Application of evolutionary computation techniques for the identification of innovators in open innovation communities

M.R. Martínez-Torres *

Facultad de Turismo y Finanzas, University of Seville, Avda. San Francisco Javier s/n, 41018 Sevilla, Spain

ARTICLE INFO

Keywords:
Open innovation
Innovation communities
Evolutionary computation
Social network analysis

ABSTRACT

Open innovation represents an emergent paradigm by which organizations make use of internal and external resources to drive their innovation processes. The growth of information and communication technologies has facilitated a direct contact with customers and users, which can be organized as open innovation communities through Internet. The main drawback of this scheme is the huge amount of information generated by users, which can negatively affect the correct identification of potentially applicable ideas. This paper proposes the use of evolutionary computation techniques for the identification of innovators, that is, those users with the ability of generating attractive and applicable ideas for the organization. For this purpose, several characteristics related to the participation activity of users though open innovation communities have been collected and combined in the form of discriminant functions to maximize their correct classification. The right classification of innovators can be used to improve the ideas evaluation process carried out by the organization innovation team. Besides, obtained results can also be used to test lead user theory and to measure to what extent lead users are aligned with the organization strategic innovation policies.

1. Introduction

The concept of open innovation, launched by Chesbrough (2003) and others, has become increasingly popular among scholars and industry practitioners since the term was coined. Open innovation refers to the use of external sources and actors to achieve innovation, and it is based on the idea that companies should not just rely on internally developed ideas and knowledge, but increasingly also on ideas and knowledge developed externally (Chesbrough, Vanhaverbeke, & West, 2006; Tödtling, Prud'homme van Reine, & Dörhöfer, 2011). It assumes that useful knowledge is widely diffused and abundant. There is, for example, a growing availability of knowledge from multiple innovation actors, including universities, specialized suppliers, inventors and knowledge brokers. In this conditions, the "old" model of closed innovation where innovation processes are controlled by the company needs to be changed in favor of the detection and assimilation of externally developed knowledge (Barge-Gil, 2010; De Jong, Kalvet, & Vanhaverbeke, 2010; Poetz & Schreier, 2012). Previous studies agree that open innovation is not a general phenomena and depends on certain company characteristics as well as external conditions. Chesbrough (2003) identified various external factors that explain why enterprises increasingly adopt the open paradigm. The availability of a strong public knowledge base, a mobile and educated working population or the availability of ample external finance for innovation are the three conditions enabling open innovation to emerge.

From the viewpoint of the organization, there are various mechanisms and channels used for sourcing and acquiring external knowledge such as the absorption of local knowledge spillovers, collaboration in R&D and innovation with firms and universities, relations to spin-off companies, informal knowledge interactions, customer contributions through design toolkits or idea competitions (Keeble & Wilkinson, 2000; Schwab, Koch, Flachskampf, & Isenhardt, 2011; Tödtling, Lehner, & Trippl, 2006; von Hippel & Katz, 2002). The strategic challenge is how firms can best organize the sourcing, codification and exploitation of the internal and external knowledge and informational resources to maximize and sustain innovation (Love & Roper, 2009). One of the most popular mechanism for open innovation implementation is user innovation communities (Dahlander, Frederiksen, & Rullani, 2008). Firms such as Microsoft, Dell, IBM, BMW, and Nokia increasingly invest in virtual communities to solicit user contributions as part of their innovation processes. This trend is explained by the increase in digitalization and the decrease in the costs of communication that have lead to an exponential growth of user innovation platforms (Mahr & Lievens, 2012). Internet have facilitated the accessibility of these platforms by users geographically distributed all over the world. However, this accessibility is also causing the

* Tel.: +34 954 55 43 10; fax: +34 954 55 69 89.
E-mail address: rmtorres@us.es

generation of a huge amount of information which it is difficult to process and evaluate by the innovation departments or experts within organizations. Posted ideas must be evaluated one by one by the innovation department or even some specific experts of the organization, and this evaluation consists of reading the idea, assessing its applicability attending to the strategic innovation policies of the organization and planning their possible implementation in case they are finally accepted. The problem is that online user innovation communities can generate hundreds or even thousands of solutions in a short period of time, saturating the capacity of internal evaluators and hiding the really attractive innovations. That is the reason why online user innovation communities typically include some type of scoring systems so that the community can evaluate potential solutions. This scoring scheme is based on the idea that the crowd can do a better evaluation than individuals since they own a group-based intelligence which can outperform individual knowledge (Surowiecki, 2004).

However, strategic innovation policies of organizations are not always aligned with users' desires. Some non affordable ideas can be excellent for users but prohibitive for the company, and these ideas would probably receive a high score by other community users. In this sense, it is much more useful for the company the identification of users posting ideas that will be finally adopted. This information can be easily collected from innovation communities websites as they usually inform users about the status of their posted ideas. The purpose of this paper consists of the identification of these innovators, defined as those users generating ideas that will be finally adopted by the company. The condition of being innovator or non innovator is a dichotomic property of each user. Therefore, the identification of innovators is a classification problem that can be solved using a discriminant function over a set of variables characterizing the activity and behavior of users within the community, which on the other hand is the main available information. The main problem associated to this identification is that the considered dependent variable contains a high number of zeros (non-innovators), leading to the so called zero inflated problem. To solve this issue, a optimization procedure consisting of finding the values of the variables coefficients so that the discriminant function can maximize the percentage of correct classification of innovators and non innovators is formulated. Three different evolutionary computation techniques are used to solve the problem for evaluating the reliability of results. Additionally, a bootstrapping technique has also been implemented to obtain the confidence intervals of the resulting coefficients.

