Expert Systems with Applications 38 (2011) 5794–5800

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Identifying financially successful start-up profiles with data mining

David Martens [a,*], Christine Vanhoutte [c], Sophie De Winne [c,d], Bart Baesens [b,e], Luc Sels [c], Christophe Mues [e]

[a] Faculty of Applied Economics, University of Antwerp, Belgium
[b] Department of Decision Sciences and Information Management, K.U. Leuven, Belgium
[c] Research Centre for Organisation Studies, K.U. Leuven, Belgium
[d] Lessius University College, Belgium
[e] School of Management, University of Southampton, United Kingdom

ARTICLE INFO

Keywords:
Data mining
Active learning
Start-up companies
Ideal configurations

ABSTRACT

Start-ups are crucial in the modern economy as they provide dynamism and growth. Research on the performance of new ventures increasingly investigates initial resources as determinants of success. Initial resources are said to be important because they imprint the firm at start-up, limit its strategic choices, and continue to impact its performance in the long run. The purpose of this paper is to identify configurations of initial resource bundles, strategy and environment that lead to superior performance in start-ups. To date, interdependencies between resources on the one hand and between resources, strategy and environment on the other hand have been neglected in empirical research. We rely on data mining for the analysis because it accounts for premises of configurational theory, including reversed causality, intradimensional interactions, multidimensional dependencies, and equifinality. We apply advanced data mining techniques, in the form of rule extraction from non-linear support vector machines, to induce accurate and comprehensible configurations of resource bundles, strategy and environment. We base our analysis on an extensive survey among 218 Flemish start-ups. Our experiments indicate the good performance of rule extraction technique ALBA. Finally, for comprehensibility, intuitiveness and implementation reasons, the tree is transformed into a decision table.

1. Introduction

Start-up companies are very important for economic dynamism and growth but often have difficulties to stay in the market (Davidsson, Lindmark, & Olofsson, 2006). In order to understand this high failure rate among start-ups, researchers within the entrepreneurship research domain have focused on identifying the determinants of new venture performance in the last couple of decades. Thus far, initial resources have been designated as important determinants of new venture performance (see e.g. the work of Bamford, Dean, & McDougall (2000), Cooper, Gimeno-Gascon, & Woo (1994)).

Resources, defined as "all tangible and intangible assets that are tied to the firm in a relatively permanent fashion" (Wernerfelt, 1984), are not only indispensable for the basic functioning of a firm, they can also serve as sources of competitive advantage. This is the key idea of the resource-based view (RBV) in which it is posited that resources can only be sources of (sustained) competitive advantage if they are valuable, rare, costly to imitate and properly exploited by the organization (Barney, 1991). Moreover, initial resources, i.e., the firm's resources present at the point of inception, imprint the firm at start-up and thus affect its future competitive position (Bamford, Dean, & Douglas, 2004; Boeker, 1989).

Previous research has mainly focused on the identification of resources leading to superior performance. The majority of these studies concentrated on testing universalistic (independent of other resources and context) or contingency (independent of other resources, dependent on the organizational or the environmental context) models. Despite providing useful insights, both perspectives neglect two important issues. Firstly, resources interact with each other. This means that the strategic value of a resource is dependent on other resources. Therefore, one should assess the strategic value of resources at the resource bundle level (Black & Boal, 1994). Existing research, however, has been restricted to the analysis of additive effects of resources on the overall performance. Secondly, organizations face multiple contingencies at the same time. The value of a firm's resources needs to be evaluated within the context of the firm's strategy, as well as the specific market environment.

Existing studies neglect this multivariate dependency of a firm's performance on resources, organizational and environmental factors. In this paper, we address both of these concerns by taking an inductive (data mining) approach, based on survey and financial data, to identify configurations of resource bundles, strategy and environment that yield good start-up performance. The remainder

* Corresponding author.
E-mail addresses: David.Martens@econ.kuleuven.be, David.Martens@ua.ac.be (D. Martens).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.10.052

of the paper is structured as follows. The next section further describes the problem setting: identifying resource combinations that impact new venture performance. In Section 3, the data gathering process, experimental setup and data mining techniques are elaborated on. Section 4 provides the results of our analysis, while Section 5 concludes the paper.

2. Resources and new venture performance

2.1. Initial resource bundles

New venture performance research increasingly takes a resource-based view and investigates initial resources as determinants of success. Initial resources are said to be important because they imprint the firm at start-up, limit its strategic choices, and continue to impact its performance in the long run (Bamford et al., 2000; Boeker, 1989; Cooper et al., 1994). Even though firms are presented as 'bundles of resources' in the RBV, the majority of studies in this area treat resources as isolated elements. Yet, it can be argued that the strategic value of resources resides at the resource bundle level because interdependencies might exist between them.

