CLASSIFICATION OF ENTREPRENEURIAL INTENTIONS BY NEURAL NETWORKS, DECISION TREES AND SUPPORT VECTOR MACHINES

Marijana Zekić-Sušac
University of J.J. Strossmayer in Osijek
Faculty of Economics
Gajev trg 7, 31000 Osijek, Croatia
Phone: ++ 385 31 224 456; E-mail: marijana@efos.hr

Sanja Pfeifer
University of J.J. Strossmayer in Osijek
Faculty of Economics
Gajev trg 7, 31000 Osijek, Croatia
Phone: ++ 385 31 224 442; E-mail: pfeifer@efos.hr

Ivana Đurđević
University of J.J. Strossmayer in Osijek
Faculty of Teacher Education
L. Jägera 12, 31000 Osijek, Croatia
Phone: ++ 385 31 200 602; E-mail: idjurdjevic@ufos.hr

Abstract

Entrepreneurial intentions of students are important to recognize during the study in order to provide those students with educational background that will support such intentions and lead them to successful entrepreneurship after the study. The paper aims to develop a model that will classify students according to their entrepreneurial intentions by benchmarking three machine learning classifiers: neural networks, decision trees, and support vector machines. A survey was conducted at a Croatian university including a sample of students at the first year of study. Input variables described students' demographics, importance of business objectives, perception of entrepreneurial carrier, and entrepreneurial predispositions. Due to a large dimension of input space, a feature selection method was used in the pre-processing stage. For comparison reasons, all tested models were validated on the same out-of-sample dataset, and a cross-validation procedure for testing generalization ability of the models was conducted. The models were compared according to its classification accuracy, as well according to input variable importance. The results show that although the best neural network model produced the highest average hit rate, the difference in performance is not statistically significant. All three models also extract similar set of features relevant for classifying students, which can be suggested to be taken into consideration by universities while designing their academic programs.

Key words: *classification, entrepreneurial intentions, decision trees, neural networks, support vector machines*

1. INTRODUCTION

Following the statment of Smith (1990) that the the role of university is to foster creativity and responsiveness to change, Grigg (1994) has shown that universities need to adopt an entrepreneurial approach if they are to fulfil their mission and maintain both intellectual and social role in the society. Identifying entrepreneurial intentions could help universities and policy makers to create an environment that will stimulate such intentions, rather than neglecting them.

The paper aims to develop a model that will classify students according to their entrepreneurial intentions by benchmarking three machine learning classifiers: neural networks (NNs), classification and regression decision trees (CART), and support vector machines (SVM). The purpose of the model is to recognize entrepreneurial intentions of students, as well as to investigate the important predictors of such intentions. The next section contains the overview of previous research, while the methodology of NNs, CART, and SVM is described in section 3, following by the description of data and the experiments. Finally, the results of individual models and comparative results are reported and discussed.

2. OVERVIEW OF PREVIOUS RESEARCH

A number of researchers were focused on the factors which determine career choice of students and the factors that explain inconsistency between attitudes and intentions. Luethje and Franke (2003) tested a covariance structure model on the sample of engineering students. It was shown that entrepreneurial attitude of students, as well as perceived barriers and support factors in the entrepreneurship-related context are strongly linked with the intention to start a new venture. However, students' personality was not found directly connected with entrepreneurial intentions. Krueger et al. (2000) tested a competitive approach of the two most popular intention models based on the Ajzen's theory (describing perceptions of personal attractiveness, social norms, and feasibiliy), and Shapero's theory (variables describing perceptions of personal desirability, feasibility, and propensity to act). They found strong statistical support for both models. For these reason, in this paper we combine the two models in the choice of input variables. Regarding methodology used in previous research, it can be noticed that most of the authors in the area of entrepreneurial intentions used multiple regression and structural modelling. Machine learning methods have not been used in this area, although they were frequently tested in other problem domains. Lin (2006) used a fuzzy NN to test the influences on entrepreneurial-behavioral trends of environmental uncertainties, decision styles and inter-organizational relations. NNs outperformed discriminant analysis (St.John et al., 2000) in categorizing firms acccording to wealth creation measured as market value added (MVA). SVMs were also compared to NNs in financial failures, machine fault detection (Yeh et al., 2010; Shin et al., 2005), medicine etc. In addition to NNs and SVM, decision trees are a method that is frequently used in classification. Lee

(2010) used decision trees in conjuction with data envelopment analysis (DEA) for efficiency analysis of B2C controls.

