Expert Systems with Applications 40 (2013) 5439–5447

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Analysing user trust in electronic banking using data mining methods

F. Liébana-Cabanillas [a], R. Nogueras [b], L.J. Herrera [c], A. Guillén [c,*]

[a] Department of Marketing and Market Research, University of Granada, Granada 18071, Spain
[b] MSc in Computer Science, University of Málaga, Málaga, Spain
[c] Department of Computer Technology and Architecture, CITIC-UGR, University of Granada, Granada 18071, Spain

ARTICLE INFO

Keywords:
Financial sector
Electronic banking
Trust
Variable selection
Multi-objective optimisation
MOEAs
NSGA-II
MOGA
Genetic algorithms

ABSTRACT

The potential fraud problems, international economic crisis and the crisis of trust in markets have affected financial institutions, which have tried to maintain customer trust in many different ways. To maintain these levels of trust they have been forced to make significant adjustments to economic structures, in efforts to recoup their investments and maintain the loyalty of their customers. To achieve these objectives, the implementation of electronic banking for customers has been considered a successful strategy. The use of electronic banking in Spain in the last decade has been fostered due to its many advantages, giving rise to real integration of channels in financial institutions. This paper reviews different methods and techniques to determine which variables could be the most important to financial institutions in order to predict the likely levels of trust among electronic banking users including socio-demographic, economic, financial and behavioural strategic variables that entities have in their databases. To do so, the most recent advances in machine learning and soft-computing have been used, including a new selection operator for multiobjective genetic algorithms. The results obtained by the algorithms were validated by an expert committee, ranking the quality of them. The new methodology proposed, obtained the best results in terms of optimisation as well as the highest punctuation given by the experts.

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 (spsacctoremoveAPAÁlvarez, 2008). 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 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 (Liébana-Cabanillas et al., 2011).

The Spanish financial sector has already started to change as a result of this situation thanks to the Bank Restructuring Fund (FROB[1]), and new regulations which will be introduced in 2013 with the advent of Basilea III[2] and more recently the Royal Decree for restructuring of the Spanish Savings Banks. According to the latest report on "Individual Financial Behaviour in Spain 2009" developed by Inmark (2009), 55.1% of the sample says their trust in the Spanish financial sector has worsened compared with 0.9% stating that it has improved and 40.1% who say there has been no change. 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.

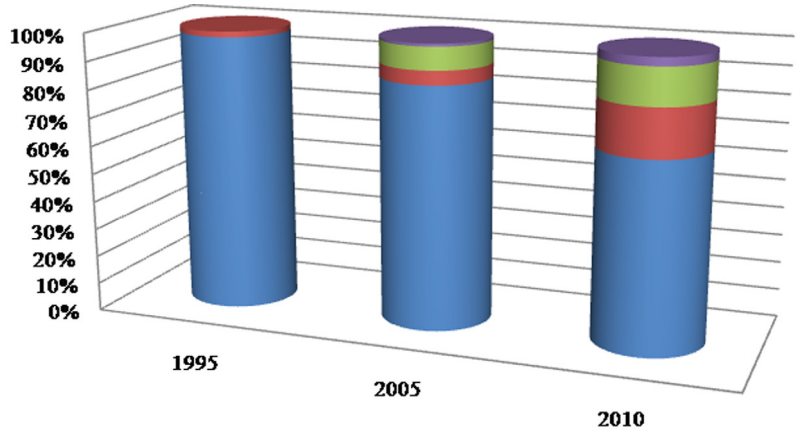
Traditionally, financial products and services have been distributed through bank branches due to their proximity to customers, the large number of services they perform, the added value that the client receives at the branch, and the important role bank branches play in decisions made by customers. In spite of this,

* Corresponding author.
E-mail address: aguillen@ugr.es (A. Guillén).

[1] The Bank Restructuring Fund was established by Royal Decree-Law 9/ 2009 of June 26, 2009, restructuring banks and strengthening the resources of credit institutions. The objective of this Fund is to manage bank restructuring processes and help strengthen their resources. The initial funding provided for this Fund is 9.000 million euros.

[2] Basilea III requires financial institutions to increase reserves to 7% of their of holding to be able to handle crisis situations.

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.03.010



	1995	2005	2010
ATM	0%	1%	3%
Telephone banking	10%	8%	13%
Electronic banking	10%	5%	17%
Bank branches	98%	86%	67%

Fig. 1. Percentage of use of the main channels for banking operations. Source: Word Retail Banking Report.

however, this conventional channel has begun to be replaced by a more agile and dynamic channel as reflected in the data of the World Retail Banking Report[3] (2010) on the percentage of use of the main channels (see Fig. 1).

From the 90s to the present, electronic banking has become the distribution channel with the greatest potential for financial institutions (Karjaluoto et al., 2002). 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 (Azcorra et al., 2001; Berrocal, 2009; Climent and Momparles, 2006; Hsu, 2008).

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 (Delgado and Nieto, 2002; Muñoz-Leiva, 2008). 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 (García et al., 2008b; Lam et al., 2004).