The rest of the paper is structured as follows. Section 2 details previous works related to the open innovation paradigm and the identification of users with special profiles. Section 3 describes the formulation of the problem in the form of an optimization problem and presents the three proposed evolutionary computation techniques: simulated annealing, particle swarm optimization and genetic algorithms. The three algorithms are then applied to the case study of IdeaStorm website, which is introduced in Section 4, as well as those variables measuring the activity and behavior of users within this innovation community. Obtained results are discussed in Section 5. Finally, conclusions are provided in Section 6.

2. Related work

Online user innovation communities make use of Internet as the prime communication channel, allowing company-to-customer as well as customer-to-customer communications (Di Gangi & Wasko, 2009; Rohrbeck, Steinhoff, & Perder, 2008). This strategy assumes that new product developments require interactions among like-minded customers who talk about their usage experiences, raise questions, present solutions, and offer answers (Fueller & Matzler, 2007). These interactions enable users to build on one another's knowledge and experiences, which plays a critical role in developing ideas (Rowley, Kupiec-Teahan, & Leeming, 2007). Previous research on open innovation communities have been mainly focused on their operational level. The first studies in this line discussed the characters of user innovation community based on the example of open-source software projects, which was a relatively well-developed and very successful form of internet-based innovation community (Martínez-Torres, 2012; Von Hippel, 2001; West & O'mahony, 2008). Later, Von Hippel and Von Krogh (2003) proved that user innovation communities illustrate a "private-collective" model of innovation incentive. However, there are some differences between open source software communities and open innovation communities. The most important one is that open source communities works based on user requests for information or help, while open innovation communities are more impersonal and users share information with others but not responding to any specific request of information. The effectiveness of open innovation is another interesting issue treated in the previous literature. Laursen and Salter (2006) found a non linear relationship between open innovation and performance, concluding that too much open innovation hurts organization performance. More specifically, they found that innovative performance is curvilinearly related through an inverted U-shape to the number of sources. The size is explained because firms gain innovative opportunities as they implement a wider and deeper search over a huge number of sources. However, innovation search is not costless and can be time consuming, expensive, and laborious. In the case of online innovation communities, the most important cost is the one associated to evaluation of posted ideas. Although the marginal evaluation cost of each idea is low, the cumulative evaluation cost when thousands of ideas are posted can be tremendous. Community based voting methodologies like simple discussion forums, community ratings (Carbone, Contreras, Hernández, & Gomez-Perez, 2011; Frey, Lüthje, & Haag, 2011) or more complex methodologies like prediction markets (Blohm, Riedl, Leimeister, & Krcmar, 2011; Spann & Skiera, 2009) that are based on stock-market trading algorithms can help to solve this challenge. Any assessment system based on a community scoring model, in which ideas can be awarded with a specific number of points, can help selecting the best idea (Hüsig & Kohn, 2011). However, obtained results through these procedures may be contradictory with the innovation strategic policies of organizations. Some of the top ranked ideas may be in the opposite direction of organization priorities or their implementation costs can be prohibitive. An alternative to the scoring model is the identification of best ideas through the identification of a particular subset of users called lead users. Lead users are characterized because they anticipate early on innovative characteristics, which are relevant only much later for other customers (Von Hippel, 1986, 1988). Additionally, lead users have the ability to develop a fully functional solution for their needs (Mahr & Lievens, 2012; Morrison, Roberts, & Midgley, 2004). They hence possess not only need information, but equally also solution information. Previous research about lead user have been focused on issues like their identification (Urban & von Hippel, 1988). The behavior of lead users have been described in the literature by several characteristics. For instance, their ability to bear innovative solutions is fundamentally linked to a person's individual creativity (Amabile, Barsade, Mueller, & Staw, 2005). They also make regular contributions to the community as they are actively engaged in problem-solving. However, some authors criticizes that their behavior is biased by their interest in obtaining a benefit from their proposed solutions. For instance, Berthon, Pitt, McCarthy, and Kates (2007) consider the idea of creative users as opposed to lead users. Creative users do not necessarily face needs that will become general as lead

users do. They innovate as an exercise of creativity and not to solve some specific need. Besides, creative users don't need to benefit directly from their innovations, although they may obviously benefit indirectly through thanks, peer recognition, and so forth. In summary, creative users constitute a more wider category than lead users, although obviously there is an overlap between both groups. This paper follows a similar approach to creative users and considers a wider scope for innovators definition using those users whose ideas have been implemented by the company. They can be identified using the available information through open innovation websites, which basically is information related to their participation activity and interactions with other users.

3. Methodology: Evolutionary computation techniques

The dependent variable in this study is the condition of being an innovator. Typically, open innovation communities generate a huge number of ideas but only a small fraction of them are finally adopted by the company. As a result, a high proportion of non-innovators is prevalent in a representative sample. Such high proportion of zeros in the dependent variable leads to zero-inflated problems. The application of regression methods or discriminant analyses with a disproportionately high number of zeros in the dependent variable may result in biased parameter estimates and misleading inferences (Lee, Wang, Scott, Yau, & McLachlan, 2006). To address this problem, this section propose the search of discriminant rules for optimizing the classification problem using three different evolutionary computation techniques.