3. METHODOLOGY

Artificial neural networks (NNs) were successfully used for classification, prediction, and association in different problem domains (Paliwal and Kumar, 2009). The main advantage of NNs is the ability to approximate any nonlinear mathematical function (Masters, 1995). The most common type of NN was tested in this research - the multilayer perceptron (MLP), a feed forward network that is able to use various algorithms to minimize the objective function, such as backpropagation, conjugate gradient, and other algorithms. The backpropagation algorithm is based on deterministic gradient descent algorithm originally developed by Werbos in 1974, extended by Rumelhart et al. (in Masters, 1995).

The input layer of a NN consists of n input units $x_i \in R$, $i=1,2,...,\ n$, and randomly determined initial weights w_i usually from the interval [-1,1]. Each unit in the hidden (middle) layer receives the weighted sum of all x_i values as the input. The output of the hidden layer denoted as y_c is computed by:

$$y_c = f\left(\sum_{i=1}^{n} w_i x_i\right) \tag{1}$$

where f is the activation function selected by the user, which can be logistic, tangent hyperbolic, exponential, linear, step or other (Masters, 1995). The computed output is compared to the actual output y_a, and the local error ε is computed. The error is then used to adjust the weights of the input vector according to a learning rule, usually the Delta rule (Masters, 1995). The above process is repeated in a number of iterations (epochs), where the gradient descent or other algorithm is used to minimize the error. In order to produce probabilities in the output layer, a softmax activation function is added for classification purposes. The output layer of all NN models in our experiments consisted of a binary variable (valued as 1 for the existence of entrepreneurship intention, and 0 for the absence of entrepreneurship intention). The backpropagation and conjugate gradient algorithms were tested, by varying the activation function in the hidden layer (sigmoid, tangens hyperbolic, exponential, and linear). The number of hidden units varied from 2 to 40. The NN structure and training time was optimized by a split-sample procedure. The maximum number of training epochs was set to 1000.

Decision trees i.e. classification trees are frequently used in datamining, due to its ability to find hidden relationships among data. Benchmarking NNs to decision trees is also present in previous research (Bensic et al., 2005; Lee, 2010). The aim of this method is to build a binary tree by splitting the input vectors at each node according to a function of a single input. The two algorithms are the most popular for building a

decision tree: discriminant-based univariate splits, and classification and regression trees (CART or C&RT). CART algorithm was pioneered in 1984 by Breiman et al. (in Witten and Frank, 2000). Questier et al. (2005) summarized CART steps as: (1) assign all objects to root node, (2) split each explanatory variable at all possible split points, (3) for each split point, split the parent node into two child nodes by separating the objects with values lower and higher than the split point for the considered explanatory variable, (4) select the variable and split point with the highest reduction of impurity, (5) perform the split of the parent node into the two child nodes according to the selected split point, (6) repeat steps 2–5, using each node as a new parent node, until the tree has maximum size, and (7) prune the tree back using cross-validation to select the right-sized tree. The evaluation function used in this research for splitting is the Gini index defined as (Apte, 1997):

$$Gini(t) = 1 - \sum_i p_i^2 \tag{2}$$

where t is a current node and p_i is the probability of class i in t. The CART algorithm considers all possible splits in order to find the best one by Gini index. The C&RT style exhaustive search for univariate splits was used in our experiments, with Gini index, equal prior probabilities, and equal missclassification costs. Prune of missclassification error was used as the stopping rule, with minimum n=5, and *standard error rule*=1. The 10-fold CV procedure was used during the training phase in order to find the right-sized tree with the minimal CV cost.