Black and Boal (1994) identify three different types of interactive relationships. First, resources may compensate for each other implying that a change in the level of one resource is offset by a change in the level of another resource. Chandler and Hanks (1998), for example, show that for start-up companies human and financial capital compensate each other. Second, resources may enhance each other. This implies that the 'combination' of resources improves performance. Barney's organizational resources (e.g. compensation policy) for example, are by themselves limited in creating a competitive advantage, but enable a firm to exploit its other resources that do carry the potential for a competitive advantage (such as knowledge workers). And finally, resources may suppress each other. In this case, the presence of one resource diminishes the impact of another. Organizations that have such resources are better off with only one type or none of the resources. A very centralized, hierarchical firm structure, for example, might hinder the firm's capability for innovation.

To conclude, the successful implementation of a competitive strategy is not only dependent on the suitability of the individual resources to achieve a specific goal, but also on the relationships among them.

2.2. Resource, strategy and environment interactions

In searching for initial resources that confer success to start-ups, researchers should take into account that resources alone do not create value. Only when they are used to implement a strategy that creates valuable products or services for customers, they can be deemed valuable. Contingency theory puts forward that the need for resources differs across strategies. For example, innovation strategies require highly creative employees whereas cost strategies rather require low-cost labor.

Furthermore, the competitiveness of a strategy highly depends on the environment in which the company operates. Dynamic and uncertain environments generally require more innovation than do stable environments.

Given that *resources* are only valuable when they support a certain *strategy* that exploits opportunities (or neutralizes threats) in the company's environment, Barney and Clark (2007) argue that the value of a company's resources is not only contingent upon the strategy of the firm, but also upon its environmental context. Hence, firms should simultaneously align their resources, strategy and environment in order to achieve a competitive advantage. This in essence is a configurational view on competitive advantage, in which a configuration is defined as a multivariate combination of firm characteristics and in which an ideal configuration is defined as a configuration in which the key attributes are tightly interrelated and mutually reinforcing.

From the above, we can conclude that research on the initial resource- performance relationship should move beyond universalistic and contingency perspectives. It should take into account that a competitive advantage will most likely lie in the complexity of resource bundles and the capability of simultaneously aligning a firm's internal and external attributes: its resources, its strategic orientation, and its environment. In this study, we investigate which configurations of initial resource bundles, strategy and environment are associated with financially healthy start-ups. We do so by means of classification trees (Martens, Van Gestel, & Baesens, 2009; Quinlan, 1993). Rules are thus formed indicating how different combinations lead to financially healthy start-ups. Consequently, tree analysis provides an assessment of the most discriminating resources (as well as strategy and environment characteristics) with respect to performance, in sequence and in combination, and an indication of how extensively these resources should be present. The principal advantage of tree analysis is that it allows for a direct search of optimal resource configurations as opposed to cluster analysis. In addition, this technique is easily interpretable and applicable to practice.

3. Experimental setup

3.1. Data collection process

The dataset on which to apply the tree induction technique was obtained via the procedure illustrated in Fig. 1. All independent resource-based variables were taken from the START 2003 survey. The target variable ($Y2$) was obtained by applying a start-up specific failure prediction model to the 2004 financial data of the START 2003 companies. This failure prediction model (with target variable $Y1$, indicating whether the company survived or not), was built using data obtained from the BELFIRST database. The final target variable $Y2$ was obtained by dummy-encoding the predicted probability p with the sector trimmed average. As such, $Y2$ is an indication of the global financial health of the firm. More details on the procedure are described next.

3.1.1. Data sources

The data on resources, strategy and environment were collected by the Flemish Policy Research Center for Entrepreneurship, Enterprises and Innovation in 2003 (START). The START 2003 survey targeted incorporated Flemish firms in all economic sectors founded between September 1st 2001 and September 1st 2002 and employing 1–49 people. Of the 2679 start-ups, 512 could not be reached and 637 filled in the questionnaire after two reminders (one by mail and one by telephone in October–November 2003). Because of item non-response and the removal of non-independent firms, the analyses are based on 218 observations.

Financial data on all starters during the same time period as the START 2003 companies (data set 1) were retrieved from the BELFIRST database which contains the annual accounts of 320,000 Belgian and 4000 Luxembourg incorporated firms. Using annual accounts for performance measurement offers the advantage of a broad spectrum of audited performance measures that are comparable across organizations.