3.1. Formulation of the problem

Mathematically, the proposed optimization problem can be formulated as follows. Estimated innovators (*Innovators**) represent a dichotomous variable that is defined in terms of the activity and behavior of users within the community.

$$Innovators^* = \begin{cases} 1 & \text{if} \quad \Pi_{i=1}^{n}\theta_i \quad Var_i + C > 0 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where Var_i are the set of features (variables) characterizing the behavior of users, θ_i are the set of coefficients for each variable and C is a constant term. If *Innovators* represent the vector of users whose ideas have been actually implemented according to the available information, the optimization problem consists of selecting a set of θ_i values so that the confusion matrix between *Innovators** and *Innovators* maximize the identification ratios.

Table 1 details the confusion matrix. TP and TN are true positives and negatives, respectively, and they refer to those innovators/non innovators that were correctly classified while FP and FN are false positives and negatives, respectively, and they refer to innovators/non innovators incorrectly classified. The last column of Table 1 details the TN rate, TP rate, and the percentage of correct classification. The last metric has been used as the cost function for the optimization problem.

Table 1
Confusion matrix.

Observed		Estimated		
		Innovators*		Percentage correct
		.00	1.00	
Innovators	.00	TN	FP	TN/(TN + FP)
	1.00	FN	TP	TP/(FN + TP)
Total percentage correct				TN/(TN + FP)+TP/(FN + TP)

3.2. Simulated annealing (SA)

Simulated annealing is a method for solving unconstrained and bound-constrained optimization problems based on the analogy between the physical annealing of metals and the process of searching for the optimal solution in a combinatorial optimization problem (Cerny, 1985; Kirkpatrick, Gelatt, & Vecchi, 1983).

SA randomly generates a new point at each iteration of the algorithm. The distance of the new point from the current point is based on a probability distribution with a scale proportional to the temperature. The algorithm accepts all new points that lower the objective, but also, with a certain probability P given by Eq., (2) points that raise the objective.

$$P = \frac{1}{1 + e^{\frac{\Delta}{\max(T)}}} \tag{2}$$

where Δ is the difference between current and last objective function values and T is the current temperature. Since both Δ and T are positive, the probability of acceptance is between 0 and 1/2. Smaller temperature leads to smaller acceptance probability. Also, larger Δ leads to smaller acceptance probability.

By accepting points that raise the objective, the algorithm avoids being trapped in local minima, and is able to explore globally for more possible solutions. An annealing schedule is selected to systematically decrease the temperature as the algorithm proceeds. As the temperature decreases, the algorithm reduces the extent of its search to converge to a minimum. Eqs. (3)–(5) details several options to update the temperature:

$$T = T_0 \cdot 0,95^k \tag{3}$$

$$T = T_0/k \tag{4}$$

$$T = T_0/\log(k) \tag{5}$$

being k the iteration number and T_0 the initial temperature.

The algorithm stops when the average change in the objective function is smaller than a fixed value or a minimum temperature is attained.

3.3. Genetic algorithms

Genetic algorithms are a family of computational models inspired by evolution (Goldberg, 1989; Holland, 1975). These algorithms encode a potential solution to specific problem on a simple chromosome-like data structure and apply genetic operators to these structures in order to preserve critical information (Martínez-Torres & Toral-Marín, 2010). An initial population P_i composed of N_t chromosomes is considered. Goldberg (1989) studied the optimum number of chromosomes for a population according to the chromosome's length. His main conclusion was that the optimum population's size value gets higher as the chromosome's length increases. This initial population is generated randomly in order to preserve the diversity in the population and the fitness function is calculated to evaluate the goodness of each chromosome. The mechanism for generating the subsequent generations is based on the selection scheme from $(\mu + \lambda)$ evolution strategy (Reina, Toral, Johnson, & Barrero, 2012). The μ best chromosomes are included directly in the next generation. The crossover and mutation operations are responsible of generating λ chromosomes of a new population. The crossover consists of using two members of a population P_j to generate two new members of the next population P_{j+1} by crossing their genetic information. The new chromosomes contain genetic information from the predecessors. The purpose of mutation is to change the genetic information of a chromosome included in P_j to generate a new chromosome of P_{j+1}.

The fitness function quantifies the suitability of each chromosome as a solution. Genetic operators make selections based on

individual fitness. That means that chromosomes with high fitness value have more chances of being selected, passing their genetic material (via reproduction, crossover or mutation) to the next generation. As a result, the fitness function provides the pressure for evolution towards a new generation with chromosomes of higher fitness than the previous ones.

3.4. Particle swarm optimization

Particle swarm optimization (PSO) is an evolutionary computation technique developed by Kennedy and Eberhart (1995), which was inspired by the social behavior of bird flocking and fish schooling. It is based on the swarm intelligence concept, which refers to artificial intelligence systems where the collective behaviors of unsophisticated agents that are interacting locally with their environment create coherent global functional patterns (Del Valle, Venayagamoorthy, Mohagheghi, Hernandez, & Harley, 2008). Basically, PSO uses a population of particles that fly through the problem hyperspace. All the particles have fitness values which are evaluated by the fitness function to be optimized, and have velocities which direct the flying of the particles. These velocities are stochastically adjusted according to the historical best position for the particle itself and the neighborhood best position (Del Valle et al., 2008; Majhi and Panda, 2011). The particles fly through the problem space by following the current optimum particles.