Support vector machine (SVM) is a machine learning method aimed to be used for non-linear mapping of the input vectors into a high-dimensional feature space. It produces a binary classifier, so-called optimal separating hyperplanes, and results in a uniquely global optimum, high generalization performance, and does not suffer from a local optima problem (Behzad et al., 2009). The basic principle of learning in SVM is that it searches for an optimal hyperplane which satisfies the request of classification, then uses an algorithm to make the margin of the separation beside the optimal hyperplane maximum while ensuring the accuracy of correct classification (Yeh et al. 2010). The principle of SVM can be described as follows. Suppose we are given a set of training data $x_i \in R^n$ with the desired output $y_i \in \{+1,-1\}$ corresponding with the two classes, and assume there is a a separating hyperplane with the target functions $w \cdot x_i + b = 0$, where w is the weight vector, and b is a bias. We want to choose w and b to maximize the margin or distance between the parallel hyperplanes that are as far apart as possible while still separating the data. In the case of linear separation, the linear SVM for optimal separating hyperplane has the following optimization problem (Yeh et al. 2010):

$$\text{Minimize } \phi(w) = \frac{1}{2} w^T w \tag{3}$$

$$\text{subject to } y_i(x_i \cdot w + b) \geq 1, \text{ i=1,2,...,n.} \tag{4}$$

The solution to above optimization problem can be converted into its dual problem. The non-negative Lagrange multipliers can be searched by solving the following optimization problem if the problem is nonlinear:

$$\text{Maximize } Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i x_j) \tag{5}$$

$$\text{subject to } \sum_{i=1}^{n} \alpha_i y_i = 0,\ 0 \le \alpha_i \le C,\ \text{i= i=1,2,...,n.} \tag{6}$$

where C is the nonnegative parameter chosen by users. The final classification function is:

$$f(x) = \text{sgn}\left\{ \sum_{i=1}^{n} \alpha_i^* y_i K(x_i, x_j) + b^* \right\} \tag{7}$$

where K is a kernel function, which can be linear, sigmoid, RBF or polynomial.

SVMs are able to select a small and most proper subset of data pairs (support vectors). Since their performance depends mostly on the choice of kernel function and hyper parameters, a cross-validatin procedure is used as a successfull tool for adjusting those parameters (Min and Lee, 2005; Behzad et al., 2009). Linear, polynomial, RBF, and exponential kernels were used, where gamma coefficient for polynomial and RBF kernel was 0.0625, degree was 3, coefficient varied from 0 to 0.1, c=10.

Each of the three methods was tested on two sets of input variables: Model 1 - all available data, and Model 2 - variables selected by a feature selection procedure based on the *chi-square* statistics and *p* value for each predictor variable ($p<0.05$ was used as the criterion for selecting important variables in our experiments). After the training phase, all three methods were tested on the same out-of-sample data (i.e. the test set). The performance of all models is measured by the hit rate of class 0 (i.e. the "lack of entrepreneurial intentions" - hit_0, or “negative hit rate”)), hit rate of class 1 (i.e. the "existence of entrepreneurial intentions" - hit_1 or “positive hit rate”), and the average hit rate (*ave hit*) according to:

$$hit_0 = \frac{c_0}{t_0},\ hit_1 = \frac{c_1}{t_1},\ ave\ hit = \frac{hit_0 + hit_1}{2} \tag{8}$$

where c_0 is the number of students accurately predicted to have output 0, t_0 is the number of students with actual (target) 0 output, c_1 is the number of students accurately predicted to have output 1, and t_1 is the number of students with actual output 1. The sensitivity and specificity ratios were computed according to (Simon and Boring, 1990):

$$sensitivity = \frac{c_1}{(c_1 + d_0)}\ ,\quad specificity = \frac{c_0}{(c_0 + d_1)} \tag{9}$$

where d_0 is the number of false negatives (the number of students falsely predicted to have output 0), and d_1 is the number of false positives (the number of students falsely predicted to have output 1). The type I and type II errors were calculated in order to compare the cost of misclassification produced by three models.

