DE GRUYTER OPEN

A Comparison of Machine Learning Methods in a High-Dimensional Classification Problem

Marijana Zekić-Sušac, Sanja Pfeifer, Nataša Šarlija
University of Josip Juraj Strossmayer in Osijek, Faculty of Economics, Croatia

Abstract

Background: Large-dimensional data modelling often relies on variable reduction methods in the pre-processing and in the post-processing stage. However, such a reduction usually provides less information and yields a lower accuracy of the model. **Objectives:** The aim of this paper is to assess the high-dimensional classification problem of recognizing entrepreneurial intentions of students by machine learning methods. **Methods/Approach:** Four methods were tested: artificial neural networks, CART classification trees, support vector machines, and k-nearest neighbour on the same dataset in order to compare their efficiency in the sense of classification accuracy. The performance of each method was compared on ten subsamples in a 10-fold cross-validation procedure in order to assess computing sensitivity and specificity of each model. **Results:** The artificial neural network model based on multilayer perceptron yielded a higher classification rate than the models produced by other methods. The pairwise t-test showed a statistical significance between the artificial neural network and the k-nearest neighbour model, while the difference among other methods was not statistically significant. **Conclusions:** Tested machine learning methods are able to learn fast and achieve high classification accuracy. However, further advancement can be assured by testing a few additional methodological refinements in machine learning methods.

Keywords: machine learning; support vector machines; artificial neural networks; CART classification trees; k-nearest neighbour; large-dimensional data; cross-validation

JEL main category: C
JEL classification: C45; C55; L26
Paper type: Research article

Received: 15 December, 2013
Accepted: 18 May, 2014

Citation: Zekić Sušac, M., Pfeifer, S., Šarlija, N. (2014), "A Comparison of Machine Learning Methods in a High-Dimensional Classification Problem", Business Systems Research, Vol. 5 No. 3, pp. 82-96.
DOI: 10.2478/bsrj-2014-0021

Introduction

Usually the problem of large-dimensional data modelling has been solved by variable reduction methods in the pre-processing and in the post-processing stage.

Methods such as t-test, Cronbach's alpha, chi-square, principal component analysis (PCA), genetic algorithms, and others are able to reduce the dimension of input vector (Paliwal and Kumar, 2009). However, such reduction usually provides less information and yields a lower accuracy of the model. Based on a previous research (Zekić-Sušac et al, 2012), it was found that such situation exists in a dataset collected within an international survey on entrepreneurship self-efficacy and identity. Based on proven instruments which measure certain attributes of students, such as their motivation, social norms, self-efficacy, and other factors which influence entrepreneurial intentions according to a conceptual framework given by researchers in the area of entrepreneurship (Kolvereid and Isaksen, 2006; Thompson, 2009; Krueger, 2000), a large number of input variables is used to provide a basis for finding an efficient model that will be able to classify students according to their entrepreneurial intentions.

Our previous investigations (Zekić-Sušac et al., 2012) showed that feature selection methods based on Cronbach's alpha and PCA produced models with lower accuracy than the model that used all available input space. Also, it was found that non-linear machine learning methods such as artificial neural networks (ANNs) could be efficient in the area of modeling entrepreneurial intentions of students (Zekić-Sušac et al., 2010). In this research, a multilayer perceptron neural network with a softmax activation function in the output layer is used to classify students into one of the two categories: 1 - students with entrepreneurial intention, and 0 – students with no entrepreneurial intention.

The purpose of this paper is to compare the accuracy of ANNs to the accuracy of other machine learning methods, such as support vector machines (SVMs), decision trees, and k-nearest neighbour in a classification type of problem with a large number of variables. Majority of social phenomena including entrepreneurial career choice require taking into account datasets with a huge number of predictors that can interact on a variety of levels and directions. Therefore, this paper contributes to the variety of stakeholders interested in social phenomena such as: researchers, policy makers, academic stuff and practitioners and enable them to use alternative methods for reducing the number of predictors or constructs relevant models for particular phenomena.

The paper starts with an overview of previous research in this area, explains the methodology used in experiments, describes the main results followed by discussion and conclusion.