Mathematically, PSO is formulated as follows. First, a set of P particles (population) is randomly initialized, where the position of each particle represents a solution to the problem, represented by a d-dimensional vector in problem space $s_i = (s_{i1}, s_{i2}, \ldots, s_{id})$, $i = 1, 2, \ldots, P$, $s \in \Re$. Thus each particle is randomly placed in the d-dimensional space as a candidate solution, and its performance is evaluated using a predefined fitness function. The velocity of the ith particle $v_i = (v_{i1}, v_{i2}, \ldots, v_{id})$, $v \in \Re$, is defined as the change of its position.

The information available for each individual is based on its own experience and the knowledge of the performance of other individuals in its neighborhood. Therefore, each particle adjusts its trajectory based on its own previous best position and the previous best position attained by any particle of the swarm, namely p_{id} and p_{gd}. The velocities and positions of particles are updated using Eqs. (6) and (7) respectively:

$$v_{id}(t+1) = w\, v_{id}(t) + c_i rand_i(p_{id} - s_{id}(t)) + c_2 rand_2(p_{gd} - s_{id}(t)) \quad (6)$$

$$s_{id}(t+1) = s_{id}(t) + v_{id}(t) \quad (7)$$

where t is the iteration counter, w is the inertia weigh, c_1 and c_2 are the acceleration coefficients, and $rand_1$, $rand_2$ are two random numbers in $[0,1]$. The inertia weight w controls the impact of previous histories of velocities on current velocity, and it is used to control the convergence behavior of the PSO. To reduce this weight over the iterations, allowing the algorithm to exploit some specific areas, the inertia weight w is updated according to Eq. (8):

$$w = w_{max} - \frac{w_{max} - w_{min}}{iter_{max}} iter \quad (8)$$

where w_{max}, w_{min} are the maximum and minimum values that the inertia weight can take, $iter$ is the current iteration of the algorithm and $iter_{max}$ is the maximum number of iterations. The acceleration coefficients c_1 and c_2 control how far a particle will move in a single iteration. Typically, these are both set to 2, although assigning different values to c_1 and c_2 sometimes leads to improved performance. The velocity update Eq. in (6) has three major components. The first one is the inertia, which models the tendency of the particle to continue in the same direction it has been traveling. The second component is usually referred as memory, and it is the linear attraction towards the best position ever found by the given particle p_{id} scaled by a random weigh $c_1 rand_1$. Finally, the third component usually referred as cooperation or social knowledge is the linear attraction towards the best position found by any particle p_{gd}, scaled by another random weight $c_2 rand_2$.

The newly formed particles are evaluated according to the fitness function and the algorithm iterates for a predetermined number of iterations, or until a convergence criterion has been met. Finally, the best solution obtained across all iterations is returned.

4. The data

4.1. Description of the case study

Dell IdeaStorm (http://www.dellideastorm.com) is an open innovation community where end users freely reveal and share innovative ideas with community members and Dell (Di Gangi & Wasko, 2009). Using IdeaStorm website, customers can post their ideas about existing or new Dell products, services and operations (Lambropoulos, Kampylis, & Bakharia, 2009). Moreover, users can also make comments about previously posted ideas, for instance refining or supporting proposed innovations, or just commenting their suitability. As a result, a debate is generated around those ideas arousing more interest among community members. IdeaStorm website also provides the possibility of scoring posted ideas using a collective intelligence procedure consisting of promoting or demoting ideas. Promotion means adding ten points to the current rating of the idea while demotion means subtracting ten points. Using all this information, Dell shares the ideas with top management, department managers, and key employees that work within relevant subject domains. The period 2008–2010 was considered for collecting data. During this period of time, a total of 6720 ideas were posted on IdeaStorm website by 3987 different users.

4.2. Definition of variables

Innovators are defined as those users whose ideas have been implemented by Dell. Collected results show that during the period 2008–2010 a total of 309 ideas posted by 228 different users were implemented or partially implemented by the company. That means that the immense majority of community users are non innovators and the dependent variable contains a high number of zeros. Regarding the 228 innovators, Table 2 shows the distribution of their posted ideas. It can be observed that more of 90% of innovators only post one or two ideas.

Activity of users through open innovation communities is associated to the available types of participation, that is, posting, commenting, promoting and demoting ideas. A set of eight variables has been collected from IdeaStorm website using a specific developed crawler. First of all, the crawler extracts the alias of community users. Users are required to be registered with an alias to participate in the community. Once registered, they can decide about the way they want to participate. That means users are not

Table 2
Distribution of posted ideas per author.

Posted ideas	Frequencies	Percentage	Cumulative percentage
1	203	89,04	89,04
2	11	4,82	93,86
3	6	2,63	96,49
4	1	0,44	96,93
5	2	0,88	97,81
6	3	1,32	99,12
9	1	0,44	99,56
25	1	0,44	100,00
Total	1381	100	

necessarily required to post ideas. In this work, we are specifically interested in those users who have posted at least one idea and have the potential of being an innovator. Therefore, the following variables have been considered for this subset of users:

- *N_ideas:* Number of posted ideas by each user.
- *Cat:* Number of categories that posted ideas by a given author are covering. Whenever a user posts an idea to the IdeaStorm website, it should be classified attending to a limited number of tags.
- *Comments_r:* Number of comments that posted ideas by a given user have received.
- *Comments_s:* Number of comments that each user have sent to other users' posted ideas.
- *Prom_r:* Number of promotions that posted ideas by a given user have received.
- *Prom_s:* Number of promotions that each user have sent to other users' posted ideas.
- *Dem_r:* Number of demotions that posted ideas by a given user have received.
- *Dem _s:* Number of demotions that each user have sent to other users' posted ideas.