4. DATA

The total dataset consisted of 237 regular students of business administration at the first year of study at University of J.J. Strossmayer in Osijek, Croatia. The survey was conducted at July 2010. The choice of input variables in this paper was lead by two theoretically recognized intention-based models: Ajzen's theory of planned behavior (TPB) and Shapero's model of the entrepreneurial event (SEE) (Krueger, 2000). Therefore, besides students' demographics, other predictors were used and grouped into the three main components: entrepreneurial outcome expectation, social norms and beliefs on entrepreneurship and entrepreneurial career, and predispositions (entrepreneurial self efficacy). The total number of 46 input variables was used in Model 1, while the Model 2 consisted of 17 variables selected by a preprocessing procedure. The four input variables were categorical, while the rest of them were continuous valued from 1 to 5. For the purposes of model training and testing, the dataset is divided into the train and test sample in the CART and SVM models, while the train set is further divided into the train and selection subsample in the NN models. To ensure equal prior probabilities of both output classes, a stratified sampling is used, to preserve equal distribution of students with positive and negative intentions in the train sample. The structure of samples is presented in Table 1.

Table 1: Sample structure.

Subsample	1 (existence of intentions)	0 (lack of intentions)	Total	%
Train and selection	69	69	138	58.23
Test	82	17	99	41.77
Total	151	86	237	100.00

5. RESULTS

It can be seen from Table 2 that the best model (with the highest average hit rate) is the MLP neural network Model 2, which uses the logistic activation function and produces the average hit rate of 74.43%. The RBF network also works better with Model 2 than with Model 1 yielding the average hit rate of 72%. CART decision tree shows a stability of results across models, and the average hit rate of 71.27% is obtained for Model1 and Model2. The selected decision tree had only two splits, and the variable "Family prior entrepreneurial experience" was the only splitting variable. The split category is 1, meaning that all students in the train set that have entrepreneurial history (i.e. entrepreneurs in the family) were classified to have positive entrepreneurial intentions, while the others were classified as not having those intentions.

The best SVM model is Model 2 with polynomial kernel, obtained by the following parameters: degree=3.0, gamma= 0.0625, coefficient=0. The total number of extracted support vectors was 86, where 43 were used for class 0 and 43 for class 1. It can be noticed that the results of the SVM model are lower than the results of

the models produced by other two methodologies, and that the accuracy depends on the choice of the kernel function. Polynomial kernel produced the highest average hit rate of all SVM kernels (68.33%).

Table 2: Results of the models obtained by NNs, CART and SVM on the test sample.

Model	Model 1 – all input variables			Model 2 – selected variables		
	Hit 0 (%)	Hit 1 (%)	Ave, hit (%)	Hit 0 (%)	Hit 1 (%)	Ave, hit (%)
MLP – logistic	52.00	88.00	70.00	84.15	64.71	74.43*
MLP – tangent hyperbolic	70.00	70.00	70.00	46.00	70.00	58.00
MLP – exponential	51.00	82.00	66.50	59.00	76.00	67.50
MLP – linear activation function	68.00	70.00	69.00	42.00	94.00	68.00
RBF network	0.00	100.00	50.00	68.00	76.00	72.00
CART decision tree	71.95	70.59	71.27	71.95	70.59	71.27
SVM linear kernel	63.41	52.94	58.18	59.76	64.71	62.23
SVM RBF kernel	57.32	64.71	61.01	63.41	64.71	64.06
SVM sigmoid kernel	60.98	70.59	65.78	63.41	70.59	67.00
SVM polynomial kernel	71.95	64.71	68.33	71.95	64.71	68.33

* The most accurate model

In order to compare the classification accuracy of the models, the test of differences in proportions (t-test) was conducted and its results are presented in Table 3.