Theoretical Background

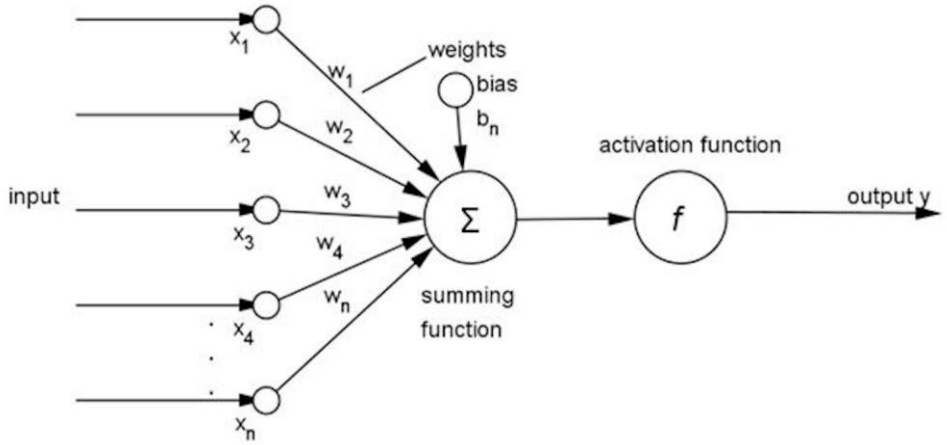
Theoretical and methodological background of the paper is focused on machine learning methods that were successfully used for classification, such as artificial neural networks, decision trees, support vector machines and k-nearest neighbour.

Artificial neural networks

Artificial neural networks (ANNs) have been successfully used for classification, prediction, and association in different problem domains (Paliwal and Kumar, 2009). ANNs have the ability to approximate any nonlinear mathematical function, which is useful especially when the relationship between the variables is not known or is complex (Masters, 1995). However, there are some limitations of ANNs such as time-consuming experimentation needed to determine network structure and learning parameters, and a lack of interpretability of the weights obtained during the model building process. The most common type of ANN was tested in this research - the multilayer perceptron (MLP), a feed forward network that can use various algorithms

to minimize the objective function, such as backpropagation, conjugate gradient, and other. A simplified architecture of a MLP ANN is presented in Figure 1.

Figure 1
Architecture of the MLP network



Source: modified from Haykin, 1999.

The input layer of an ANN consists of *n* input units with values $x_i \in R$, *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 summing the inputs multiplied with their weights, according to:

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

where f is the activation function selected by the user (sigmoid, 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 according to:

$$\Delta w_i = \eta \cdot y_c \cdot \varepsilon \quad (2)$$

where Δw_i is the weight adjustment, η is the learning parameter that could be experimentally determined. The above process is repeated in a number of iterations (epochs), where the three different algorithms were tested to minimize the error: gradient descent, conjugate gradient descent, and Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm. Conjugate gradient descent is faster than gradient descent and performs a series of line searches through error space, therefore avoiding a local minimum. BFGS belongs to the second-order algorithms with very fast convergence but memory intensiveness due to storing the Hessian matrix (Dai, 2002). In order to produce probabilities in the output layer, a softmax activation function is added. The output layer of all ANN 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 number of hidden units varied from 2 to 20, and the training time is determined in an early-stopping procedure

which iteratively trains and tests the networks on a separate test sample in a number of cycles, and saves the network which produces the lowest error on the test sample.

Support vector machines

Support vector machine (SVM) is aimed to be used for non-linear mapping of the input vectors into a high-dimensional feature space. 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). 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 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 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):

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

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

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:

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

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

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

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

where *K* is a kernel function, which can be linear, sigmoid, RBF or polynomial. Since the successfulness of SVM depends on the choice of kernel function K and hyper parameters, a cross-validation procedure should be performed for adjusting those parameters (Min and Lee, 2005; Behzad et al., 2009). Linear, polynomial, RBF, and exponential kernels were used in our experiments, where gamma coefficient for polynomial and RBF kernel was 0.0625, degree was 3, coefficient varied from 0 to 0.1, c=10. The advantage of SVMs is that they are able to select a small and most proper subset of data pairs (support vectors) (Yu et al., 2003).

Decision trees

Decision trees i.e. classification trees are frequently used in data mining, 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):

$$\text{Gini}(t) = 1 - \sum_i p_i^2 \tag{8}$$

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 misclassification costs. Prune of misclassification 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.