5. Results

The three considered optimization techniques have been applied to solve the optimization problem consisting of finding the optimum coefficient values θ_i from Eq. (1) able to maximize the percentage of correct classification of IdeaStorm community users. Table 3 details the parameters setting for the proposed optimization algorithms.

Obtained results are expressed in the form of a confusion matrix, Table 4. Columns show the estimated non innovators and innovators using the three optimization techniques and how they are classified with respect to the real non innovators/innovators according to the data extracted from IdeaStorm website. The last column shows the percentage of correct classifications. The maximum possible value of the total percentage correct is 2, and it would corresponds to a perfect classification of both innovators and non innovators users.

According to Table 4, simulated annealing only reach an optimum value of 1.2948 which is a logical consequence of its lower performance in terms of exploration capabilities. Fig. 1 shows the best value evolution for three different temperature functions. The exponential temperature function reach the optimum value in 5482 iterations, much faster than the two other temperature functions. The best results in terms of percentage of correct classification are provided by genetic algorithms and particle swarm optimization, with best values of 1.4351 and 1.4664, respectively. GA requires 85 generations to converge while PSO reaches the optimum solution after 264 iterations (Figs. 2 and 3, respectively).

Table 3
Parameter settings for the three considered optimization techniques.

Simulated annealing Parameter	Value
Acceptance function	$1/1+e^{\frac{\Delta}{\max(T)}}$
Temperature	$T_0 \cdot 0,95^k$
Error gradient tolerance	1e-6
Genetic algorithms	
Population size	100
Generations	100
Crossover fraction	0.8
Migration fraction	0.15
Mutation factor	0.05
Error gradient tolerance	1e-6
Particle swarm optimization	
Number of particles	75
PSO Mode	Common PSO with inertia weights
Acceleration constants $[c_1,c_2]$	[2.1,2.1]
Inertia weights $[w_{max},w_{min}]$	[0.9,0.3]
Error gradient tolerance	1e-6
Iterations without error gradient change	40
Maximum number of iterations	400

The obtained set of coefficients and their confidence intervals for the three methods are detailed in Table 5. It can be noticed that SA gives always higher coefficients and wider confidence intervals when compared with GA and PSO, which on the other hand provides the best results in terms of classification. The obtained coefficients define the discriminant functions able to distinguish innovators from non innovators, see Eq. (1). According to the definition of innovators given by Eq. (1), a positive coefficient value means that the corresponding variable positively discriminates innovators while a negative coefficient value acts in the opposite direction.

The bootstrap method was used to compute the estimated mean and confidence bounds for the set of coefficients. The bootstrap method is a computer-based method which is useful for estimating a parameter when the underlying distribution function of the parameter is unknown (Efron, 1979). Bootstrapping uses resampling with replacement to create m resampled data sets (also known as bootstrap samples) that contain the same number of observations (n = 9 in this case) as the original data set. To perform resampling with replacement, an observation or data point is randomly selected from the original data set and then copied into the resampled data set being created. As a result, the same observation may be included in the resampled data set one, two, or more times, or not at all. Next the statistic of choice, in this case the mean value, is computed for each resampled data set. A confidence interval for the mean value is calculated from the collection of values obtained for the statistic. There are several options for computing confidence intervals. In this case, the bias-corrected and accelerated (BCa) method has been applied running 25 times each algorithm and using m = 2000 resampled data sets (Efron & Tibshirani, 1998).

The obtained mean values and confidence intervals of Table 5 shows that three of the eight coefficients are positive, three negatives and two of them cannot be guaranteed to be positive or negative for the three considered techniques. The positive coefficient define the main features of innovators:

Received comments: The number of received comments is related to the debate generated around posted ideas. That means that innovators are not only active users involved with the development of the community, but they also post ideas that arouse the interest of other users.

Sent comments: It is a measure of activity and involvement of users within the community. Active participation is one of the distinctive characteristics of lead users according to Von Hippel (1986) lead user theory.

Sent demotions: Demotions are also part of IdeaStorm scoring system and consists of negatively evaluating ideas posted by other users. This variable considers to what extent users are critical with ideas posted by other users. The obtained positive coefficient means that innovators are critical with other users' ideas.

As a difference, the following features are not applicable to innovators according to the results of Table 5

Number of posted ideas: The number of posted ideas is the most creative way in which users can participate. Although it is another way of participation and according to Von Hippel (1986) lead user theory, this variable should be an antecedent of the condition of being an innovator, our results clearly show a negative dependence. The explanation can be found in the distribution of posted ideas per author of Table 2. Almost the 90% of users have only

Table 4
Confusion matrix obtained by the three considered optimization techniques.