Table 3: Statistical comparison of the NN, CART, and SVM models.

Hypothesis	Model	Average hit rates	t-test results
H0: NN=CART	NN	74.43	p=0.3089
	CART	71.27	
H0: NN=SVM	NN	74.43	p=0.1718
	SVM	68.33	
H0:CART=SVM	CART	71.27	p=0.3264
	SVM	68.33	

The test of differences in proportions shows that the average hit rate of the MLP network is not significantly different from the average hit rate of the CART model, as well from the average hit rate of the SVM model. Based on this test, it cannot be concluded that any of the three models outperforms the other two in prediction accuracy. However, the suggestions of the best model for the observed dataset could be directed to other criteria, such as type I and type II errors (or sensitivity and specificity). Table 4 presents the sensitivity and specificity (at the diagonal of each matrix), as well as type I and type II errors of the best NN, CART, and SVM models.

Table 4: The sensitivity and specificity of the best NN, CART, and SVM models.

Actual output	Output predicted by the NN model		Output predicted by the CART model		Output predicted by the SVM model	
	1	0	1	0	1	0
1 (positive)	0.65	0.35	0.71	0.29	0.65	0.35
0 (negative)	0.16	0.84	0.28	0.72	0.28	0.72

It can be seen from Table 4 that specificity of the NN model is 0.65, while its sensitivity is 0.84. The false positive rate (i.e. the type I error) of the NN model is 0.35, while the false negative rate (i.e. the type II error) is 0.16. It reveals that the NN model is more sensitive than specific, tending to misclassify more students that actually had no entrepreneurial intentions into the category of positive ones with entrepreneurial intentions. When the same coefficients of the CART model are observed, it can be noticed that the specificity of this model is the highest (0.71) among all three models, while the sensitivity of CART is the same as the sensitivity of the SVM (0.72), which is lower than the sensitivity of the NN model. The lowest false positive hit rate is produced by both NN and SVM models (0.35), while the lowest false negative hit rate is obtained by the NN model. The CART model has the smallest difference between type I and type II errors, implying that this model is able to balance between sensitivity and specificity. In order to get better insight into the importance of input variables in predicting student entrepreneurial intentions, the sensitivity analysis was performed on the best NN model by using the out-of-sample data. The most important variable is "Family prior entrepreneurial experience" meaning if there is an entrepreneur in the family, student intentions are likely to be positive. Other highly ranked variables by the NN model are "Prior working exposure", "Understanding for entrepreneurial way of thinking", "Perceived control", and other variables mostly from the group of "Entrepreneurial predispositions and entrepreneurial self efficacy". Although the ranking by CART model varies than the one by the NN model, some predictors were found highly important by both models, such as "Socializing with entrepreneurial persons", "Prior working exposure", "People think I am a person with unusual/unconventional ideas", and "Understanding for entrepreneurial way of thinking".

4. CONCLUSION

This paper aims to analyze the performance of three machine learning methods in modelling entprepreneurial intentions of students based on demographic variables, entrepreneurial outcome expectation, social norms and beliefs on entrepreneurship and entrepreneurial career, and entrepreneurial predispositions (entrepreneurial self efficacy). The multilayer perceptron neural networks, CART decision trees, and support vector machines with different kernel functions were trained and tested on the same out-of-sample data by using all available input space, as well as preselected set of variables. Although the model with the highest accuracy in classification is the NN model, its average hit rate is not significantly different from the hit rates of the other two models. The analysis shows that the NN model shows highest sensitivity (produces the lowest type II error), while the CART model is the most specific one (produces the lowest type I error) in recognizing positive entrepreneurial intentions. The NN model with a high sensitivity could be used for screening for the students with intentions, since it has tendency to misclassify more students into the group of positive ones. The CART model with a high specificity could be used for confirming the test results, since it is more specific in recognizing the actual positive students. The results are limited for the observed dataset, and further research will include testing the methodology on more dataset in order to generalize the conclusions.

