Comparison of Classification Performance of Selected Algorithms Using Rural Development Investments Support Programme Data

Mehmet Ali ALAN 1 Cavit YEŞİLYURT 2 Saadettin AYDIN 3 Erol AYDIN 4

1 Cumhuriyet University, Faculty of Economics and Administrative Sciences, Department of Management Information Systems, TR-58140 Sivas - TURKEY
2 Kafkas University, Faculty of Economics and Administrative Sciences, Department of Management, TR-36100 Kars - TURKEY
3 Institute of Strategic Thinking, TR-06460 Ankara - TURKEY
4 Kafkas University, Faculty of Veterinary Medicine, Department of Livestock Economics and Management, TR-36100 Kars - TURKEY

Makale Kodu (Article Code): KVFD-2013-10154

SUMMARY

It is not always possible to solve a large size of data via traditional statistical techniques. In order to solve these kinds of data special tactics like data mining are needed. Data mining may meet these kinds of needs with both categorizing and piling tactic. In this study, we have used data mining by using Rural Development Investment Support Program (RDISP) data with various categorizing algorithms. The most prospering categorizing algorithm was tried to determine by using present data. At the end of analysis, it has been understood that MLP (multilayer perceptron), a nerve net model, is the best algorithm that makes the best categorizing.

Keywords: Data mining, MLP, Nerve net model, RDISP, Rural development

INTRODUCTION

Databases are rich with hidden information that can be used for intelligent decision making. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Such analysis can help provide us with a better understanding of the data at large [1].

The different disciplines on database’s data make statistical, mathematical, machine learning and visual analyses with different purposes. One of those analyses techniques is data mining. There are a lot of algorithms in data mining.

Data mining is an interdisciplinary field, the confluence of a set of disciplines, including database systems, statistics, machine learning, visualization, and information science [1].

Data mining, the science and technology of exploring data in order to discover previously unknown patterns, is a part of the overall process of knowledge discovery in...
Comparison of Classification 

...
layer. These inputs pass through the input layer and are then weighted and fed simultaneously to a second layer of “neuronlike” units, known as a hidden layer. The outputs of the hidden layer units can be input to another hidden layer, and so on. The number of hidden layers is arbitrary, although in practice, usually only one is used [1].

The weighted outputs of the last hidden layer are input to units making up the output layer, which emits the network’s prediction for given tuples. The units in the input layer are called input units. The units in the hidden layers and output layer are sometimes referred to as neurons, due to their symbolic biological basis, or as output units.

Multilayer feed-forward networks are one of the most important and most popular classes of ANNs in real-world applications. Typically, the network consists of a set of inputs that constitute the input layer of the network, one or more hidden layers of computational nodes, and finally an output layer of computational nodes. The processing is in a forward direction on a layer-by-layer basis [13].

A multiplayer perceptron has three distinctive characteristics:

1. The model of each neuron in the network includes usually a nonlinear activation function, sigmoidal or hyperbolic.
2. The network contains one or more layers of hidden neurons that are not a part of the input or output of the network. These hidden nodes enable the network to learn complex and highly nonlinear tasks by extracting progressively more meaningful features from the input patterns.
3. The network exhibits a high degree of connectivity from one layer to the next one.

The multilayer perceptron is the most commonly used architecture for predictive data mining. It is a feed-forward network, with possibly several hidden layers, one input layer and one output layer, totally interconnected. It can be considered as a highly non-linear generalization of the linear regression model when the output variables are quantitative, or of the logistic regression model when the output variables are qualitative [1].

In this study, the algorithms named as MultiLayer Perceptron, Ridor, DTNB, ADTree, LADTree, SPEgassos, SQO, Dagging, IBk, FT, LMT, LBR, Voted Perceptron, OneR, I81, VFI, Decorate, Bayes Net, RBF Network, Naïve Bayes were used to select the algorithm has the best classification performances by benefiting the Rural Development Investments Support Programme (RDISP) data. In order to determine the best algorithm, Correctly Classified Instances, Kappa statistic, Mean absolute error, Root mean squared error, Relative absolute error, Root relative squared error, TP Rate, FP Rate, F-Measure, classification timing values of the algorithms has taken into consideration.