	Observed		Estimated		
			Innovation solvers $User^{*}_{isolvers}$		Percentage correct
			.00	1.00	
Simulated annealing (SA)	Innovation solvers	.00	2394	1365	0.6369
	$User_{isolvers}$	1.00	78	150	0.6579
	Total percentage correct				1.2948
Genetic algorithm (GA)	Innovation solvers	.00	2394	1365	0.6369
	$User_{isolvers}$	1.00	46	182	0.7982
	Total percentage correct				1.4351
Particle swarm optimization (PSO)	Innovation solvers	.00	2462	1297	0.6550
	$User_{isolvers}$	1.00	43	185	0.8114
	Total percentage correct				1.4664

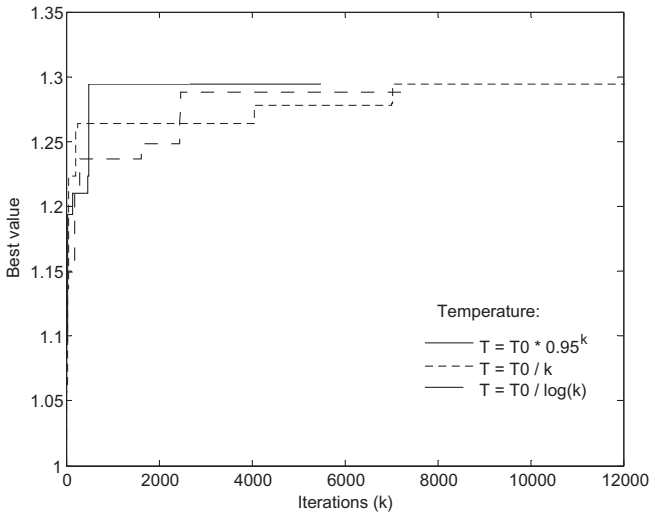

Fig. 1. Simulated annealing best value evolution.

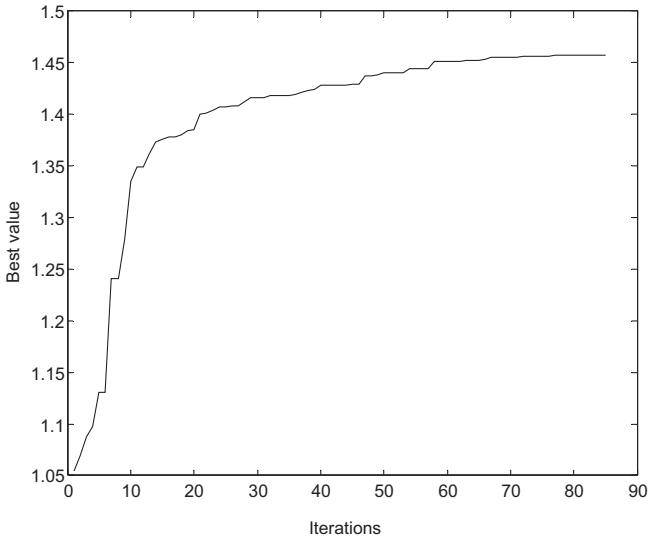

Fig. 2. GA best value evolution.

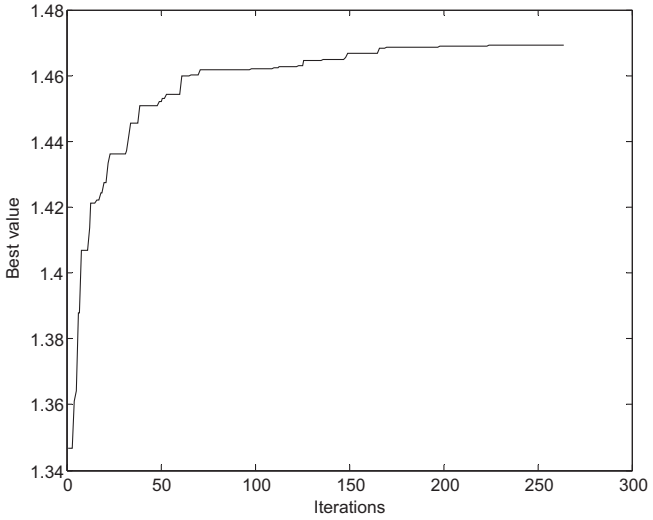

Fig. 3. PSO best value evolution.

posted one idea. Therefore, the number of posted ideas cannot be a distinctive feature of an innovative profile within the population. Users posting a high number of ideas could be considered as outliers in our sample.

Number of categories covered by posted ideas: The number of categories shows the scope of the posted ideas. Users are able to classify their posted ideas but using a set of predefined tags, and they can choose among several topics of products. The obtained results is that those ideas with a wider scope are less likely to be adopted by the company. This results seems to be contradictory with the idea that innovations are usually the result of combining pieces of knowledge from different areas. However, the list of tags provided by Dell is short and they refer to clearly independent areas or products, so those ideas focused on several tags tends to be quite generic. This fact can explain the negative value for this coefficient.

Received promotions: Promotions are part of IdeaStorm scoring system. It is a way of putting in practice the evaluation of ideas using collective intelligence. The negative coefficient value for promotions means that the community is evaluating ideas following their possible benefits and not the applicability of these ideas. Ideas that receive a high number of votes are those ideas who satisfy to the majority of users, but they are usually difficult to be adopted by the company. Consequently, the number of received promotions is not a good indicator of innovators. Notice that we have defined innovators are those users proposing ideas finally adopted by Dell. However, strategic innovation policies of Dell are not always aligned with users' desires. Some non affordable ideas can be excellent for users but prohibitive for the company.