3.1.2. Target variable definition and measurement

Our target variable is a measure for the competitive advantage of a firm in terms of global financial condition, i.e., the

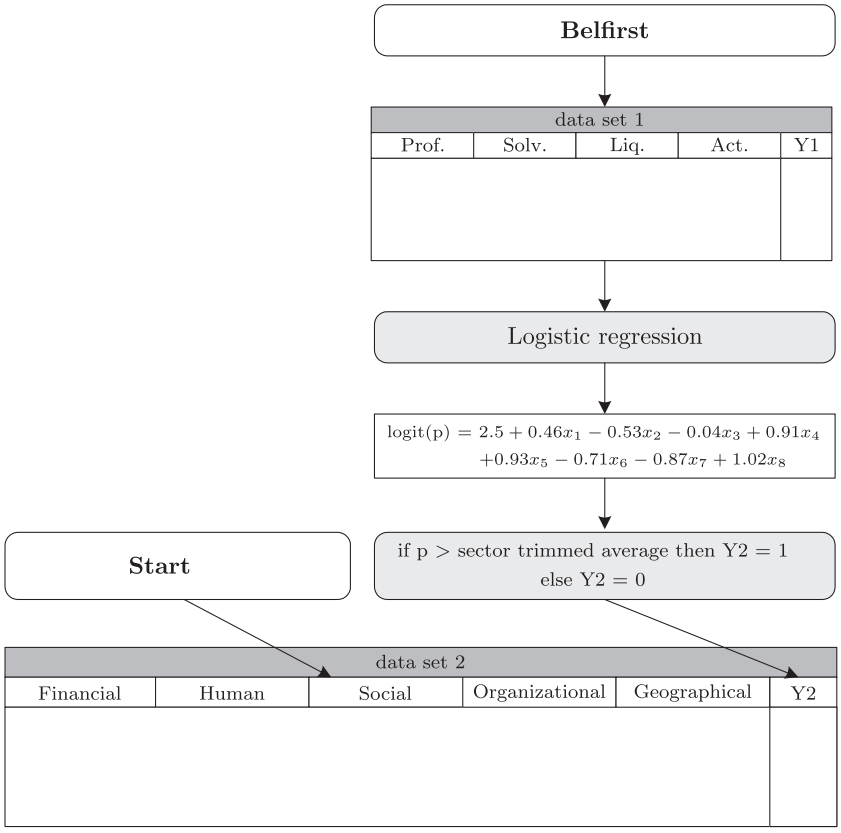

Fig. 1. Data collection process.

likelihood of organizational survival. This measure holds a number of advantages over more commonly used measures such as Return On Assets (ROA), Return on Equity (ROE), and growth (Murphy, Trailer, & Hill, 1996). Firstly, this measure is appropriate in a start-up context since companies are primarily preoccupied with the struggle to survive in the first years of their life cycle. Secondly, this measure is limitedly subject to goal differences and hence highly comparable among new businesses. Thirdly, this measure unites several performance dimensions such as profitability, solvency, liquidity, and activity into one figure, reflecting the overall financial condition of the firm (Altman, 1993; Pompe & Bilderbeek, 2005). This is an important feature since resource-based theory always refers to performance outcomes in general terms, e.g. success, competitive advantage, and superior performance. We thus assume that firms should score well on multiple dimensions to have a competitive advantage. It is difficult to attribute a firm a (realized) competitive advantage when it has an above average profitability but at the same time a poor liquidity or solvency. These advantages strengthen our view that the likelihood of organizational survival is a valuable alternative to other performance measures such as ROA, ROE, and growth.

The likelihood of organizational survival is obtained by applying logistic regression that predicts the survival for a set of companies. Since failure is said to be geographically bounded (Pompe & Bilderbeek, 2005), the included companies are all Flemish incorporated and for profit firms that were founded between January 1st 2001 and December 31st 2002. Additional criteria are imposed with respect to the number of employees (less than 50 employees in 2003) and the presence of annual accounts (presence required in 2004). A distinction is further made between failed and surviving firms. A firm is classified as failed if the firm went bankrupt in 2005 or 2006. A firm is considered as surviving if it is still active in the same capacity in 2005 or 2006. Firms that were taken over or voluntarily discontinued their activity in this period are omitted from the database because it is not clear what their financial situation was. The dummy indicating whether a firm was surviving or failing in the 2005–2006 period serves as the dependent variable in a logistic regression analysis modeling the probability of survival. The set of independent variables are financial indicators that emerge as essential predictors of start-up failure in previous research (Altman, 1993; Martens, Bruynseels, Baesens, Willekens, & Vanthienen, 2008; Pompe & Bilderbeek, 2005). This initial set contains different measures of profitability, solvency, liquidity and activity for the fiscal year 2004. A stepwise logistic regression resulted in a model with eight significant parameters:

- x_1: net operating result/total assets (profitability)
- x_2: industry-adjusted ratio quick assets/total assets (liquidity)
- x_3: cash/amounts payable within 1 year (liquidity)
- x_4: cash/current assets (liquidity)
- x_5: trade debts/total assets (liquidity)
- x_6: stocks/current working assets (liquidity)
- x_7: dummy variable with value 1 when overdue short-term priority debts is positive and 0 when overdue short-term priority debts is equal to 0 (liquidity)
- x_8: dummy variable with value 1 when equity is positive or 0 and 0 when equity is negative (solvency)

This model was subsequently applied to the financial data of the START 2003 companies. The resulting probability of survival gives

an indication of the financial health of the firm in 2004. We subsequently dichotomized the probability score by allocating firms to the '1' class when their probability score was above the sector trimmed average (the threshold for realized competitive advantage) and to the '0' class when their probability score was below the sector trimmed average. Firms that effectively failed by the end of 2004 were allocated to the lower class. The result is a variable representing the competitive (dis)advantage of the firm in terms of global financial condition. In our final sample 143 firms were classified as financially superior and 75 firms were classified as financially inferior.

