Análisis empírico de la influencia ejercida por la usabilidad percibida, la satisfacción y la confianza del consumidor sobre la lealtad a un sitio web. In: XVI Encuentros de Profesores Universitarios de Marketing, pp. 209–226. Esic (2004)
14. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation 6(2) (2002)
15. Srinivas, N., Deb, K.: Multi-objective Optimization using Nondominated sorting in Genetic Algorithms. Evolutionary Computation 2(3), 221–248 (1994)
16. Eirola, E., Liitiainen, E., Lendasse, A., Corona, F., Verleysen, M.: Using the Delta Test for Variable Selection. In: ESANN 2008 Proceedings, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning, Bruges, Belgium (2008)
17. Gallant, S.I.: Perceptron-based learning algorithms. IEEE Transactions on Neural Networks 1(2), 179–191
18. Guillen, A., Sovilj, D., Lendasse, A., Mateo, F., Rojas, I.: Minimising the Delta Test for Variable Selection in Regression Problems. International Journal High Performance Systems Architecture 1(4) (2008)
19. Pi, H., Peterson, C.: Finding the Embedding Dimension and Variable Dependencies in Time Series. Neural Computation 6(3), 509–520 (1994)
20. Lendasse, A., Corona, F., Hao, J., Reyhani, N., Verleysen, M.: Determination of the Mahalanobis matrix using nonparametric noise estimations. In: ESANN, pp. 227–232 (2006)
21. Kraskov, A., Stögbauer, H., Grassberger, P.: Estimating mutual information. Physics Review (June 2004)
22. Kwak, N., Choi, C.H.: Input Feature Selection by Mutual Information Based on Parzen Window. IEEE Trans. Pattern Analysis and Machine Intelligence 24(12), 1667–1671 (2002)