### MATERIAL and METHODS

In this study the data from Sivas Provincial Directorate of Agriculture were used in the frame of Rural Development Investment Support Program. There are 14859 data belong to 1143 appeal. Data have been taken into Excel format. Then necessary regulations, transformations are made by using Excel macros and the file saved in the name of “kkydp.arff”. Both individual and corporate applications of RDISP are kept as B/K (I/C). Lands owned by applicants have been identified as “up to 30 decare”, “between 31-40 decare” and “more than 40 decare” and applicant’s loan land has been identified similarly as “reach up 30 decare”, “between 31-40 decare” and “more than 40 decare”. Applicant’s request type is set as “M1/M2” (M1: local, M2: Imported). For city or county “i” value and for village “k” value was assigned as application location. Utilization statement was assigned as “twice”, “once” or “never” from RDISP. Results were determined to be “Positive” if the application has been accepted and “Negative” if it has not. It was assigned whether there is no value belong to a variable or it’s undefined “?” value.

Variable definitions in pre-processing step of prepared dataset are given below: @relation kkydp, @attribute BT (B,K), @attribute S30 (E,H), @attribute S3040 (Y,N), @attribute S40 (Y,N), @attribute K30 (Y,N), @attribute K3040 (Y,N), @attribute K40 (Y,N), @attribute MType (M1,M2), @attribute location (i,k), @attribute twice (Y,N), @attribute once (Y,N), @attribute notbenefited (Y,N), @attribute Class (Positive, Negative), @DATA B,N,Y,N,N,Y,M1,K,N,N,Y, Positive; B,N,Y,N,N,N,M1,K,N,Y, Negative.

### RESULTS

In this study, WEKA (Waikato Environment for Knowledge Analysis) program’s 3.6.9 version that was developed in Waikato University was used [14]. WEKA program is open source code software. This program supports a lot of categorizing, piling and coupling rules algorithm. Instead of text based arff., arff.gz, names, data, csv, c45, libsvm, dat, bsi, xrrf, xrrf.gz file types WEKA supports databases and URL addresses that include data.

An Intel i5 model and 1.7 Ghz CPU, 6 Gb RAM and 64 bit Win 8 operating systemic laptop was used during the application.

There are different but in equal number of values for every defined variables as it is shown in Fig. 1. In addition, every variable has been take by using two different values as yes or no (Y,N) and these values represented as two different colors.

Results that come from post prepared data set used in WEKA program is given in the chart. As it can be understood from the Table 1, Multilayer Perception algorithm is the
Comparison of Classification ... best algorithm that makes the best categorizing with 992 correctly classified instances. This algorithm's kappa statistic is 0.7321, True Positive rate is 0.868, and False Positive rate is 0.124 and F-measure is 0.869. This algorithm's categorizing time is 1.99 seconds. Ridor algorithm follows this algorithm with 973 true categorizing numbers.