K-nearest neighbour technique

The k-nearest neighbour technique (KNN) is aimed to classify the outcome of a query point based on a selected number of its nearest neighbours. It can be used for both classification and regression types of problems. For a given query point, the method estimates the outcome by finding k examples that are closest in distance to the query point (i.e. its neighbours). For regression problems, predictions are based on averaging the outcomes of the k nearest neighbours, while for classification problems, it uses a majority of voting. In estimating the model, it is important to select the appropriate value of k, because it can affect the quality of predictions such that a small value of k will lead to a large variance in predictions, while a large value of k may lead to a large model bias. One way to select the optimal value of k is to use cross-validation procedure to smooth the k parameter, i.e. to find the value of k that is the optimal trade off (Bishop, 1995).In order to find the neighbours of a point, a distance metrics needs to be used. The most common is the Euclidean, while others possible metrics are Euclidean squared, City-block, and Chebychev distances. In this paper, the Euclidean distance is used according to (Bishop, 1995):

$$D(x-p) = \sqrt{(x-p)^2} \tag{9}$$

where x is a query point, and p is a case from the sample. A popular approach to improve the prediction is to use distance weighting which uses large values of k with

more importance given to cases closest to the query point. A set of weights *w* is introduced, one for each nearest neighbour, where *w* denotes the relative closeness of each neighbour with respect to the query point. Weights are computed according to (Bishop, 1995):

$$W(x,p_i)=\frac{\exp(-D(x,p_i))}{\sum_{i=1}^{k}\exp(-D(x,p_i))} \quad (10)$$

where $D(x, p_i)$ is the distance between the query point x and the *i*th case p_i of the sample. All the weights sum to 1, and for the classification problems, the output of the case with the maximum weight is assigned as the output value to the query point x. In our experiments, an initial value k=10 was used, and it was optimized in a cross-validation procedure.

Previous research

Methodological tools for modeling entrepreneurial intentions mostly included multiple regression and structural modelling. Machine learning methods have not been investigated in this area, although they were frequently tested in other problem domains. Lin (2006) used a fuzzy neural network (NN) to test the influences on entrepreneurial-behavioral trends of environmental uncertainties, decision styles and inter-organizational relations. ANNs outperformed discriminant analysis (St.John et al., 2000) in categorizing firms according to wealth creation measured as market value added (MVA). Support vector machines (SVMs) were also compared to ANNs in financial failures, machine fault detection (Yeh et al., 2010; Shin et al., 2005), medicine etc. In addition to ANNs and SVMs, decision trees are a method that is frequently used in classification (Lee, 2010), as well as the k-nearest neighbour technique which has been used as a standard classification technique in many areas.

However, there is a lack of comparative studies of machine learning methods, especially in case with a large number of predictors combined with a relatively small sample size. One of the first papers that dealt with such comparison (Brown et al., 1993) investigated multi-modal classification problems by testing decision trees and backpropagation neural networks for emitter classification and digit recognition. They used two types of real-world problems: one with few features and a large data set; and the other one with many features and a small data set. The authors obtained that both methods produce comparable error rates but that direct application of either method will not necessarily produce the lowest error rate. They suggest multi-variable splits, feature selection, and node identification to improve the results. Kuzey et al. (2014) compared two machine learning methods: artificial neural networks and decision trees in investigating relative importance of factors as predictors of firm value. They used multinationality (measured by foreign sales ratio) and fourteen other financial indicators on firm value as input variables and ranked their importance by the sensitivity analysis based on information fusion. Their research shows that both methods extracted similar set of important predictors as important, but the accuracy of methods with a high-dimensional input space is still not investigated enough. Shao and Lunetta (2012) showed that SVMs had superior generalization capability over CART decision trees, particularly with respect to small training sample sizes. SVM also showed less variability when classification trials were repeated using different training sets. Bolivar-Cime and Marron (2013) analyzed

binary discrimination methods in dealing with high-dimensional data. They conducted comparison of SVM, mean difference (i.e. Centroid rule), distance weighted discrimination, maximal data piling, and naive Bayes methods in the high dimension low sample size context for Gaussian data with common diagonal covariance matrix. Their results show that, under appropriate conditions, the first four methods are asymptotically equivalent, while the Naive Bayes method can have a different asymptotic behavior when the number of variables tends to infinity.

Besides comparing the efficiency of different machine learning methods in classifying high-dimensional data, researchers were even more focused on improving the classification algorithms of SVM or ANNs to deal with a large number of variables. For example, Talukder and Casasent (2001) proposed a closed-form neural network for discriminatory feature extraction from high-dimensional data which provides more general nonlinear transforms of the input data and are suitable for cases involving high-dimensional (image) inputs where training data are limited and the classes are not linearly separable. Zanaty (2012) introduced a new kernel function for improving the accuracy of the Support Vector Machines (SVMs) classification called Gaussian Radial Basis Polynomials Function (GRPF) which was shown to be more effective than multi-layer neural networks in classifying high-dimensional data.

It can be summarized that a thorough comparison of machine learning methods classification ability with high-dimensional data is yet to be conducted. Since lots of real datasets share the characteristics of high dimensionality, bby taking into account that machine learning methods are robust and do not require rigorous statistical assumptions on predictor interdependencies, a comparison of their efficiency in solving a high-dimensional classification problem is potentially useful to researchers in the area of modeling and to practitioners in the area of recognizing entrepreneurial intentions.