3.2. ALBA: Rule extraction from non-linear support vector machine models

A decision tree algorithm can be applied to the historical data, obtained in a manner discussed before, in order to come to ideal configurations of start-ups that are of course as comprehensible and accurate as possible. Decision tree builder C4.5 is a popular decision tree induction technique that is suitable for this problem. Previous research has shown that the SVM rule extraction technique, ALBA, can improve the results obtained by C4.5 further, both in accuracy and in comprehensibility (Martens et al., 2009; Verbeke, Baesens, Martens, De Backer, & Haesen, 2009; Verbeke, Martens, Mues, & Baesens, 2011). More details about this technique are provided next.

The support vector machine (SVM) (Vapnik, 1995) is currently one of the state-of-the-art classification techniques. Benchmarking studies reveal that in general, the SVM performs best among current classification techniques (Baesens et al., 2003), due to its ability to capture non-linearities. However, its strength is also its main weakness, as the generated non-linear models are typically regarded as incomprehensible black-box models. The opaqueness of SVM models can be remedied through the use of rule extraction techniques, which induce rules that mimic the black-box SVM model as closely as possible. Through rule extraction, some insight is provided into the logics of the SVM model (Martens, Baesens, Van Gestel, & Vanthienen, 2007).

ALBA is a rule extraction algorithm that uses specific concepts of the SVM, being the support vectors, and simple rule induction techniques such as C4.5 and RIPPER (Martens et al., 2009). Active learning entails the control of the learning algorithm over the input data on which it learns. More specifically, active learning focuses on the problem areas (Cohn, Atlas, & Ladner, 1994), which for rule extraction are those areas in the input space were the noise is the highest. These regions are found near the SVM decision boundary, which marks the transition of one class to another. First, we can change the labels of the data instances by the SVM predicted labels. In this manner the induced rules will mimic the SVM model and all noise is omitted from the data, removing any apparent conflicts in the data. Second, to incorporate the active learning approach additional data instances are generated close to the decision boundary. For this explicit use is made of the support vectors, which are typically close to the decision boundary. The support vectors are thus used as proxies for the decision boundary by generating additional data instances close to the support vectors. Since the distribution of the support vectors will follow the data distribution, more support vectors will be found in dense input areas, and less in more sparse ones. This implicit incorporation of the existing data distribution in the extra data generation step, eliminates the necessity to explicitly take into account density measures. The Active Learning Based Approach is described formally in Algorithm 1. A full discussion on ALBA can be found in Martens et al. (2009).

Algorithm 1 Pseudo-code of ALBA algorithm

1: preprocess data $\mathscr{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$
2: split data in training data $\mathscr{D}_{tr}$, and test data $\mathscr{D}_{te}$ in a 2/3, 1/3 ratio
3: tune SVM parameters with gridsearch on $\mathscr{D}_{tr}$
4: train SVM on $\mathscr{D}_{tr} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_{tr}}$, providing an oracle SVM mapping a data input to a class label
5: change the class labels of the training data to the SVM predicted class
6: *% Calculate the average distance $distance_k$ of training data to support vectors, in each dimension k*
7: **for** $k = 1$ to n **do**
8: $distance_k = 0$
9: **for all** support vectors $\mathbf{sv}_j$ **do**
10: **for all** training data instance d in $\mathscr{D}_{tr}$ **do**
11: $distance_k = distance_k + |d_k - sv_{j,k}|$
12: **end for**
13: **end for**
14: $distance_k = \frac{distance_k}{\#\mathbf{sv} \times N_{tr}}$
15: **end for**
16: *% Create 1000 extra data instances*
17: **for** $i = 1$ to 1000 **do**
18: randomly choose one of the support vectors $\mathbf{sv}_j$
19: *%Randomly generate an extra data instance $\mathbf{x}_i$ close to $\mathbf{sv}_j$*
20: **for** $k = 1$ to n **do**
21: $x_{i,k} = sv(j,k) + \left[(rand - 0.5) \times \frac{distance_k}{2}\right]$ with *rand* a random number in $[0,1]$
22: **end for**
23: provide a class label y_i using the trained SVM as oracle: $y_i = \mathrm{SVM}(\mathrm{x}_i)$
24: **end for**
25: run rule induction algorithm on the data set containing both the training data $\mathscr{D}_{tr}$, and newly created data instances $\{(\mathbf{x}_i, y_i)\}_{i=1:1000}$
26: evaluate performance in terms of accuracy, fidelity and number of rules, on $\mathscr{D}_{te}$

After improving the input data, the tree induction technique C4.5 is applied to induce trees. C4.5 is a popular decision tree builder (Quinlan, 1993; Tan, Steinbach, & Kumar, 2005; Witten & Frank, 2000) where each leaf assigns a class label to observations. Each of these leaves can be represented by a rule and therefore C4.5 builds comprehensible classifiers.