Acknowledgements
This work was supported by the Ministry of Science, Education and Sports, Republic of Croatia, through research grant 010-0101195-0872, "Transformation of entrepreneurial potential into entrepreneurial behavior."

REFERENCES

1. Apte, C. and S. Weiss (1997), "Data Mining with Decision Trees and Decision Rules", *Future Generation Computer Systems* , Vol. 13, pp. 197-210.
2. Behzad, M., Asghar, K., Eazi, M. and M. Palhang, (2009), „Generalization performance of support vector machines and neural networks in runoff modeling", *Expert Systems with Applications*, Vol. 36, pp. 7624–7629.
3. Bensic, M., Sarlija, N. and M. Zekic-Susac (2005), „Modeling Small Business Credit Scoring Using Logistic Regression, Neural Networks, and Decision Trees, *Intelligent Systems in Accounting, Finance and Management*, Vol 13, No. 3, pp. 133-150.
4. Haykin, S.(1999), "*Neural Networks: A Comprehensive Foundation*", Prentice Hall International, Inc., New Jersey, USA.
5. Krueger, N.F. JR., Reilly, M.D. and A.L. Carsrud (2000), "Competing Models of Entrepreneurial Intentions", *Journal of Business Venturing,* Vol. 15, pp. 411–432.
6. Lee, S.(2010), "Using data envelopment analysis and decision trees for efficiency analysis and recommendation of B2C controls", *Decision Support Systems*, Vol.49, pp. 486–497
7. Lin, W.B. (2006), „A comparative study on the trends of entrepreneurial behaviors of enterprises in different strategies: Application of the social cognition theory", *Expert Systems with Applications*, Vol. 31, pp. 207–220.
8. Luethje, C. and N. Franke (2003), "The 'Making' of an Entrepreneur: Testing a Model of Entrepreneurial Intent Among Engineering Students at MIT", *R&D Management*, Vol. 33, No. 2, pp. 135-147.
9. Masters, T. (1995), *Advanced Algorithms for Neural Networks, A C++ Sourcebook,* John Wiley & Sons, Inc., New York, USA.
10. Paliwal, M. and U.A. Kumar (2009), "Neural networks and statistical techniques: A review of applications", *Expert Systems with Applications*, Vol. 36, pp. 2–17.
11. Questier, F., Put, R., Coomans, D., Walczak, B. and Y. Vander Heyden (2005), "The use of CART and multivariate regression trees for supervised and unsupervised feature selection", *Chemometrics and Intelligent Laboratory Systems*, Vol. 76, pp. 45-54.
12. Shin, H.J., Eom, D.-H. and S.-S. Kim (2005), „One-class support vector machines - an application in machine fault detection and classification", *Computers & Industrial Engineering*, Vol. 48, pp. 395–408.
13. Shin, K.-S, et al.(2005), "An application of support vector machines in bankruptcy prediction model", *Expert Systems with Applications*, Vol. 28, pp. 127–135.

14. Simon, D. and J.R. Boring III (1990), "Sensitivity, Specificity, and Predictive Value", *Clinical Methods: The History, Physical, and Laboratory Examinations*, Butterworths, Boston, pp. 49-54.
15. Smith, R. (1990), „The modern Australian university: Challenges for leadership and management", *Journal of Tertiary Educational Administration*, Vol. 12, No. I, pp.: 243-254.
16. St. John, C.H., Balakrishnan, N. and J.O. Fiet (2000.), „Modeling the relationship between corporate strategy and wealth creation using neural networks", *Computers & Operations Research*, Vol. 27, pp. 1077-1092.
17. Witten, I.H. and E. Frank (2000), *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation*, Morgan Kaufman Publishers, San Francisco.
18. Yeh, C.-C, et al (2010), „A hybrid approach of DEA, rough set and support vector machines for business failure prediction", *Expert Systems with Applications*, Vol. 37, pp. 1535–1541.