In this context, this paper reviews different methods and techniques to determine which variables could be the most important to financial institutions in order to predict the likely levels of trust among electronic banking users including socio-demographic, economic, financial and behavioural strategic variables that entities have in their database.

This paper is organized as follows. Section 2 describes the Electronic Banking at the European Union as well as the concept of costumer trust. Section 3 introduces the research methodology, the data description, and the models and algorithms used to analyse the data. Then, Section 4, presents the results of the analysis and its validation by a set of expert. Finally, Section 5, draws the conclusions from a business management perspective and further work.

[3] Available in http://www.capgemini.com.

2. Effect of the trust crisis in electronic banking and fraud problems

2.1. The electronic banking sector

A recent study by Orange Foundation (Fundación Orange, 2011) demonstrates the importance of the technological innovations taking place in the financial sector and how Internet banking continues to be one of the on-line tools most commonly used by the Spanish. According to the Orange Foundation the percentage of Spanish people using on-line banking compared to the percentage of people using Internet is growing is growing faster than the EU average as can be seen in the Fig. 2.

Despite this increase, Fig. 2 demonstrates that Internet penetration in general is still below the average for the European Union (EU), indicating there is still significant growth potential in the Spanish Internet banking sector for it to reach EU levels. An analysis of the evolution of e-banking compared to other uses of the Internet shows that it is used far less than other Internet services (Fundación Orange, 2011) (see Fig. 3) such as e-mail with an 85% rate of use in Spain and 89% on average in other EU countries in 2010; Internet searches (85% and 81% respectively in Spain and in the EU in 2009); and in downloading and reading newspapers (62% and 50% respectively). On the other hand, the percentage of users downloading software (33% and 31% respectively), doing job searches, making telephone calls and having video conferences (24,5% and 22% respectively) is higher in Spain than in the EU.

In addition to this data, the last study published in October 2010 by the company comScore[4], revealed that Spain has become the eighth country in the world in terms of e-banking penetration after Canada, Holland, France, Sweden, the United Kingdom, New Zealand and Belgium, beating the United States and Australia (Fig. 3).

This report differs from prior data (Fundación Orange, 2011) which placed the penetration rate in Spain in 42% and the EU average at 52%.

[4] http://www.comscoredatamine.com/2010/10/top-10-countries-by-online-bankingpenetration.

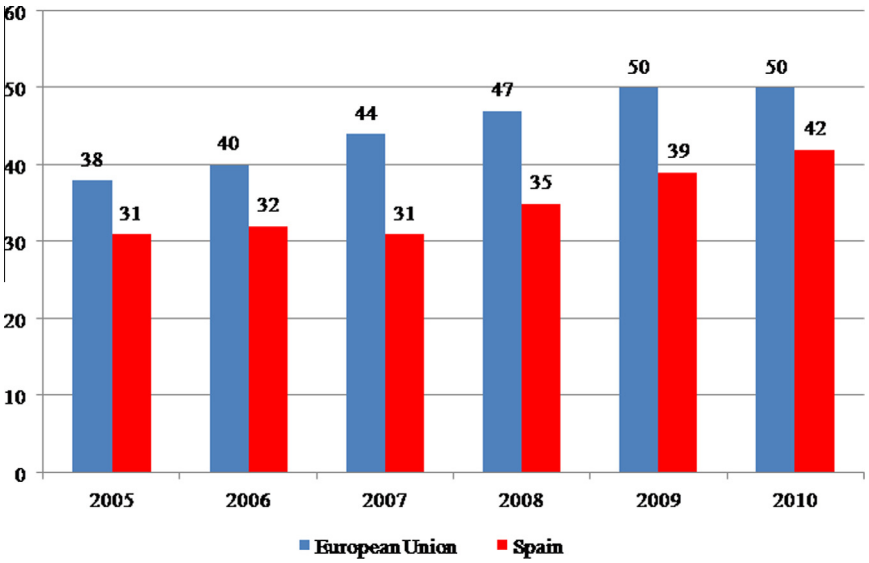


Fig. 2. Percentage of people using on-line banking in Spain and in Europe (Orange Foundation, 2011).