3.3. Oversampling

The distribution of the class variable of the dataset is not balanced, with 66% of the data consisting of financially healthy start-up companies. This leads to classification techniques experiencing difficulties in classifying both the financially healthy and unhealthy correctly. With oversampling, observations of the minority class in the training set are copied and added to the training set. Only the training data is oversampled and the test set is not, in order to provide an unbiased indication of the performance of the model towards future predictions.

Oversampling leads to a change in the model and thus in the resulting accuracy (percentage of instances correctly classified), sensitivity (percentage of financially healthy that are correctly predicted) and specificity (percentage of financially unhealthy that are correctly classified). Table 1 illustrates this, showing the change in performance metrics when including oversampling. The results will be discussed next.

Table 1
Results from predictive data mining techniques in terms of (test) accuracy, sensitivity, specificity and number of rules.

	Acc	Sensitivity	Specificity	Nb Rules
No oversampling				
SVM	**0.65**	0.89	*0.15*	
C4.5	*0.56*	*0.67*	*0.34*	8.9
ALBA	0.65	**0.91**	*0.13*	11.7
Oversampling twice				
C4.5	*0.53*	*0.57*	0.44	30.1
ALBA	*0.52*	*0.50*	**0.56**	20.6

4. Results and discussion

4.1. Output and performance of data mining models

For each technique, 10 randomizations were created, followed by a two third, one third dataset split-up in training and test set. Table 1 shows the average results of our experiments (on the test set). We then use a paired *t*-test to test the performance differences. For each performance metric, the best performance measure is underlined and in bold face. Performances that are not significantly different at the 5% level from the top performance with respect to a one-tailed paired *t*-test are tabulated in bold face. Statistically significant underperformances at the 1% level are emphasized in italics. Performances significantly different at the 5% level but not at the 1% level are reported in normal script. Although the observations of the randomizations are not independent, we remark that this standard *t*-test is used as a common heuristic to test the performance differences (Dietterich, 1998).

When considering test accuracy, the SVM classifiers perform the best, as expected (Baesens et al., 2003; Lessmann, Baesens, Mues, & Pietsch, 2008; Martens et al., 2007; Van Gestel et al., 2004). The trees extracted by the ALBA technique (with no oversampling) however yield a similar accuracy. Compared to C4.5, both SVM and ALBA perform significantly better (at 1% level). When considering sensitivity, ALBA performs even better, with the overall best result. With regards to specificity, oversampling puts a higher focus on predicting the financially unhealthy correctly, and hence results in an improved specificity. Once again ALBA performs best, but C4.5 does not perform significantly worse. Although oversampling provides us with more rules (or more leaves in the tree), the size is still quite small and surely acceptable in this setting.

Overall, we can conclude that the ALBA technique performs very well and is able to combine the accuracy of the SVM with the comprehensibility of rules. The tree thereof is shown in Fig. 2. In what follows, we will convert the tree into a decision table.

4.2. Decision tables

Decision tables are a tabular representation used to describe and analyze decision situations (Vanthienen, Mues, & Aerts, 1998) and consist of four quadrants, separated by double-lines, both horizontally and vertically. The horizontal line divides the table into a condition part (above) and an action part (below), while the vertical line separates subjects (left) from entries (right). The condition subjects are the problem criteria (the variables) that are relevant to the decision-making process. The action subjects describe the possible outcomes of the decision-making process; i.e., the classes of the classification problem: financial health = good or bad. Each condition entry describes a relevant subset of values (called a state) for a given condition subject (variable), or contains a dash symbol ('–') if its value is irrelevant within the context of that column. Subsequently, every action entry holds a value assigned to the corresponding action subject (class).