Methodology

Data

The dataset for this research was collected in an international survey on entrepreneurial intentions at the summer semester 2010 and 2012. It consisted of 443 regular students of business administration at the first year of study at University of Osijek, Croatia. The survey design was based on the instruments tested in the previous research on entrepreneurial intentions. A number of researches confirmed reliability of the instruments that are valid for measuring students attitudes, values and career choices such as: (1) entrepreneurial intentions (Thompson, 2009), (2) altruistic values and empathy (Smith, 2009), (3) subjective norms (Kolvereid and Isaksen, 2006), (4) entrepreneurial self-efficacy (McGee et al., 2009), (4) allocentrism/idiocentrism (Triandis and Gelfand, 1998), (5) prior family business exposure (Carr and Sequeira, 2007), (6) entrepreneurial outcome expectations (Krueger, 2000), (7) strength of entrepreneur identity aspiration (Farmer et al. 2011), and (8) social entrepreneurship self-efficacy (Nga, 2010). Following these suggestions for measuring entrepreneurial intentions, a prospective researcher often need some suggestions how to solve high dimensional classification problems and how to construct models that will reduce the hundreds of variables to more operable number of variables. For the purpose of this study, the total number of 94 input variables was selected as the most relevant. The sample includes 48.76% of respondents with intentions to start a business, and 51.24% of them with no intentions to start a business. For the purposes of ANNs training and testing, the total dataset is

divided into three subsamples: train, test and validation subsample in the ANN models, while the SVM, CART and k-nearest neighbour models used the train and test sets together for analysis purposes and the validation sample for the final testing. The structure of samples is presented in Table 1.

Table 1
Sample Structure Used for the ANN, SVM, CART and k-nearest Neighbour Models

	ANN models	**SVM, CART, and k-nearest neighbour models**
Subsample	Total	%
Train	355	80.14
Test	44	9.93
Validation	44	9.93
Total	443	100.00

Source: Authors' work

For the purpose of testing the generalization ability of the models, 10 different datasets were randomly generated from the initial dataset by the 10-fold cross-validation (CV) procedure such that a different 44 cases of data is used for validation purpose. Each of the four classification methods were conducted on 10 datasets generated in the 10-fold CV procedure.

Modelling procedure

Each of the four methods was trained (estimated) and tested in the 10-fold cross-validation procedure such that each method uses the same subsets of data for training and testing in order to enable the comparison of results.

The performance of all models on each of the 10 test samples is measured by the total classification rate (i.e. the proportion of correctly classified cases in the test set), and a 10-fold cross-validation procedure for testing generalization ability of the models was conducted. The cross-validation procedure (or leave *k* cases out, where *k*=1/10 of the total sample) is used in this paper because it produces no statistical bias of the result since each tested sample is not the member of the training set (Liu et al., 2007). According to Witten and Frank (2000), extensive tests on numerous datasets have shown that 10 is sufficient value for *n* in the *n*-fold cross validation. After the 10-fold cross-validation procedure, the average of the total classification rate is computed, which is used to estimate the generalization error of a model. Also, the classification rate of class 0 (i.e. the "lack of entrepreneurial intentions" or "negative hit rate"), classification rate of class 1 (i.e. the "existence of entrepreneurial intentions" or "positive hit rate") were also observed in order to compute the sensitivity and specificity of the models. The sensitivity and specificity ratios were computed according to Simon and Boring (1990):

$$\text{sensitivity} = \frac{c_1}{(c_1 + d_0)}, \quad \text{specificity} = \frac{c_0}{(c_0 + d_1)} \tag{11}$$

where c_0 is the number of students accurately predicted to have output 0, c_1 is the number of students accurately predicted to have output 1, 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 error (α *=1-specificity)* and type II error (β *=1-sensitivity)* were calculated in order to compare the cost of misclassification produced by each of the models,

while the likelihood for positives and likelihood for negatives in classification is computed according to:

$$L_1 = \frac{\text{sensitivity}}{\alpha}, \quad L_0 = \frac{\text{specificity}}{\beta} \tag{12}$$

where L_1 is likelihood for the class of positive entrepreneurial intentions (class 1), while L_0 is the likelihood for the class of no entrepreneurial intentions (class 0).

Results

The results of the ANN model, CART model, SVM model, and KNN model are presented in Table 2, where the total classification rate of each model is computed as the proportion of correctly classified cases in the validation sample.