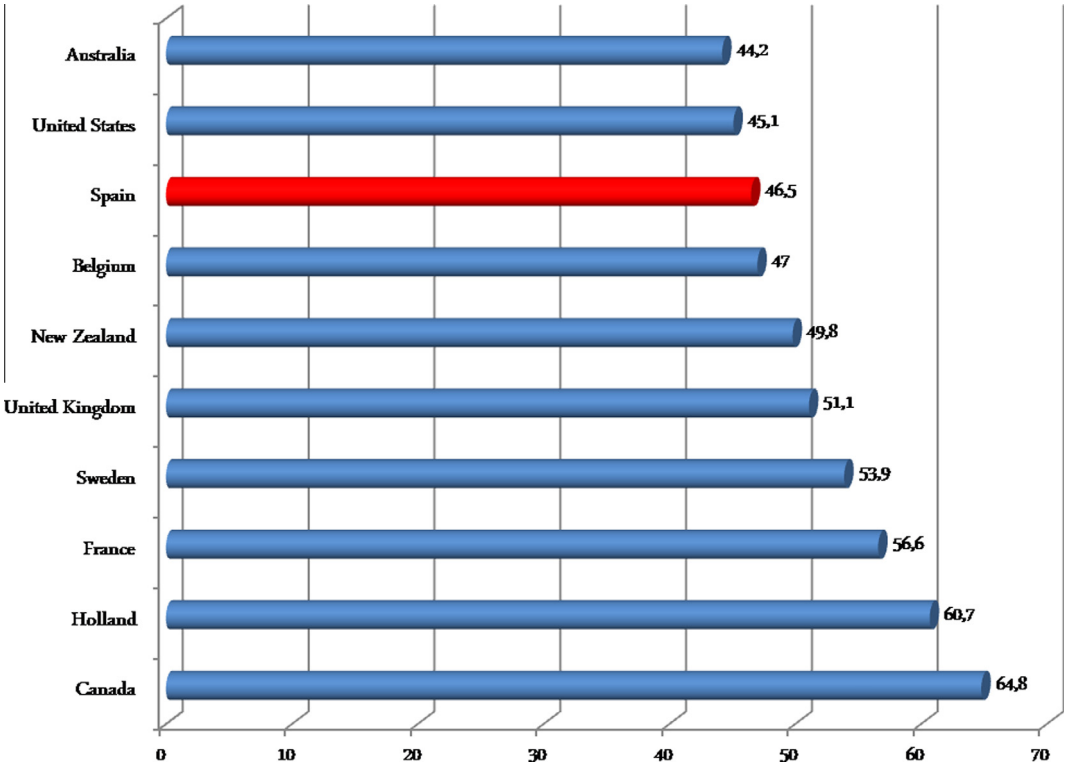


Fig. 3. Top ten contries in e-banking penetration. Source: comScore.

2.2. Perceived trust in electronic banking

The behaviour of users of electronic banking (Fundación Orange, 2011) 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, 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 (Milne and Boza, 1999).

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.

Therefore trust is one of the most influential variables on the buying behaviour via the Internet (Pavlou, 2003). Many studies have shown that a lack of trust in on-line relationships damage electronic commerce (Akhter et al., 2005; Bellman, 1999; Gefen, 2000; Hu et al., 2002; Lee and Wu, 2011; McKnight et al., 2002; Mcnight et al., 2004; Shankar et al., 2002; Tzortzatos and Boulianne, 2005; Wakefield and Whitten, 2006) and virtual environment (Wei et al., 2008). However, the figures for the number of transactions and the amount of currency moved through electronic banking in recent years, demonstrate the trust that customers have in this channel and consequently, their trust in a financial institution.

Trust is a construct widely studied in scientific literature with multiple dimensions such as cognitive and behavioural. (Dwyer et al., 1987) defined the cognitive perspective "as the belief that the word or the promise of a party is reliable and that it will meet its obligations in a relational exchange," while the behavioural perspective is defined as "the willingness of one party to be vulnerable to the actions of the other party based on the hope that the other will perform a particular action important to the trust, regardless of ability to monitor or control the other party" (Mayer et al., 1995).

We state that trust, in accordance with the principles of González et al. (2011)and Belanger et al. (2002), is the degree to which users of financial institution's electronic channels expect the institution to be honest, benevolent, competent and provide security measures to reduce the uncertainty and lack of user privacy.

The dimensions usually attributed to trust are (Doney and Cannon, 1997; Flavián and Guinalíu, 2007): integrity (adherence to generally accepted ethical principles), honesty (belief that the other party will fulfil its promises and obligations), benevolence (belief that the other party is interested in getting mutual benefits and will not make decisions or actions which cause damage to that trust) and competence (level of training a party should have to undertake specific tasks) (Doney and Cannon, 1997; Flavián and Guinalíu, 2007); all vital issues for financial institutions in order to maintain the trust of customers and users of electronic banking.

The latest research done in the last years have shown how trust is an element that helps to maintain the intention of use (Gu et al., 2009) in a variety of scenarios (Ho et al., 2010) as well as the commercial relationships. This last element is really important in the business field (Bigné and Blesa, 2003; García et al., 2008a).

In virtual environments trust is affected by the security and privacy problems (Ha, 2004; Laroche et al., 2005). However, it seems that these problems are being overcome because, according to the Survey on Equipment and Information and Communication Technology Use in Households (Instituto Nacional de Estadística, 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.

Therefore, we conclude that the trust has three important effects on business management (Segarra, 2007): (1) Reduction of perceived risk associated with opportunistic behaviour by the vendor, (2) a greater increase in self-trust, and (3) a reduction of transaction costs in commercial relationships.

3. Methodological approach

3.1. Fieldwork of study and 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 onto 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).

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 (Muñoz-Leiva et al., 2010); site characteristics (web design, information availability, ease when navigating the site, privacy and security especially in those places where you can perform financial transactions) (Flavián et al., 2004); and consumer characteristics (Muñoz-Leiva, 2008). In our research we analysed a total of 34 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.