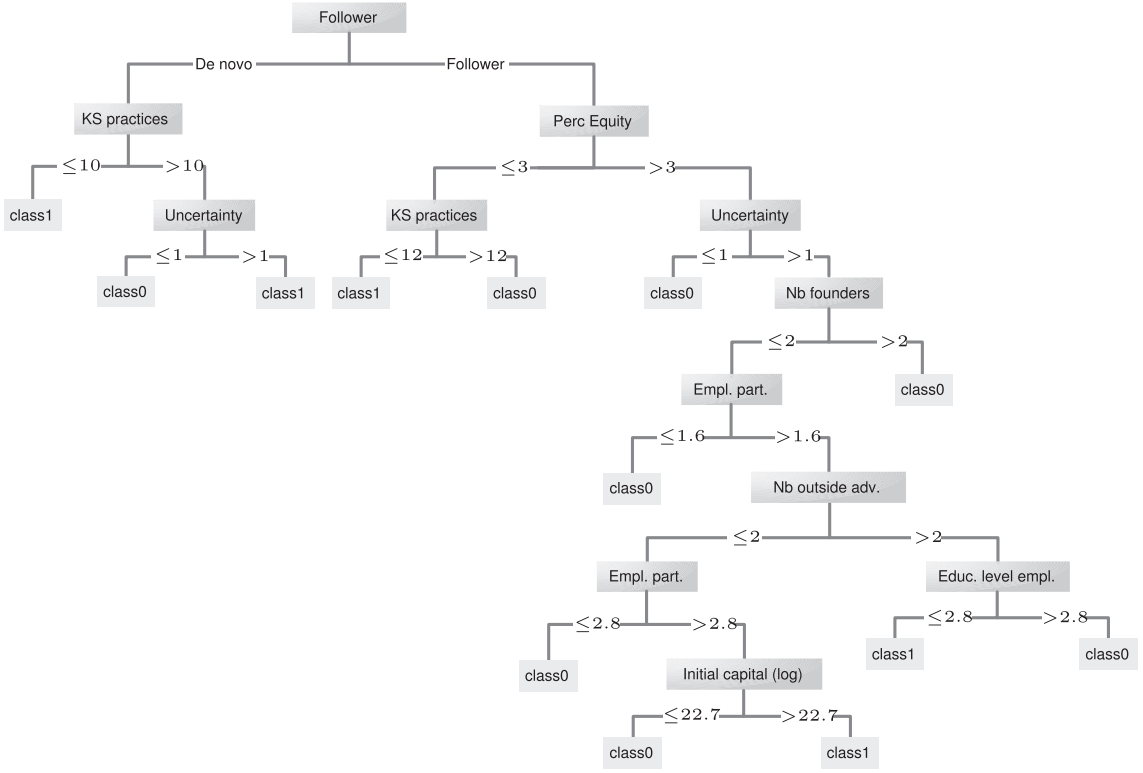

Fig. 2. Decision tree extracted with the ALBA technique.

Every column in the entry part of the decision table thus comprises a classification rule, indicating what action(s) apply to a certain combination of condition states. For example, in Table 2, the first column tells us that a configuration where we have a novice start-up company (follower business = 0) with number of knowledge sharing practices $\leqslant 10$ is considered financially healthy. Decision tables can be contracted by combining logically adjacent (groups of) rows that lead to the same action configuration.

Recent empirical research suggests that the tabular layout of decision tables offers some advantages over other representations such as decision trees or rules in terms of the accuracy and response time by which human users can apply them to various problem-solving tasks such as determining the correct classification for a given case or validating certain properties of the classifier (Huysmans, Dejaeger, Mues, Vanthienen, & Baesens, in press). Secondly, decision tables that are constructed according to the criteria above have shown to be advantageous with respect to verification and validation (Vanthienen, Mues, & Aerts, 1998). On the one hand, decision tables can be easily checked for logical anomalies such as inconsistencies (i.e., different classes being assigned to the same case) or incompleteness (i.e., columns having no action entries). In addition, by readily providing algorithms to reorder conditions, decision tables also facilitate checking for violations of variable monotonicity constraints that may be imposed onto the domain (Martens et al., 2006). Any parts of the input space where these constraints are violated are clearly indicated by moving the attribute for which a monotonic relationship with the target class is assumed to the bottom position in the table condition order. Thirdly, another potential advantage of decision tables is that they are supported by most commercially available business rule management systems (see e.g. Halle (2001)) and hence naturally bridge the gap to a practical software implementation using a standard business rule engine and to their subsequent maintenance.

4.3. Discussion

From the decision tree in Fig. 2 and the decision table in Table 2, it can be observed that start-ups can achieve an advantage vis-a-vis the competition in five different ways while they can end up with a competitive disadvantage in eight different ways. Together, these different combinations form the 13 rules that fully cover the dataset. Looking inside the rules, we see that nine different variables were selected by the algorithm for their discriminatory properties. Of these, eight are resource-related variables representing organizational, human, social and financial capital and one is an environment-related variable (demand uncertainty). These results bring along a number of interesting insights both on the higher and the lower theoretical level.

On the higher theoretical level, remarkable findings emerge with respect to equifinality (multiple ways exist to reach a competitive advantage), the role of resource bundles and the role of multivariate fit between resources, strategy and environment in determining competitive advantage. First, the analysis shows that there exist five different ways to reach a competitive advantage. This is an important finding because it proves that the basic assumption of equifinality in configurational theory is pertinent and hence that the same final state (i.e., competitive advantage) can be reached by a variety of paths. Second, all five ideal configurations contain more than one resource property confirming the importance of the 'resource bundle level'. As such, our findings show that the value of resources indeed lies in their combination and that business success is a matter of a whole rather than a separate piece. A third remarkable finding of this study is that resources occupy the highest positions in the decision tree. This substantiates the RBV assertion that resources are the prime determinants of competitive advantage. It even turns out that in two of the five ideal configurations resources are the only discriminators which seems to suggest that there exist universalistic resource combinations which work equally well under any strategic orientation and in any environmental context, at least for the variables under study. Yet, the results also indicate that certain resource combinations are contingent upon the degree of market demand uncertainty. It suggests that the remaining three ideal resource bundles are only applicable when market demand is uncertain. We can conclude that, for the constructs under study, some initial resource bundles are ideal in all strategic and environmental contexts while other resource bundles are only ideal in particular contexts. It must be noted that this result does not detract from the fit assumption (namely that a firm's resources, strategy and environmental conditions should be aligned to one another). It only suggests that there is no differential impact of the initial resource bundles on performance across the contingencies under consideration.