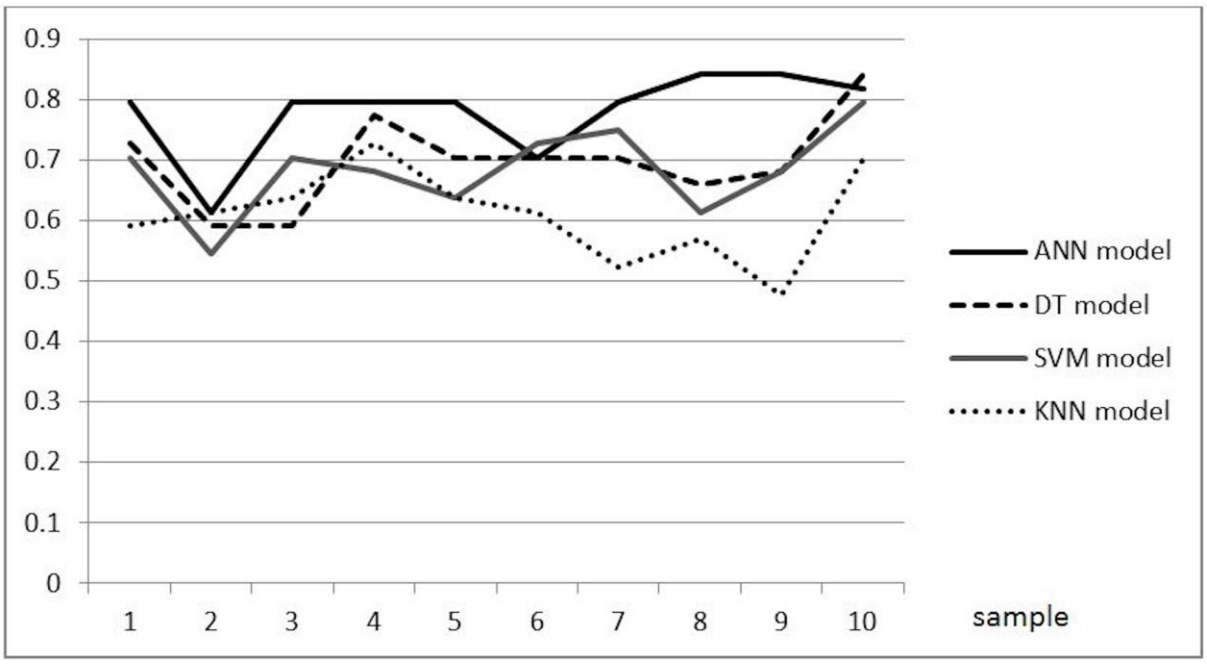
Table 2
Results of the 10-fold Cross-Validation Procedure

Sample	***Total classification rate***			
	ANN model	***CART model***	***SVM model***	***KNN model***
1	0.7955	0.7273	0.7045	0.5909
2	0.6136	0.5909	0.5455	0.6136
3	0.7955	0.5909	0.7045	0.6364
4	0.7955	0.7727	0.6818	0.7273
5	0.7955	0.7045	0.6364	0.6364
6	0.7045	0.7045	0.7273	0.6136
7	0.7955	0.7045	0.7500	0.5227
8	0.8409	0.6591	0.6136	0.5682
9	0.8421	0.6818	0.6818	0.4773
10	0.8182	0.8409	0.7955	0.7045
Average classification rate	0.7797	0.6977	0.6841	0.6091
Standard deviation of classification rates	0.0696	0.0758	0.0714	0.0756

Source: Authors' work

It can be seen from Table 2 that the classification rate across ten samples in the 10-fold CV procedure varied in each model. The highest average classification rate was obtained by the ANN model (0.7797), followed by the CART model with the average classification rate of 0.6977. The lowest average rate was produced by the k-nearest neighbour technique (0.6091). The ANN model also had the smallest standard deviation (0.0696), implying that this model is the most accurate and most stable across 10 samples. The variation of each model's accuracy is graphically presented in Figure 2 showing that k-nearest neighbour technique performed particularly low in most of the samples, while the ANN model outperformed others in all samples except in the sample 6 where the SVM model was more accurate.

Figure 2
Classification Rate of the Four Tested Models across 10 Validation Samples



Source: Authors' illustration

Statistical significance of difference in the accuracy of the tested models is tested by the t-test of difference in proportion. The results of the t-test are shown in Table 3 indicating that the p-value is significant on the 5% level only for the difference between the ANN and the KNN models. There is no statistically significant difference between the results of other models.