Table 1. Performance ratings of selected algorithms in Decision Tree Analysis

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Correctly Classified Instances</th>
<th>Kappa Statistic</th>
<th>Mean Absolute Error</th>
<th>Root Mean Squared Error</th>
<th>Relative Absolute Error %</th>
<th>Root Relative Squared Error %</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>F-Measure</th>
<th>Time (in seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MultiLayer Perceptron</td>
<td>992</td>
<td>0.7321</td>
<td>0.1465</td>
<td>0.2767</td>
<td>30.1889</td>
<td>56.1843</td>
<td>0.868</td>
<td>0.124</td>
<td>0.869</td>
<td>1.99</td>
</tr>
<tr>
<td>Ridor</td>
<td>973</td>
<td>0.6972</td>
<td>0.1487</td>
<td>0.3857</td>
<td>30.6551</td>
<td>78.3030</td>
<td>0.851</td>
<td>0.145</td>
<td>0.852</td>
<td>0.06</td>
</tr>
<tr>
<td>DTNB</td>
<td>965</td>
<td>0.6737</td>
<td>0.2085</td>
<td>0.3161</td>
<td>42.9768</td>
<td>64.1741</td>
<td>0.844</td>
<td>0.181</td>
<td>0.843</td>
<td>0.47</td>
</tr>
<tr>
<td>ADTree</td>
<td>965</td>
<td>0.6796</td>
<td>0.2562</td>
<td>0.3243</td>
<td>52.7993</td>
<td>65.8508</td>
<td>0.844</td>
<td>0.163</td>
<td>0.844</td>
<td>0.06</td>
</tr>
<tr>
<td>LADTree</td>
<td>962</td>
<td>0.6584</td>
<td>0.2294</td>
<td>0.3200</td>
<td>47.2783</td>
<td>64.9635</td>
<td>0.842</td>
<td>0.212</td>
<td>0.835</td>
<td>0.13</td>
</tr>
<tr>
<td>SPegasos</td>
<td>927</td>
<td>0.6036</td>
<td>0.1890</td>
<td>0.4347</td>
<td>38.9500</td>
<td>88.2634</td>
<td>0.811</td>
<td>0.218</td>
<td>0.809</td>
<td>0.11</td>
</tr>
<tr>
<td>SMO</td>
<td>927</td>
<td>0.6036</td>
<td>0.1890</td>
<td>0.4347</td>
<td>38.9500</td>
<td>88.2634</td>
<td>0.811</td>
<td>0.218</td>
<td>0.809</td>
<td>0.17</td>
</tr>
<tr>
<td>Dagging</td>
<td>926</td>
<td>0.5991</td>
<td>0.2203</td>
<td>0.4106</td>
<td>45.4056</td>
<td>83.3628</td>
<td>0.810</td>
<td>0.225</td>
<td>0.807</td>
<td>0.14</td>
</tr>
<tr>
<td>IBk</td>
<td>924</td>
<td>0.5807</td>
<td>0.2188</td>
<td>0.3597</td>
<td>45.0871</td>
<td>73.0424</td>
<td>0.808</td>
<td>0.261</td>
<td>0.797</td>
<td>0.01</td>
</tr>
<tr>
<td>FT</td>
<td>923</td>
<td>0.5952</td>
<td>0.2644</td>
<td>0.3889</td>
<td>54.5024</td>
<td>78.9647</td>
<td>0.808</td>
<td>0.224</td>
<td>0.805</td>
<td>0.36</td>
</tr>
<tr>
<td>LMT</td>
<td>921</td>
<td>0.5923</td>
<td>0.2809</td>
<td>0.3758</td>
<td>57.8871</td>
<td>76.2947</td>
<td>0.806</td>
<td>0.224</td>
<td>0.803</td>
<td>1.78</td>
</tr>
<tr>
<td>LBR</td>
<td>916</td>
<td>0.5682</td>
<td>0.2457</td>
<td>0.3819</td>
<td>50.6500</td>
<td>77.5334</td>
<td>0.801</td>
<td>0.263</td>
<td>0.791</td>
<td>0.02</td>
</tr>
<tr>
<td>Voted Perceptron</td>
<td>914</td>
<td>0.5788</td>
<td>0.2003</td>
<td>0.4476</td>
<td>41.2942</td>
<td>90.8754</td>
<td>0.800</td>
<td>0.232</td>
<td>0.797</td>
<td>0.02</td>
</tr>
<tr>
<td>OneR</td>
<td>904</td>
<td>0.5831</td>
<td>0.2091</td>
<td>0.4573</td>
<td>43.0974</td>
<td>92.8438</td>
<td>0.791</td>
<td>0.188</td>
<td>0.792</td>
<td>0.01</td>
</tr>
<tr>
<td>IB1</td>
<td>897</td>
<td>0.5270</td>
<td>0.2152</td>
<td>0.4639</td>
<td>44.3597</td>
<td>94.1936</td>
<td>0.785</td>
<td>0.291</td>
<td>0.770</td>
<td>0.01</td>
</tr>
<tr>
<td>VFI</td>
<td>882</td>
<td>0.5119</td>
<td>0.4655</td>
<td>0.4675</td>
<td>95.9414</td>
<td>94.9301</td>
<td>0.772</td>
<td>0.278</td>
<td>0.765</td>
<td>0.02</td>
</tr>
<tr>
<td>Decorate</td>
<td>870</td>
<td>0.5007</td>
<td>0.4171</td>
<td>0.4361</td>
<td>85.9614</td>
<td>88.5406</td>
<td>0.761</td>
<td>0.267</td>
<td>0.759</td>
<td>1.38</td>
</tr>
<tr>
<td>Bayes Net</td>
<td>842</td>
<td>0.4456</td>
<td>0.3048</td>
<td>0.4151</td>
<td>62.8232</td>
<td>84.2842</td>
<td>0.737</td>
<td>0.300</td>
<td>0.733</td>
<td>0.05</td>
</tr>
<tr>
<td>RBF Network</td>
<td>840</td>
<td>0.4409</td>
<td>0.3252</td>
<td>0.4086</td>
<td>67.0371</td>
<td>82.9662</td>
<td>0.735</td>
<td>0.304</td>
<td>0.731</td>
<td>0.31</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>806</td>
<td>0.3801</td>
<td>0.3323</td>
<td>0.4529</td>
<td>68.4844</td>
<td>91.9612</td>
<td>0.705</td>
<td>0.705</td>
<td>0.701</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Although this algorithm’s categorizing time is shorter than MultiLayer Perceptron the other values are worse. Other algorithms follow them.