3.2. Problem definition

This subsection describes the algorithms and paradigms that have been applied to identify the most relevant elements in costumer trust. This problem is known in the literature as variable selection and falls into de category of data pre-processing before building classification or prediction models.

The problem can be formally stated as: Given a set of N input/output pairs $\left(\vec{x}_i^j, y_i\right)$ where $i = 1 \ldots N, j = 1 \ldots d, \vec{x}_i^j \in R^d$ 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.

The process of variable selection (Guillén et al., 2008), as it is known nowadays consists in reducing the number of relevant variables to a smaller subset of variables and is very important in many multidimensional real world problems. The more variables are in the input vectors, the more number of them is required to sample properly the input vector space, increasing the number of computations when designing the model (Herrera et al., 2006).

Therefore, it is clear that this procedure must be applied although there are several approaches divided into *filter* (Guillén et al., 2008) and *wrapper* methods (Alberto et al., 2009). The first ones perform the variable selection before designing the regression model separating the two stages. The second ones, perform the variable selection during the setting of the parameters defined by each model. This last approach has two drawbacks:

- The variable selection performed is "ad hoc" to the model that is being designed, so the selection could not be representative for other regression methods.
- The number of models to be generated in order to explore the possible selections is large and makes this approach infeasible when the number of input vectors or dimensions is large.

Regarding the *filter* approaches, the main problem is the definition of a criterion that determines if a subset is adequate or not. Some authors design a complete model for each combination of variables, however, the problems faced with the *wrapper* methods arise again. Other authors (Eirola et al., 2008) propose the use of non-parametric methods that are much more efficient to evaluate, allowing the optimization algorithm to explore a large number of solutions. These non-parametric approaches have the property of being model independent so, the selection performed would be appropriate for any model selected afterwards.

As was commented before, it is desired to optimize the criterion that determines if a subset of variables is appropriate as well as reduce the number of variables in the solution so we face a multiobjective optimization problem that is the process of optimizing two or more objectives subject to certain constraints. A multiobjective 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

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

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

$$y = f(x) = \{f_1(x), f_2(x), \ldots, f_k(x)\} \tag{1}$$

Subject to

$$\begin{aligned} g(x) &= \{g_1(x), g_2(x), \ldots, g_m(x)\} \geqslant 0 \\ h(x) &= \{h_1(x), h_2(x), \ldots, h_p(x)\} = 0 \end{aligned} \tag{2}$$

where

$$\begin{aligned} x &= \{x_1, x_2, \ldots, x_n\} \in X \\ y &= \{y_1, y_2, \ldots, y_k\} \in Y \end{aligned} \tag{3}$$

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, $x_2, \ldots, x_n$) in the decision space X. A function $f: X \rightarrow Y$ evaluates the quality of a specific solution by assigning it an objective vector ($y_1, y_2, \ldots, y_k$) 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.

A decision vector x_1 is said to dominate another decision vectors x_2 ($x_1 \prec x_2$) 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.

$$\begin{aligned} &\forall i \in \{1, 2, \ldots, k\}, f_i(x_1) \leqslant f_i(x_2) \wedge \exists j \\ &\in \{1, 2, \ldots, k\} | f_j(x_1) < f_j(x_2) \end{aligned} \tag{4}$$

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:

$$P^* = \{x \in X_f | \neg \exists x' \in X_f \wedge x' \succ x\} \tag{5}$$

and its image in objective space as Pareto front $Y^* = f(X^*) \subseteq Y_1$.

In order to solve this problem, Genetic Algorithms (GAs) are a quite adequate technique due to its straight forward application. An adaptation of classical GAs was done in order to evolve several objectives so a Pareto of non-dominated solutions is generated, defining a new class of GAs: Multi-objective GAs (MOGAs). Among the MOGAs, the NSGA-II (Nondominated Sorting Genetic Algorithm) (Deb et al., 2000, 2002; Srinivas and Deb, 1994) which is an updated version of the classical NSGA. 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 successfully applied to a wide variety of problems providing an excellent performance.

In this context, this paper reviews different methods and techniques to determine which variables could be the most important to financial institutions in order to predict the likely levels of trust among electronic banking users including socio-demographic, economic, financial and behavioural strategic variables that entities have in their database.

3.3. Criteria to evaluate solutions

This subsection presents the different methods selected to determine if a subset of variables is adequate or not. The common characteristic is that they do not rely in the construction of a model that approximates the output, saving time and skipping the selection of the model and its parameters.

3.3.1. Delta test

This method (Pi and Peterson, 1994) 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) (Lendasse et al., 2006).

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

$$\delta_{n,k} = \frac{1}{2n} \sum_{i=1}^{N} (y_i - y_{nn[i,k]})^2 \tag{6}$$

where $nn[i,k]$ is the index of the kth nearest neighbour to x_i. The criterion to determine how close the input vectors can be selected from a wide variety, however, the most common is 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.