The two variables in which confidence intervals do not clearly show a positive or negative value are 'Sent promotions' and 'Received demotions'. The first one is related to the activity of users. This result suggests that the activity of author commenting ideas is quite different from the activity promoting ideas. The fact of commenting an idea requires a good understanding and knowledge of the posted idea, while promotions are frequently done as a result of a feeling about the possible outcomes. Therefore, commenting is a form of participation more related to an innovative

Table 5
Coefficients' optimum values.

Coefficients	SA	95% confidence interval	GA	95% confidence interval	PSO	95% confidence interval
N_ideas	−2472	[−4726 −0217]	−0329	[−0584 −0074]	−0827	[−1135 −0519]
Cat	0380	[−1327 2087]	−0146	[−0228 −0063]	−0235	[−0358 −0112]
Comments_r	8438	[7948 8929]	3329	[2906 3752]	8741	[8246 9235]
Comments_s	3038	[1111 4965]	0208	[0073 0344]	0960	[0186 1733]
Prom_r	−0712	[−2844 1421]	−0608	[−0827 −0389]	−1980	[−3394 −0565]
Prom_s	1531	[0671 2391]	0017	[−0011 0046]	−0057	[−0168 0053]
Dem_r	−1904	[−3987 0179]	−0254	[−0500 −0009]	0063	[−1270 1396]
Dem_s	3111	[1346 4877]	0515	[0398 0632]	2161	[1257 3066]
Constant	−6132	[−7447 −4817]	−1221	[−1529 −0913]	−5747	[−7077 −4416]

character. With respect to the 'Received demotions', it is again the result of applying a collective intelligence method for evaluating ideas and the same that received promotions is not revealing an innovative profile.

6. Conclusion

This paper deals with the problem of identifying innovators in open innovation communities using variables related to their activity. Mathematically, innovators are estimated using discriminant functions obtained by a linear combination of the selected variables. Due to the zero inflated characteristic of the dependent variable, the problem has been formulated as an optimization problem consisting of determining the coefficient values of the discriminant function so that the innovators and non innovators identification ratios are maximized. Three different optimization techniques have been used for this purpose in order to validate the results. Each algorithm was executed 25 times to average the coefficients' values and a bootstrapping technique was then applied to obtain the 95% confidence interval. From the viewpoint of the methodology, obtained results show that GA and PSO solve the optimization problem better than SA, leading to best results in terms of classification as well as in terms of smaller ranges in the confidence intervals. From the viewpoint of the application problem, obtained results clearly show that the interactions among users through comments (both sent and received) are better indicators of innovative profiles than the interactions though the scoring system.

References

Amabile, T., Barsade, S. G., Mueller, J. S., & Staw, B. M. (2005). Affect and creativity at work. *Administrative Science Quarterly, 50*, 367–403.

Barge-Gil, A. (2010). Open, semi-open and closed innovators: Towards an explanation of degree of openness. *Industry and Innovation, 17*(6), 577–607.

Berthon, P. R., Pitt, L. F., McCarthy, I., & Kates, S. M. (2007). When customers get clever: Managerial approaches to dealing with creative consumers. *Business Horizons, 50*(1), 39–47.

Blohm, I., Riedl, C., Leimeister, J. M., & Krcmar, H. (2011). Idea evaluation mechanisms for collective intelligence in open innovation communities: Do traders outperform raters?. In *Proceedings of 32nd international conference on, information systems* (pp. 1–24).

Carbone, F., Contreras, J., Hernández, J. Z., & Gomez-Perez, J. M. (2011). Open innovation in an enterprise 3.0 framework: Three case studies. *Expert Systems with Applications, 39*(10), 8929–8939.

Cerny, V. (1985). Thermodynamical approach to the travelling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications, 45*, 41–51.

Chesbrough, H. (2003). *Open innovation: The new imperative for creating and profiting from technology*. Boston, MA: Harvard Business School Press.

Chesbrough, H., Vanhaverbeke, W., & West, J. (2006). *Open innovation: Researching a new paradigm*. Oxford: Oxford University Press.

Dahlander, L., Frederiksen, L., & Rullani, F. (2008). Online communities and open innovation. *Industry and Innovation, 15*(2), 115–123.

De Jong, J. P. J., Kalvet, T., & Vanhaverbeke, W. (2010). Exploring a theoretical framework to structure the public policy implications of open innovation. *Technology Analysis & Strategic Management, 22*(8), 877–896.

Di Gangi, P. M., & Wasko, M. (2009). Steal my idea! organizational adoption of user innovations from a user innovation community: A case study of Dell IdeaStorm. *Decision Support Systems, 48*(1), 303–312.

Del Valle, Y., Venayagamoorthy, G. K., Mohagheghi, S., Hernandez, J.-C., & Harley, R. G. (2008). Particle swarm optimization: Basic concepts, variants and applications in power systems. *IEEE Transactions on Evolutionary Computation, 12*(2), 171–195.

Efron, B. (1979). Bootstrap methods. Another look at the jackknife. *The Annals of Statistics, 7*, 1–26.

Efron, B., & Tibshirani, R. J. (1998). *An introduction to the bootstrap*. New York: Chapman & Hall/CRC.

