Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 34 (2008) 1227–1234
Expert Systems with Applications
www.elsevier.com/locate/eswa

Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size

Yong Soo Kim *

CI Division, SK telecom, 11, Euljiro 2-ga, Jung-gu, Seoul, 100-999, Republic of Korea

Abstract

In this article, the performance of data mining and statistical techniques was empirically compared while varying the number of independent variables, the types of independent variables, the number of classes of the independent variables, and the sample size. Our study employed 60 simulated examples, with artificial neural networks and decision trees as the data mining techniques, and linear regression as the statistical method. In the performance study, we use the RMSE value as the metric and come up with some additional findings: (i) for continuous independent variables, a statistical technique (i.e., linear regression) was superior to data mining (i.e., decision tree and artificial neural network) regardless of the number of variables and the sample size; (ii) for continuous and categorical independent variables, linear regression was best when the number of categorical variables was one, while the artificial neural network was superior when the number of categorical variables was two or more; (iii) the artificial neural network performance improved faster than that of the other methods as the number of classes of categorical variable increased.

Keywords: Data mining; Statistical method; Artificial neural network; Decision tree; Linear regression

1. Introduction

The difficulties posed by prediction problems have resulted in a variety of problem-solving techniques. For example, data mining methods comprise artificial neural networks and decision trees, and statistical techniques include linear regression and stepwise polynomial regression. It is difficult, however, to compare the efficacy of the techniques and determine the best one because their performance is data-dependent.

A few studies have compared data mining and statistical approaches to solving prediction problems. Gorr, Nagin, and Szczypula (1994) compared linear regression, stepwise polynomial regression, and neural networks in the context of predicting student GPAs. Although they found that linear regression performed best overall, none of the methods performed significantly better than the ordering index used by the investigator. Shuhui, Wunsch, Hair, and Giesselmann (2001) reported that neural networks performed better than linear regression for wind farm data, while Hardgrave, Wilson, and Walstrom (1994) experimentally showed that neural networks did not significantly outperform statistical techniques in predicting the academic success of students entering the MBA program. Subbanarasimha, Arinze, and Anadarajan (2000) demonstrated that linear regression performed better than neural networks when the distribution of the dependent variable was skewed, and Kumar (2005) expanded on Subbanarasimha et al. (2000) result, developing a hybrid method that improved the prediction accuracy.

These comparison studies have mainly considered a specific data set or the distribution of the dependent variable. Other unexplored criteria, however, affect the performance

* Tel.: +82 2 6100 5987; fax: +82 2 6100 7911.
E-mail address: yskim95@kaist.ac.kr

0957-4174/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2006.12.017

of decision problem techniques, such as sample size and characteristics of the independent variables. We empirically compared the performance of data mining and statistical techniques while varying the number of independent variables, the types of independent variables, the number of classes of the independent variables, and the sample size. Our study employed 60 simulated examples, with artificial neural networks and decision trees as the data mining techniques, and linear regression as the statistical method.

In addition to these general comparison results, we used the RMSE value as the metric and determined the following: for continuous independent variables, a statistical technique (i.e., linear regression) was superior to data mining (i.e., decision tree and artificial neural network) regardless of the number of variables; for continuous and categorical independent variables, linear regression was best when the number of categorical variables was one, while the artificial neural network was superior when the number of categorical variables was two or more; and the artificial neural network performance improved faster than that of the other methods as the number of classes of categorical variable increased.

The article is organized as follows. Section 2 illustrates the generation of the data sets and analysis methods for the empirical study. The experimental results are described in Section 3, and the conclusions and future research directions are presented in Section 4.

2. Data analysis

2.1. Data generation

In this section, we describe the 60 simulated prediction problems that we generated to evaluate the performance of the decision tree, neural network, and linear regression techniques. First, Table 1 shows 12 simulated examples with continuous independent variables.

These 12 examples were obtained from the linear model, where x_i was randomly selected in the range $[0,1]$, and ε was normally distributed with mean 0 and standard deviation 1. The number of independent variables was set to one, three, or five, and the sample size was set to 100, 500, 1000, or 10,000.

Some continuous variables in Table 1 were converted to categorical variables in Tables 2 and 3. Three and five independent variables were considered in Tables 2 and 3, respectively. In the case of two categorical variables, a value of a continuous variable was converted to category 'A' when it is less than 50% and as 'B' when greater than 50%. For three categorical variables, each continuous variable was categorized as 'A' when it was less than 25%, as 'B' when greater than 75%, and as 'C' otherwise.

2.2. Data analysis methods

In this section, the artificial neural network (ANN), decision tree analysis (DT), and linear regression (LR) techniques are applied to the 60 simulated examples to evaluate their prediction accuracy. Each example was randomly divided into two sets, a training set and a test set. The training set consisted of 70% of the data while the remainder was assigned to the test set. For simplicity, our performance comparisons only considered the root mean square error of the test set. The analyses were performed using "SAS Enterprise Miner".

The ANN employed in this study was a multilayer feedforward network trained by a backpropagation algorithm. The number of hidden layers was set to either one or two. For each hidden layer, the number of hidden neurons varied between one and ten to identify the best ANN structure. The learning rate and momentum were set to 0.1 and 0.9, respectively. A low learning rate ensures a continuous descent on the error surface, and a high momentum is able to speed up the training process (Sarle, 1994; Yeh, Hamey, & Westcott, 1998). These values are typically used for ANN training (Ting, Yunus, & Salleh, 2002).

For DT, we varied the splitting criterion and used two parameters for pre-pruning: 'minimum number of observations in a leaf' and 'observations required for a split search'. The splitting criterion was set to either '*F*-test at 2% significance level' or 'Variance reduction'. The 'minimum number of observations in a leaf' and 'observations

Table 1
Simulated examples with continuous independent variables

Example ID	No. of independent variables	Sample size	Relationship	Distribution of independent variables and error term
S1	1	100	$y = 1 + 5x + \varepsilon$	$x_i \sim U(0,1)$
S2	1	500	$y = 1 + 5x + \varepsilon$	$\varepsilon \sim N(0,1)$
S3	1	1000	$y = 1 + 5x + \varepsilon$	
S4	1	10,000	$y = 1 + 5x + \varepsilon$	
S5	3	100	$y = 1 + 3x_1 + 2x_2 + 2x_3 + \varepsilon$	
S6	3	500	$y = 1 + 3x_1 + 2x_2 + 2x_3 + \varepsilon$	
S7	3	1000	$y = 1 + 3x_1 + 2x_2 + 2x_3 + \varepsilon$	
S8	3	10,000	$y = 1 + 3x_1 + 2x_2 + 2x_3 + \varepsilon$	
S9	5	100	$y = 1 + 3x_1 + 2x_2 + 2x_3 + x_4 + x_5 + \varepsilon$	
S10	5	500	$y = 1 + 3x_1 + 2x_2 + 2x_3 + x_4 + x_5 + \varepsilon$	
S11	5	1000	$y = 1 + 3x_1 + 2x_2 + 2x_3 + x_4 + x_5 + \varepsilon$	
S12	5	10,000	$y = 1 + 3x_1 + 2x_2 + 2x_3 + x_4 + x_5 + \varepsilon$	