3.3.2. 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 $X = \{\vec{x}_k\}$ and $Y = \{y_k\}$ for $k = 1 \ldots n$, then the MI between X and Y can be understood as the amount of information that the subset of variables X provide

Table 2
Respondent characteristics.

Items	Data	Frequency	Percent (%)	Cumulative	(%) Cumulative
Gender	Male	634	6702	634	6702
	Female	312	3298	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

Table 3
Variables Analyzed.

Type	Variable	Var. number
Socio-demographic	Office	1
	Geographic region	2
	Age	23
	Gender	24
	Mobile telephone	25
	E-mail	26
	Zip code	27
	Province	28
Economic-financial	Profitability per customer in 2010 per entity	3
	Profitability per customer in 2009 per entity	4
	Average liability balance within client account 2010	5
	Average liability balance within client account 2009	6
	Average liability balance outside of client account in 2010	7
	Average liability balance outside of client account in 2009	8
	Average balance of client assets in 2010	9
	Average balance of client assets in 2009	10
	Number of products purchased in 2010	11
	Number of products purchased in 2009	12
	Linked products per client in 2010	13
	Linked products per client in 2009	14
	Business volume per client in 2010	15
	Business volume per client in 2009	16
	Customer profitability in 2010	17
	Customer profitability in 2009	18
	Direct deposit for paychecks	19
	Direct deposit for pensions	20
	Debit card	21
	Credit card	22
	Months of experience with Ruralva	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	33

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:

$$H(Y) = -\int \mu_Y(y)\log\mu_Y(y)dy, \tag{7}$$

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:

$$\mu_Y(y) = \int \mu_{X,Y}(x,y)dx. \tag{8}$$

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

$$-\int \mu_X(x)\int \mu_Y(y|X=x)\log\mu_Y(y|X=x)dydx. \tag{9}$$

Using the properties of the entropy, the MI can be reformulated as: $I(X,Y) = H(X) + H(Y) - H(X|Y)$ leading to:

$$I(X,Y) = \int \mu_{X,Y}(x,y)\log\frac{\mu_{X,Y}(x,y)}{\mu_X(x)\mu_Y(y)}dxdy. \tag{10}$$

Then, to obtain the MI value it is only needed to estimate the joint probability density function (PDF) between X and Y. Among the several methods to estimate it, below are the two most common: k-Nearest Neighbours and Parzen Window.

*3.3.2.1. Computing the MI using k-**NN**.* One of the methods to compute the PDF uses the k-Nearest Neighbours (k-NN) algorithms which has several advantages over other methods based on histograms and binnings (Kraskov, Stgbauer, & Grassberger, 2004; Stogbauer, Kraskov, Astakhov, & Grassberger, 2004). The algorithm proposed by the authors is available at http://www.klab.caltech.edu/~kraskov/MILCA/.

The only value that requires for its execution if the number of neighbours to be considered, k, however, the authors recommend to set the value upper to 6 so the experiments where carried out following these guidelines so $k = 20$.

3.3.2.2. Computing the MI using Parzen window. In Peng et al. (2005) a novel approach for feature selection is proposed considering both the relevance and the redundancy. The computation of the PDF ($\hat{p}(x)$) is based in Parzen window as described in (Kwak and Choi (2002)) so it has the following form:

$$\frac{1}{n}\sum i = 1N\delta(x - x_i, h), \tag{11}$$

where $\delta(.)$ is the Parzen window function and h is the window size. This window is usually chosen to be Gaussian (Kwak and Choi, 2002).

3.4. Multi-objective Selection Genetic Algorithm: MSGA

The NSGA-II has a good behaviour although it is quite expensive in terms of 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.

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

3.4.1. Selection operator

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. Starting from 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.

3.4.2. Crop operator

As in the problem of variable selection 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 number of variables are discarded. The α value should be selected manually considering the expert's opinion.

4. Experiments, results and discussion

The experiments consisted in 15 executions of each of the GAs and, afterwards, a tab search was applied to the bests solutions. The parameters of the GA were:

- population size: 100
- crossover: two-points
- crossover probability: 0.85
- mutation: simple gene mutation
- mutation probability: 0.1
- selection operator: Binary tournament/ new MO selection
- stop criterion: no modification of the Pareto front for several iterations

The results obtained are shown in Table 4.

From the computer science point of view, the results show how the modifications introduced by the MGA lead to more robust results as the standard deviation is smaller in all the cases. It is remarkable that the number of variables that optimises the solution is higher in the MGA than in the NSGA-II. Regarding to the optimization capacity, both algorithms perform adequately and, depending on the criteria, one behaves better than the other.

In order to validate the results, five experts in the subject with experience in different national and international companies were consulted. All of them have over 10 years of experience and the last three they have been involved in commercial tasks. The expert's profile is shown in Table 5.

The validation process was divided into three stages through the first semester of 2012:personal interview with each one, evaluation of the methods and evaluation of the results. In the personal interview, the experiments were explained and a first selection of variables was performed. Afterwards, the subsets of variables proposed by the algorithms were presented to the experts and they evaluate them in a Likert scale (1 to 7). The results are shown in Table 6. The last stage consisted in the evaluation of the results of the methods applied with the perspective of business management.