On the lower theoretical level, the decision tree analysis provides insight into the specific variables, variable levels and variable interactions that discriminate between financially superior and inferior start-ups. Since going into great detail is outside the scope of this study, we highlight some interesting findings. The first ideal configuration, for example, indicates that de novo businesses (i.e., real start-ups) do financially well when their managers adopt knowledge creating and sharing practices. Since all companies in this subsample adopted at least one practice, we can state that the management of knowledge streams is good for business. This

Table 2
Decision table corresponding to the tree induced by ALBA.

Follower	De novo			Follower									
Perc. Equity	-			≤3		>3							
KS Practices	≤10	>10		≤12	>12	-							
Uncertainty	-	≤1	>1	-	-	≤1	>1						
Nb founders	-	-	-	-	-	-	≤2						>2
Nb outside advisors	-	-	-	-	-	-	≤2			>2			-
Employee participation	-	-	-	-	-	-	≤2.8	>2.8		≤1.6	>1.6		-
Initial capital (log)	-	-	-	-	-	-	-	≤22.7	>22.7	-	-		-
Educational level employees	-	-	-	-	-	-	-	-	-	-	≤2.8	>2.8	-
Class = financially healthy	x	-	x	x	-	-	-	-	x	-	x	-	-
Class = financially unhealthy	-	x	-	-	x	x	x	x	-	x	-	x	x
	1	2	3	4	5	6	7	8	9	10	11	12	13

is in line with the knowledge-based view of the firm in which knowledge resources and more importantly knowledge-creating capabilities are considered the main sources of competitive advantage (Curado & Bontis, 2006). This is because knowledge-creating capabilities express the firm's orientation towards learning which is considered essential to keep track in today's fast-paced economy (Decarolis & Deeds, 1999). Furthermore, the tacit, hence firm-specific, knowledge which is supposed to emanate from internal learning processes can provide the firm with a difficult to imitate advantage. Yet the tree also indicates that adopting too many practices does not necessarily lead to success. Only when in combination with an uncertain market demand, do more than ten knowledge-enhancing practices pay off. This shows that the benefits of practices which stimulate the creation of knowledge are finite in environments with a certain market demand. Several authors have already indicated that the continual accumulation and renewal of knowledge is much more needed in uncertain environments as compared to certain environments because uncertainty entails unexpected problems and unforeseen situations which require the coordination of different packs of knowledge through information exchange and sharing to be properly dealt with (Bhatt, 2001). Bhatt (2001) writes that "*when an environment is dynamic, and complex, it often becomes essential for organizations that they continually create, validate, and apply new knowledge into their products, processes, and services for value-addition*". As such, our results corroborate prior theoretical assertions in the management literature regarding the role of knowledge for firm performance.

5. Conclusion

This empirical work is based on an extensive survey and data gathering process among Flemish organizations. The rule set extracted by the ALBA rule extraction technique, and corresponding decision table, provide an important insight into successful configurations of resources, combined with strategy and environment, for start-up companies. The results confirm the existence of multivariate configurations and equifinality.

We foresee further research to collaborate the findings set forth by the rule-based model in a wider setting by extending the data survey to other regions and time periods, and by qualitatively validating the different configurations observed in the model. Finally, we envision a practical use of this work by setting up entrepreneurial guidelines for setting up start-up companies.

Acknowledgments

The authors thank the Flemish Research Fund for financial support to the authors (post-doctoral research grant to David Martens; FWO project G.0599.05 of Christine Vanhoutte; Odysseus grant to Bart Baesens), and the Policy Research Centre for Entrepreneurship, Enterprises and Innovation for financial support.

References

Altman, E. (1993). *Corporate financial distress and bankruptcy*. New York: John Wiley & Sons.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society, 54*(6), 627–635.

Bamford, C., Dean, T., & Douglas, E. (2004). The temporal nature of growth determinants in new bank foundings: Implications for new venture research. *Journal of Business Venturing, 19*(6), 899–919.

Bamford, C., Dean, T., & McDougall, P. (2000). An examination of the impact of initial founding conditions and decisions upon the performance of new bank start-ups. *Journal of Business Venturing, 15*(3), 253–277.

Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of Management, 17*(1), 99–120.

Barney, J., & Clark, D. (2007). *Resource-based theory: Creating and sustaining competitive advantage*. Oxford University Press.

Bhatt, G. (2001). Knowledge management in organizations: Examining the interaction between technologies, techniques, and people. *Journal of Knowledge Management, 5*(1), 68–75.

Black, J., & Boal, K. (1994). Strategic resources: Traits, configurations and paths to sustainable competitive advantage. *Strategic Management Journal, 15*, 131–148.

Boeker, W. (1989). Strategic change: The effects of founding and history. *Strategic Management Journal, 32*(3), 489–515.

Chandler, G., & Hanks, S. (1998). An examination of the substitutability of founders human and financial capital in emerging business ventures. *Journal of Business Venturing, 13*, 353–369.

Cohn, D., Atlas, L., & Ladner, R. (1994). Improving generalization with active learning. *Machine Learning, 15*(2), 201–221.

Cooper, A., Gimeno-Gascon, F., & Woo, C. (1994). Initial human and financial capital as predictors of new venture performance. *Journal of Business Venturing, 9*(5), 371–396.

Curado, C., & Bontis, N. (2006). The knowledge-based view of the firm and its theoretical precursor. *International Journal of Learning and Intellectual Capital, 3*(4), 367–381.

Davidsson, P., Lindmark, L., & Olofsson, C. (2006). *Ch. Smallness, newness and regional development*. Cheltenham: Edward Elgar (pp. 499–513).

Decarolis, D., & Deeds, D. (1999). The impact of stocks and flows of organizational knowledge on firm performance: An empirical investigation of the biotechnology industry. *Strategic Management Journal, 20*(10), 953968.

Dietterich, T. G. (1998). Approximate statistical test for comparing supervised classification learning algorithms. *Neural Computation, 10*(7), 1895–1923.

Halle, B. (2001). *Business rules applied: Building better systems using the business rules approach*. New York: Wiley.

Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., & Baesens, B. (in press). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*.

Lessmann, S., Baesens, B., Mues, C., & Pietsch, S. (2008). Benchmarking classification models for software defect prediction: A proposed framework and novel findings. *IEEE Transactions Software Engineering, 34*(4), 485–496.

Martens, D., Baesens, B., Van Gestel, T., & Vanthienen, J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research, 183*(3), 1466–1476.

Martens, D., Bruynseels, L., Baesens, B., Willekens, M., & Vanthienen, J. (2008). Predicting going concern opinion with data mining. *Decision Support Systems, 45*, 765–777.

Martens, D., De Backer, M., Haesen, R., Baesens, B., Mues, C., & Vanthienen, J. (2006). Ant-based approach to the knowledge fusion problem. In *Proceedings of the fifth international workshop on ant colony optimization and swarm intelligence*. *Lecture Notes in Computer Science* (pp. 85–96). Springer.

Martens, D., De Backer, M., Haesen, R., Snoeck, M., Vanthienen, J., & Baesens, B. (2007). Classification with ant colony optimization. *IEEE Transaction on Evolutionary Computation, 11*(5), 651–665.

Martens, D., Van Gestel, T., & Baesens, B. (2009). Decompositional rule extraction from support vector machines by active learning. *IEEE Transactions on Knowledge and Data Engineering, 21*(2), 178–191.

Murphy, G., Trailer, J., & Hill, R. (1996). Measuring performance in entrepreneurship research. *Journal of Business Research, 36*(1), 15–23.

Pompe, P., & Bilderbeek, J. (2005). The prediction of bankruptcy of small- and medium-sized industrial firms. *Journal of Business Venturing, 20*(6), 847–868.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Boston, MA: Addison Wesley.

Van Gestel, T., Suykens, J., Baesens, B., Viaene, S., Vanthienen, J., Dedene, G., et al. (2004). Benchmarking least squares support vector machine classifiers. *Machine Learning, 54*(1), 5–32.

Vanthienen, J., Mues, C., & Aerts, A. (1998). An illustration of verification and validation in the modelling phase of KBS development. *Data and Knowledge Engineering, 27*(3), 337–352.

Vanthienen, J., Mues, C., & Aerts, A. (1998). An illustration of verification and validation in the modelling phase of kbs development. *Data and Knowledge Engineering, 27*, 337–352.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York, NY, USA: Springer-Verlag New York, Inc.

Verbeke, W., Baesens, B., Martens, D., De Backer, M., & Haesen, R. (2009). Including domain knowledge in customer churn prediction using antminer+. In *9th industrial conference on data mining ICDM, workshop on data mining in marketing DMM, Leipzig, Germany, July 22*.

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications, 38*, 2354–2364.

Wernerfelt, B. (1984). A resource-based view of the firm. *Strategic Management Journal, 5*(2), 171–180.

Witten, I. H., & Frank, E. (2000). *Data mining: Practical machine learning tools and techniques with Java implementations*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.