Table 3
Statistical Comparison of the Average Classification Rates of the Four Models

Hypothesis	**t-test results**
H0: NN=DT	p=0.1919
H0: NN=SVM	p=0.1571
H0: NN=k-nearest	p=0.0430*
H0: DT=SVM	p=0.4453
H0: DT=k-nearest	p=0.1925
H0: SVM=k-nearest	p=0.2319

*significant at 0.05 level
Source: Authors' work

Besides comparing the total classification accuracy, in some problems it is more important to correctly recognize one class of output variable. In case of recognizing entrepreneurial intentions it is more important to correctly recognize the class of students with entrepreneurial intentions (class 1) than the class of students with no intention (class 0). Therefore, classification rates of class 1 and class 0 are further compared across models. Type I and type II errors (sensitivity and specificity) of each model is computed and presented in Table 4.

Table 4
The Sensitivity and Specificity of the Best NN, CART, and SVM Models

Measure of efficiency	NN model	DT model	SVM model	KNN model
Sensitivity	0.843889	0.721495	0.722853	0.635132
Specificity	0.690154	0.681263	0.654512	0.592231
Likelihood ratio L_1	2.930801	2.867496	2.144374	1.666607
Likelihood ratio L_0	0.230211	0.414052	0.422161	0.643737

The model with higher sensitivity ratio has a lower type I error in misclassifying a student with an actual positive entrepreneurial intention (class 1) into the class of students with no intention (class 0). Such error yields a greater loss for the school and for the society than the type II error, and it is more important to recognize more potential entrepreneurs than to misclassify those who have no entrepreneurial intention. Therefore, the most efficient models is the one that has the highest sensitivity, and according to Table 4, it is the ANN model with the average sensitivity of 0.843889 or the type I error of β (0.15611), and also the highest likelihood for recognizing class 1 (2.9308). Sensitivity of other models is much lower than the sensitivity of the ANN model (below 0.8). The lowest sensitivity ratio is obtained by the KNN technique. It is worth noticing that the specificity ratio is also highest in the ANN model, while the likelihood for class 0 is the highest in the model obtained by the KNN technique.

Discussion

This paper compares the efficiency of machine learning methods in a high-dimensional problem of classifying entrepreneurial intentions. Artificial neural networks, decision trees, support vector machines and k-nearest neighbour technique were trained and tested. The performance of the methods is observed by the classification rates obtained in a 10-fold cross-validation procedure. The results show that the artificial neural network method provides the most efficient model and outperforms other machine learning methods according to criteria of classification accuracy, stability over 10 samples, sensitivity, and specificity. However, the accuracy of artificial neural network is significantly higher (on the 0.05 level) only comparing to the accuracy of k-nearest neighbour method, while the difference between the artificial neural networks and other tested methods is not found to be statistically significant. The reason for successfulness of artificial neural network model could be found in its robustness and the ability to minimize the error in the iterative procedure of optimizing its parameters such as learning rate, while the other methods have predefined values of some input parameters. The disadvantages of artificial neural network over other methods are in its slower learning due to a larger number of iterations needed to achieve the accuracy, and in time consuming experiments with different activation functions. Although the support vector machines also require experimenting with different kernel functions, they converge faster comparing to neural networks. The CART decision tree, however, also learn fast and by providing a slightly lower classification average rate than artificial neural networks, are a very strong candidate for an efficient tool in this area after the neural network.

Although the above results can not directly be compared to previous research results, due to the fact that other authors used different datasets and were mostly comparing two out of four methods used in this research, certain similarities and

differences can be identified. Our findings are consistent with the results of Brown et al. (1993) showing that decision trees and artificial neural networks produce comparable error rates. However, our results differ from the results of Shao and Lunetta (2012) who obtained that SVMs had superior generalization capability and less variability comparing to CART decision trees. Our findings show that CART accuracy was not significantly different from the accuracy of SVM, but confirm that SVM method produces the model with less variability.

Conclusion

Accurate classification on real datasets with a high-dimensional input space is still not investigated enough in previous research. The aim of this paper was to provide an extensive research by comparing the accuracy of four machine learning methods in order to analyze their efficiency in recognizing entrepreneurial intentions of students with a large number of input variables. Our findings show that all four tested methods: artificial neural networks, decision trees, support vector machines and k-nearest neighbours are generally able to learn fast and achieve high classification accuracy even with a high-dimensional input space. The artificial neural networks outperformed other methods in classification accuracy, although the difference was significant (on the 0.05 level) only between the artificial neural network and the k-nearest neighbour model. The obtained results partially confirm, and partially differ from previous research findings. The consistency was found in the fact that the three tested methods do not significantly differ in their performance, therefore confirming that competitive way of using machine learning methods is not the right approach, and that further resarch should be focused to integrative approaches. Our results differ from previous research in showing that support vector machines were not found more efficient over decision trees or neural networks.