Considering only the comparison between the two algorithms (MGA vs NSGA-II), the experts support with higher ranks the solutions provided by the second algorithm for all the criteria. Showing that, for certain applications, to try to obtain the whole Pareto might not be the right strategy even though when this problem is multi-objective.

In a second interview, the experts gave the reasons for the punctuations. Regarding the NSGA-II, the main problem is that it provides a very reduced number of variables that would make quite difficult for the expert to apply them in real life. Furthermore, the nature of the selected variables does not seem to offer a high relevance for management purposes. On the other hand, the MGA provides more complete solutions that include more significant variables so better decisions can be taken from a management perspective.

To summarize, the best method to perform variable selection using the expert's criterion is the MGA using Mutual Information computed with the *k*-NN algorithm. The variables it selects are more reliable than the other methods from a management perspective in order to improve the electronic banking.

The second method chosen was MGA using Delta Test that offers a subset of significant variables although the number of those is smaller, making more difficult the decision support as experts say.

The worst method evaluated was Mutual Information computed using Parzen window independently of the GA used to optimize it. As it selects only one variable, from the management point of view, it does not provide enough information to take management decisions.

Table 4
Results obtained by the two MOGAs after applying the Tab search. Mean values and standard deviations (in brackets) of the different criteria. DT is of type *lower-is-better* MI is *higher-is-better*. The variables that appeared in the solution more than the 50% of the executions is shown in the last column.

Criterion	Mean (std)	Variables (> 50%)
NSGA-II		
DT	0.0364(0.0045)	14,21
Parzen	0.0117(1.8E−18)	27
Milca	0.0356(0.0110)	20
MGA (α = 10)		
DT	0.0315(1.4E−17)	2,12,23,30
Parzen	0.0117(1.8E−18)	27
Milca	0.0297(7.1E−18)	2,3,14,17,19,22,27,32

Table 5
Expert's profiles.

Expert #	Age	Position	Years of experience
1	34	Office Director	10
2	46	Office Director	16
3	34	Office Director	11
4	38	Office Director	14
5	36	Commercial Department Director	13

Table 6
Expert's ranking of the MOGAs results.

Criterion	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Mean
NSGA-II						
DT	4	4	2	3	3	**3.2**
Parzen	1	2	1	1	1	1.2
Milca	3	3	1	2	2	2.2
MGA (α = 10)						
DT	5	4	3	4	4	4
Parzen	1	2	1	1	1	1.2
Milca	7	6	6	7	6	**6.4**

5. Conclusions, limits and further research

During the last decade the Spanish financial system has undergone a profound transformation with the aim of reducing the bank usage rate in order to save costs. One of the most important changes in the sector has been the revolution of electronic banking among customers although the reduction was not as high as expected. Nonetheless, electronic banking currently penetrates close

to 50% of Internet users and 20% of the total population according to different sources, indicating significant societal acceptance of this service. This has forced banks to make major investments to maintain and enhance user satisfaction with this channel that is changing the traditional concept of financial institutions. The large number of advantages and few drawbacks electronic banking offers demonstrates its importance.

Considering this, the work presented aims to shed some light in this area by using several machine learning methods over a data set that was collected by a financial entity. The data consisted on information about the clients, the offices and an evaluation of the electronic banking platform performed by the costumers. The goal was to identify which variables could influence more costumer's trust.

As the variable selection problem has been studied previously in the literature, multi-objective genetic algorithms were applied using several fitness functions: Test Delta and Mutual Information. Two new operators were defined in order to obtain solutions more significant and valuable from the expert's point of view, which lead to better results. From the analysis it is interesting how Mutual Information obtains higher scores, indicating that the Test Delta might not be such a good indicator since it tends to select the less noisy variable instead of the most significant one.

Therefore, to apply this new techniques and algorithms in the process of decision taken helps to identify which variables could be the most important ones and focus on improving those to increase costumer's trust in electronic banking. The process is feasible due to the large amount of information that financial companies have about their clients.

In recent years the financial sector has carried out customer loyalty campaigns at different levels. On one hand, management systems such as CRM, together with general communication campaigns about the activities of the entity have increased the involvement of customers through the branches. On the other hand, most entities link credit concessions and interest rates extra for customer deposits to other products favouring up-selling strategies (cross-selling, reducing product and service cancellations and improving client relationships). For these reasons it is logical that three bidirectional goals are accomplished with this strategy; increased cross sales, increased asset balance and therefore the customer's gross profitability for the entity. For these reasons proper segmentation of customers is vital to optimal management of electronic banking services. When an organization knows different customer profiles and the variables that define them, it can anticipate needs and achieve increased profitability and improved levels of customer trust leading to greater brand loyalty. Finally, some future work goes through the analysis of regression models and how they behave using the solutions provided by the variable selection algorithms as well as the study of the expert's opinion regrading the interpretability of those models.