**DISCUSSION**

Nowadays, the amount of stored data extremely increases. Data storage is performed by not only private sector, but also by public enterprises such as provincial directorate of agriculture. While private enterprises are achieved in increasing customer commitment to the enterprise and customer satisfaction by using these data, especially public enterprises could not use these data effectively time to time. These data could include some beneficial hide patterns for both public and private sector.

One of the most important methods used in producing beneficial information from these data is data mining. Data mining is to produce beneficial and useful data from large scale of data.

In data mining; different methods such as; statistical methods, decision trees, genetic algorithm, fuzzy logic and artificial neural networks could be used. Despite traditional methods, in data mining, inferences could be deduced oriented to results by using the entire data. In this technique, not only numerical data but also alphanumerical data is used in analyses. A data warehouse is formed by changing both numeric and alpha numerical data to required form. This study is performed by subjecting current data to required change and a data warehouse is prepared in text format by pruning.

Both numerical results and visual results are used in data mining. In this study, there is a graphic (Fig. 1) which shows the place of each variable in the entire data besides determination of the most successful classification algorithm. There are a lot of studies as the current study and most of these are from different data sets about the subject. One of these is Palaniappan et al., categorized decision tree, Naive Bayes method and artificial nerve nets by using Heart Disease Prediction System (HDPS)’s data and they stated that these results may help nursing and medicine students. In addition; Frank et al. introduced how to use WEKA software in Bioinformatics, emphasizing that it supports important categorizing and regression techniques like decision trees, rule masses, bayes sorters, SVM (Support Vector Machines), Logistic and linear regression, MLP (Multi-Layer Perceptron) and the closest neighbour. Another study is Ngai et al. categorized Customer Relations Method (CRM) and data mining articles by scanning and probed using data mining in customer relations. Kirkos et al., presented that fraudulent statements can be determined by using decision trees, Artificial Nerve Nets and Bayes Nets from data mining algorithms. DI MIC et al. collected student data by using Moodle e-learning materials and made analyses with categorizing, piling and coupling rules technique by those data. Hsieh made analyses with artificial nerve nets and coupling rules by using the data from bank database and presented the data mining’s contribution on behavioral methods of a credit card customer in a bank. Hung et al. tried to guess customer transfers between mobile companies through both artificial nerve nets and decision trees data mining, using data from a telecom company in Taiwan.

Healthy estimation are very important in the studies. One of the most widely used techniques in data mining is the classification. Estimation techniques based on machine learning have been proved to be more successful than the traditional estimation techniques in parallel to developments in information technologies.

As a result of the study, MultiLayer Perceptron algorithm is the best classification algorithm which is an artificial neural networks model. In this study it has been introduced, artificial neural network classification performance in data mining algorithms is higher than that of the other algorithms.

**REFERENCES**

Comparison of Classification ...