However, in order to provide more insight and make general conclusions, further tests are necessary on multiple datasets and more algorithms. Future research could be focused on testing some additional methodological improvements in machine learning, such as support vector machines with hierarchical clustering, and others that could enable more thorough analysis of dealing with large dimensional data in machine learning. Such research could be valuable for data mining in education, business and other areas, which is usually based on large databases and deals with the same issue investigated in this paper.

References

1. Apte, C., Weiss, S.(1997), "Data Mining with Decision Trees and Decision Rules", Future Generation Computer Systems , Vol. 13, No.2, pp. 197-210.
2. Behzad, M., Asghar, K., Eazi, M., Palhang, M. (2009), „Generalization performance of support vector machines and neural networks in runoff modeling", Expert Systems with Applications, Vol. 36, No.4, pp. 7624–7629.
3. Bensic, M., Sarlija, N., Zekic-Susac, M. (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. Bishop, C. (1995), Neural Networks for Pattern Recognition, University Press, Oxford, UK.
5. Bolivar-Cime, A., Marron, J.S. (2013), "Comparison of binary discrimination methods for high dimension low sample size data", Journal of Multivariate Analysis, Vol. 115, No. 1, pp. 108-121.

6. Brown, D.E., Corruble, V., Pittard, C.L. (1993). "A comparison of decision tree classifiers with backpropagation neural networks for multimodal classification problems", Pattern Recognition, Vol. 26, No. 6, pp. 953-961.
7. Carr, J.C., Sequeira, J.M. (2007), "Prior family business exposure as intergenerational entrepreneurial intent: A theory of planned behavior approach", Journal of Business Research, Vol. 60, No.10, pp.1090-1098.
8. Dai, Y-H. (2002), "Convergence properties of the BFGS algorithm", SIAM Journal of Optimization, Vol. 13, No. 3, pp. 693-701.
9. Farmer, S.M., X. Yao., Kung-Mcintyre, K. (2011), "The behavioral impact of entrepreneur identity aspiration and prior entrepreneurial experience", Entrepreneurship Theory and Practice, Vo. 35, No. 2, pp. 245-273.
10. Haykin, S. (1999), Neural Networks: A Comprehensive Foundation, Prentice Hall International, Inc., New Jersey, USA.
11. Kolvereid, L., Isaksen, E. (2006), "New business start-up and subsequent entry into self-employment", Journal of Business Venturing, Vol. 21, No.6, pp. 866-885.
12. Krueger, N.F. Jr. (2000), "The Cognitive Infrastructure of Opportunity Emergence", Entrepreneurship: Theory and Practice, Vol. 24, No. 3, pp. 5-23.
13. Krueger, N.F. JR., Reilly, M.D., Carsrud, A.L. (2000), "Competing Models of Entrepreneurial Intentions", Journal of Business Venturing, Vol. 15, No.5, pp. 411–432.
14. Kuzey, C., Uyar, A., Delen, D. (2014), "The impact of multinationality on firm value: A comparative analysis of machine learning techniques", Decision Support Systems, Vol. 59, No. 1, pp. 127-142.
15. Lee, S.(2010), "Using data envelopment analysis and decision trees for efficiency analysis and recommendation of B2C controls", Decision Support Systems, Vol. 49, No.4, pp. 486–497.
16. 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, No.2, pp. 207–220.
17. Liu, G., Yi, Z., Yang, S. (2007), „A hierarchical intrusion detection model based on the PCA neural networks", Neurocomputing, Vol. 70, No.7, pp. 1561–1568.
18. Masters, T. (1995), Advanced Algorithms for Neural Networks, A C++ Sourcebook, John Wiley & Sons, Inc., New York, USA.
19. McGee, J., Peterson, M., Mueller, S., Sequeira, J.M. (2009), "Entrepreneurial self-efficacy: Refining the measure and examining its relationship to attitudes toward venturing and nascent entrepreneurship", Entrepreneurship Theory and Practice, Vol. 33, No.4, pp. 965-988.
20. Min, J.H., Lee, Y.-C. (2005), „Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters", Expert Systems with Applications, Vol. 28, No. 4, pp. 603–614.
21. Nga, J.K.H., Shamuganathan, G. (2010), "The influence of personality traits and demographic factors on social entrepreneurship start up intentions", Journal of Business Ethics, Vol. 95, No.2, pp. 259-282.
22. Paliwal, M., Kumar, U.A. (2009), "Neural networks and statistical techniques: A review of applications", Expert Systems with Applications, Vol. 36, No.1, pp. 2–17.
23. Questier, F., Put, R., Coomans, D., Walczak, B., Vander Heyden, Y. (2005), "The use of CART and multivariate regression trees for supervised and unsupervised feature selection", Chemometrics and Intelligent Laboratory Systems, Vol. 76, No.1, pp. 45-54.
24. Shao, Y., Lunetta, R.S. (2012), "Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited

training data points", ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 70, June 2012, pp. 78-87.
25. Shin, H.J., Eom, D.-H., Kim, S.S. (2005), "One-class support vector machines - an application in machine fault detection and classification", Computers & Industrial Engineering, Vol. 48, No.2, pp. 395–408.
26. Simon, D. and Boring, J.R.(1990), "Sensitivity, Specificity, and Predictive Value", in Walker, H.K., Hall, W.D., Hurst, J.W. (Eds.), Clinical Methods: The History, Physical, and Laboratory Examinations, Butterworths, Boston, pp. 49-54.
27. Smith, T.W. (2009), Altruism and Empathy in America: Trends and Correlates, National Opinion Research Center, University of Chicago, Chicago.
28. St. John, C.H., Balakrishnan, N., Fiet, J.O. (2000), „Modeling the relationship between corporate strategy and wealth creation using neural networks", Computers & Operations Research, Vol. 27, No. 11, pp. 1077-1092.
29. Talukder, A., Casasent, D. (2001), "A closed-form neural network for discriminatory feature extraction from high-dimensional data", Neural Networks, Vol. 14, No. 9, pp. 1201-1218.
30. Thompson, E.R. (2009), "Individual entrepreneurial intent: Construct clarification and development of an internationally reliable metric", Entrepreneurship Theory and Practice, Vol. 33, No. 3, pp. 669-694.
31. Triandis, H.C., Gelfand, M.J. (1998), "Converging Measurement of Horizontal and Vertical Individualism and Collectivism", Journal of Personality and Social Psychology, Vol. 74, No.1, pp.118-128.
32. Witten, I.H., Frank, E. (2000), Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation, Morgan Kaufman Publishers, San Francisco.
33. 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, No.2, pp. 1535–1541.
34. Yu, H., Yang, J., Han, J. (2003),„Classifying Large Data Sets Using SVMs with Hierarchical Clustering", in Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM New York, NY, USA, pp. 306-315.
35. Zanaty, E.A. (2012), "Support Vector Machines (SVMs) versus Multilayer Perception (MLP) in data classification", Egyptian Informatics Journal, Vol. 13, No. 3, pp. 177-183.
36. Zekic-Susac, M., Pfeifer, S., Djurdjevic, I. (2010), „Classification of entrepreneurial intentions by neural networks, decision trees and support vector machines", Croatian Operational Research Review, Vol. 1, No.1, pp. 62-71.
37. Zekic-Susac, M., Šarlija, N., Pfeifer, S. (2012), „Combining PCA analysis and neural networks in modelling entrepreneurial intentions of students", Croatian Operational Research Review, Vol. 4, No.1, pp. 306-317.

About the authors

Marijana Zekić-Sušac is a full professor at the University of J.J. Strossmayer in Osijek, Faculty of Economics in Osijek, Croatia. She has earned her doctoral degree at University of Zagreb, Faculty of Organization and Informatics Varaždin, Croatia. Her research interests include artificial intelligence, machine learning and data mining in business, education and medicine. She currently teaches several ICT courses on undergraduate, graduate and doctoral level. She can be contacted at **marijana@efos.hr**

Sanja Pfeifer is a tenant professor at the University of J.J. Strossmayer in Osijek, Faculty of Economics in Osijek, Croatia. Her main field of research interest are entrepreneurship, management, entrepreneurial education and creativity. Her teaching engagements are focused on a number of graduate and doctoral studies including entrepreneurship management, strategic management and entrepreneurship research methodology. She can be contacted at **pfeifer@efos.hr**

Nataša Šarlija is a full professor at the University of J.J. Strossmayer in Osijek, Faculty of Economics in Osijek, Croatia. Her research interests are: credit risk modeling, scoring models, credit analysis, financial modelling, and data mining. Her teaching engagements include courses on graduate and doctoral studies at Faculty of economics: Credit analysis, Financial management for entrepreneurs, Data mining and Entrepreneurship research methodology, as well as at the Department of mathematics: Managing credit risk and Financial Management. She can be contacted at **natasa@efos.hr**