References

Akhter, F., Hobbs, D., & Maamar, Z. (2005). Using fuzzy cognitive time maps for modeling and evaluating trust dynamics in the virtual enterprises. *Expert Systems with Applications, 28*, 623–628.

Alberto, Guillén, Héctor, Pomares, Jesús, González, Ignacio, Rojas, Olga, Valenzuela, & Beatriz, Prieto (2009). Parallel multiobjective memetic RBFNNs design and feature selection for function approximation problems. *Neurocomputing, 72*(16–18), 3541–3555.

Álvarez, J. M. (2008). La banca española ante la actual crisis financiera. *Estabilidad Financiera, 15*(November), 23–38.

Azcorra, A., Bernardos, C. J., Gallego, O., & Soto, I. (2001). Informe sobre el estado de la teleeducación en España, Asociación de Usuarios de Internet.

Belanger, F., Hiller, J. S., & Smith, W. J. (2002). Trustworthiness in electronic commerce: The role of privacy, security, and site attributes. *Journal of Strategic Information Systems, 11*, 245–270.

Bellman, S. G. L. (1999). Predictors of online buying behavior. *Communications of the ACM, 42*(12), 32–38.

Berrocal, M. (2009). Fidelización y venta cruzada, Informe Caja Castilla La Mancha.

Bigné, J. E., & Blesa, A. (2003). Market orientation, trust and satisfaction in dyadic relationships: A manufacturer-retailer analysis. *International Journal of Retail & Distribution Management, 31*(11), 574–590.

Climent, F., & Momparles, A. (2006). La situación de la banca on line en españa, Boletín Económico de ICE 2898, del 4 al 10 de Diciembre (pp. 27–49).

Deb, K., Agarwal, S., Pratap, A., & Meyarivan, T. (2000). A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. *Parallel Problem Solving from Nature PPSN VI, 1917*, 849–858.

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation, 6*(2).

Delgado, J., & Nieto, M. J. (2002). Incorporación de la tecnología de la información a la actividad bancaria en España: La banca por Internet, Estabilidad financiera, Banco de Espaa, 3 (noviembre), 85–105.

Doney, P. M., & Cannon, J. P. (1997). An examination of the nature of trust in buyer–seller relationships. *Journal of Marketing, 61*(April), 35–51.

Dwyer, F. R., Schurr, P. H., & Oh, S. (1987). Developing buyer–seller relationships. *Journal of Marketing, 5*(2), 11–27.

Eirola, E., Liitiainen, E., Lendasse, A., Corona, F., & Verleysen, M. (2008). Using the delta test for variable selection. In *Proceedings, european symposium on artificial neural networks-advances in computational intelligence and learning, ESANN 2008*. Belgium: Bruges.

Flavián, C., Guinalíu, M., & Gurrea, R. (2004). Análisis empírico de la influencia ejercida por la usabilidad percibida, la satisfacción y la confianza del consumidor sobre la lealtad a un sitio web. XVI Encuentros de Profesores Universitarios de Marketing (pp. 209–226). Madrid: Esic.

Flavián, C., & Guinalíu, M. (2007). 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.

Fundación Orange. (2011). eEspaña 2011: Informe anual sobre el desarrollo de la sociedad de la información en España, disponible en. <http://www.fundacionorange.es>.

García, N., Santos, M. L., Sanzo, M. J., & Trespalacios, J. A. (2008a). El papel del marketing interno como antecedente de la capacidad de innovación de la PYME. Efecto sobre los resultados empresariales. XXII Congreso anual AEDEM. Salamanca, 18, 19 y 20 de Junio.

García, N., Sanzo, M. J., & Trespalacios, J. A. (2008b). Can a good organizational climate compensate for a lack of top management commitment to new product development? *Journal of Business Research, 61*, 118–131.

Gefen, D. (2000). E-commerce: The role of familiarity and trust. *The International Journal of Management Science, 28*, 725–737.

González, E., López, M. J., & Lampón, J. F. (2011). La confianza del consumidor como factor clave en la construcción de lealtad en ambientes electrónicos, XXI Jornadas Hispano Lusas de Gestión Científica, 2–4 de Febrero, Córdoba.

Guillén, A., Sovilj, D., Lendasse, A., Mateo, F., & Rojas, I. (2008). Minimising the delta test for variable selection in regression problems. *International Journal High Performance Systems Architecture, 1*(4).

Gu, J. C., Lee, S. C., & Suh, Y. H. (2009). Determinants of behavioral intention to mobile banking. *Expert Systems with Applications, 36*, 11605–11616.

Ha, H. Y. (2004). Factors influencing consumer perceptions of brand trust online. *Journal of Product and Brand Management, 13*(5), 329–342.

Herrera, L. J., Pomares, H., Rojas, I., Verleysen, M., & Guillen, A. (2006). Effective input variable selection for function approximation. *Lecture Notes in Computer Science, 4131*, 41–50.

Ho, L., Kuo, T., Lin, C., & Lin, B. (2010). The mediate effect of trust on organizational online knowledge sharing: An empirical study. *International Journal of Information Technology & Decision Making, 9*(4), 625644.