Frey, K., Lüthje, C., & Haag, S. (2011). Whom should firms attract to open innovation platforms? The role of knowledge diversity and motivation. *Long Range Planning, 44*(5–6), 397–420.

Fueller, J., & Matzler, K. (2007). Virtual product experience and customer participation – A chance for customer-centred, really new products. *Technovation, 27*, 378–387.

Goldberg, D. E. (1989). *Genetic algorithm in search, optimization and machine learning*. Reading, MA: Addison-Wesley.

Holland, J. (1975). *Adaptation in natural and artificial systems*. Ann Arbor, MI: University of Michigan Press.

Hüsig, S., & Kohn, S. (2011). "Open CAI 2.0" – Computer aided innovation in the era of open innovation and Web 2.0. *Computers in Industry, 62*(4), 407–413.

Keeble, D., & Wilkinson, F. (Eds.). (2000). *High-technology clusters, networking and collective learning in Europe*. Aldershot: Ashgate.

Kennedy, J., & Eberhart, R. C. (1995). Particle swarm optimization. In *Proceedings of the IEEE international conference on neural networks* (pp. 1942–1948). Berlin: Springer.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science, 220*, 671–680.

Lambropoulos, N., Kampylis, P., & Bakharia, A. (2009). User innovation networks and research challenges, online communities and social computing. *Lecture Notes in Computer Science, 5621*, 364–373.

Laursen, K., & Salter, A. J. (2006). Open for Innovation: The role of openness in explaining innovation performance among UK manufacturing firms. *Strategic Management Journal, 27*(2), 131–150.

Lee, A. H., Wang, K., Scott, J. A., Yau, K. K., & McLachlan, G. J. (2006). Multi-level zero-inflated poisson regression modelling of correlated count data with excess zeros. *Statistical Methods in Medical Research, 15*, 47–61.

Love, J. H., & Roper, S. (2009). Organizing the innovation process: Complementarities in innovation networking. *Industry and Innovation, 16*(3), 273–290.

Mahr, D., & Lievens, A. (2012). Virtual lead user communities: Drivers of knowledge creation for innovation. *Research Policy, 41*(1), 167–177.

Majhi, B., & Panda, G. (2011). Robust identification of nonlinear complex systems using low complexity ANN and particle swarm optimization technique. *Expert Systems with Applications, 38*(1), 321–333.

Martínez-Torres, M. R., & Toral-Marín, S. L. (2010). Strategic group identification using evolutionary computation. *Expert Systems with Applications, 37*(7), 4948–4954.

Martínez-Torres, M. R. (2012). A genetic search of patterns of behaviour in OSS communities. *Expert Systems with Applications, 39*(18), 13182–13192.

Morrison, P. D., Roberts, J. H., & Midgley, D. F. (2004). The nature of lead users and measurement of leading edge status. *Research Policy, 33*, 351–362.

Poetz, M. K., & Schreier, M. (2012). The value of crowdsourcing: Can users really compete with professionals in generating new product ideas? *Journal of Product Innovation Management, 29*, 245–256.

Reina, D. G., Toral, S. L., Johnson, P., & Barrero, F. (2012). An evolutionary computation approach for designing mobile ad hoc networks. *Expert Systems with Applications, 39*(8), 6838–6845.

Rohrbeck, R., Steinhoff, F., & Perder, F. (2008). Virtual customer integration in the innovation process: Evaluation of the web platforms of multinational enterprises (MNE). In *International conference on management of engineering & technology, PICMET 2008* (pp. 469–478) Portland.

Rowley, J., Kupiec-Teahan, B., & Leeming, E. (2007). Customer community and co-creation: A case study. *Marketing Intelligence & Planning, 25*, 136–146.

Schwab, S., Koch, J., Flachskampf, P., & Isenhardt, I. (2011). Strategic implementation of open innovation methods in small and medium-sized enterprises. In *Proceedings of the 2011 17th international conference on concurrent enterprising (ICE 2011)* (pp. 1–8).

Spann, M., & Skiera, B. (2009). Sports forecasting: A comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *Journal of Forecasting, 28*(1), 55–72.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. Random House, Inc..

Tödtling, F., Lehner, P., & Trippl, M. (2006). Innovation in knowledge intensive industries: The nature and geography of knowledge links. *European Planning Studies, 14*(8), 1035–1058.

Tödtling, F., Prud'homme van Reine, P., & Dörhöfer, S. (2011). Open innovation and regional culture – Findings from different industrial and regional settings. *European Planning Studies, 19*(11), 1885–1907.

Urban, G. L., & von Hippel, E. (1988). Lead user analyses for the development of new industrial products. *Management Science, 34*(5), 569–582.

Von Hippel, E. (1986). Lead users: A source of novel product concepts. *Management Science, 32*(7), 791–805.

Von Hippel, E. (1988). *The sources of innovation*. New York: Oxford University Press.

Von Hippel, E. (2001). Innovation by user communities: Learning from open-source software. *MIT sloan management review*, 82–86.

Von Hippel, E., & Katz, R. (2002). Shifting innovation to users via toolkits. *Management Science, 48*(7), 821–833.

Von Hippel, E., & von Krogh, G. (2003). Open source software and the ''private-collective'' innovation model: Issues for organization science. *Organization science, 14*(2), 209–223.

West, J., & O'mahony, S. (2008). The role of participation architecture in growing sponsored open source communities. *Industry & Innovation, 15*(2), 145–168.