Hsu, S. H. (2008). Developing an index for online customer satisfaction: Adaptation of american customer satisfaction index. *Expert Systems with Applications, 34*, 3033–3042.

Hu, X., Lin, Z., & Zhang, H. (2002). Trust promoting seals in electronic markets: An exploratory study of their effectiveness for online sales promotion. *Journal of Promotion Management, 9*(1–2), 163–180.

Inmark. (2009). Comportamiento Financiero de los Particulares España.

Instituto Nacional de Estadística. (2011). Encuesta sobre Equipamiento y Uso de Tecnologías de la Información y Comunicación en los hogares.

Karjaluoto, H., Mattila, M., & Pento, T. (2002). Factors underlying attitude formation toward online banking in Finland. *International Journal of Bank Marketing, 20*(6), 261–272.

Kraskov, A., Stgbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physics Review*.

Kwak, N., & Choi, C. H. (2002). Input feature selection by mutual information based on Parzen window. *IEEE Transaction on Pattern Analysis and Machine Intelligence, 24*(12), 1667–1671.

Lam, S. Y., Shankar, V., & Murthy, M. K. (2004). Customer value, satisfaction, loyalty, and switching costs: An illustration from a business-to-business service context. *Journal of the Academy of Marketing Science, 32*(N3), 293–311.

Laroche, M., Yang, Z., Mcdougall, G. H. G., & Bergeron, J. (2005). Internet versus bricks-and-mortar retailers: An investigation into tangibility and its consequences. *Journal of Retailing, 81*(4), 251–267.

Lee, F. H., & Wu, W. Y. (2011). Moderating effects of technology acceptance perspectives on e-service quality formation: Evidence from airline websites in Taiwan. *Expert Systems with Applications, 38*, 7766–7773.

Lendasse, A., Corona, F., Hao, J., Reyhani, N., & Verleysen, M. (2006). Determination of the Mahalanobis matrix using nonparametric noise estimations. In *ESANN* (pp. 227–232).

Liébana-Cabanillas, F., Martínez-Fiestas, M., & Rejón-Guardia, F. (2011). The economic crisis in the European union: The confidence in the Spanish financial sector. In *Workshop: Crisis, Lisbon, EU Policies and member states*, 30 y 31 de Mayo, Granada.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review* (20), 709–734.

McKnight, H., Choudhury, V., & Kacmar, C. (2002). The impact of initial consumer trust on intentions to transact with a web site: A trust building model. *Journal of Strategic Information Systems, 11*, 297–323.

Mcnight, D. H., Kacmar, C. J., & Chodhury, V. (2004). Shifting factors and the ineffectiveness of third party assurance seals: A two-stage model of initial trust en a web business. *Electronic Markets, 14*(3), 252–266.

Milne, G. R., & Boza, M. E. (1999). Trust and concern in consumers' perceptions of marketing information management practices. *Journal of Interactive Marketing, 13*(1), 5–24.

Muñoz-Leiva, F. (2008). La adopción de una innovación basada en la Web. Tesis Doctoral. Departamento de Comercialización e Investigación de Mercados, Universidad de Granada.

Muñoz-Leiva, F., Luque-Martínez, T., & Sánchez-Fernández, J. (2010). How to improve trust toward electronic banking. *Online Information Review, 34*(6), 907–934.

Pavlou, P. A. (2003). Consumer acceptance of electronic commerce: Integrating trust and risk with the technology acceptance model. *International Journal of Electronic Commerce, 7*(3), 69–103.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(8), 1226–1238.

Pi, H., & Peterson, C. (1994). Finding the embedding dimension and variable dependencies in time series. *Neural Computation, 6*(3), 509–520.

Segarra, P. (2007). Influencia de la heterogeneidad del mercado en la intención de comportamiento del consumidor: Respuestas a la actividad relacional en la distribución de gran consumo. Tesis Doctoral. Departamento de Gestión de Empresas. Universidad Rovira i Virgili.

Shankar, V., Urban, G., & Sultan, F. (2002). Online trust: A stakeholder perspective, concepts, implications, and future directions. *The Journal of Strategic Information Systems, 11*(34), 325–344.

Srinivas, N., & Deb, K. (1994). Multi-objective optimization using nondominated sorting in genetic algorithms. *Evolutionary Computation, 2*(3), 221–248.

Stogbauer, H., Kraskov, A., Astakhov, S. A., & Grassberger, P. (2004). Least-dependent-component analysis based on mutual information. *Physical Review E, 70*(6), 066123.

Tzortzatos, R., & Boulianne, E. (2005). Assurance seals on Web sites aren't foolproof. *Bank Technology News, 18*(7), 44.

Wakefield, R. L., & Whitten, D. (2006). Examining user perceptions of third-party organization credibility and trust in e-retailer. *Journal of Organizational and End User Computing, 18*(2), 1–19.

Wei, Z., Lu, L., & Yanchun, Z. (2008). A fuzzy logic-based system for assessing the level of business-to-consumer (B2C) trust in electronic commerce. *Expert Systems with Applications, 35*, 1583–1592.