Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol.6, No.1, April 2018, pp. 76-83
Available online at: http://pen.ius.edu.ba

Implementation Example of the Expert system for Decision Support on Android platform based on a specific Dataset

Muzafer Saračević[1], Aybeyan Selimi[2], Mersad Mujević[3]
[1] Department of Computer sciences, University of Novi Pazar, Serbia
[2] Faculty of Informatics, International Vision University, Macedonia
[3] Public Procurement Administration, Government of Montenegro

Article Info

Article history:

Received Dec 17th, 2017
Revised Mar 16th, 2018
Accepted Aprt 10th, 2018

Keyword:

Expert systems
Decision Support
DataSet qualification problem
Model prediction
Android platforms

ABSTRACT

This paper presents the method of creating Expert system for decision support on the Android platform. The system knowledge base for the given area of expertise is generated by inductive learning methods based on examples from the *WEKA* data research system. The system was realized using the Expert System shell for the *e2gDroid lite* mobile device, based on the application area and a set of training examples, specifically based on the *Covertype DataSet* qualification problem.

Corresponding Author:

Muzafer Saračević,
Department of Computer sciences,
University of Novi Pazar,
Dimitrija Tucovica bb, Novi Pazar, Serbia,
Email: muzafers@uninp.edu.rs

1. Introduction

The main task of the *Covertype* Data Set qualification problem is to predict forest cover type only with cartographic signs, without other data. Independent variables are derived from data originally obtained from the *US Geological Survey* - *USGS* and *US Forest Service* - *USFS*. This research includes four wild regions ie areas, located in the Roosevelt National Forest north of Colorado. Some basic information for these four regions are:

1. Rawah (region 1)
2. Neota (region 2), probably has the highest altitude,
3. Comanche Peak (region 3) have a lower altitude than region 2,
4. Cashe la Poudre (region 4) has the lowest altitude.

As for the types of trees in this area: in *Neot*, the most common spruce / firs (type 1), while in *Rawah* and *Comanche Peak* is the most abundant twisted pine (type 2) as the main species, then spruce / fir and aspen (type 5) . In Cache la Poudre there are red pine (type 3), *Douglas* fir (type 6), and poplar / willow (type 4). The areas of *Rawah* and *Comanche Peak* tend to be more typical when looking at data than *Neota* or *Cache la Poudre*, precisely because of the diversity of tree species and the range of predictable values of variables such as altitude. *Cache la Poudre* is more unique than others due to less altitude value and species diversity.

In addition to four regions, twelve cartographic measures (independent variables) and seven types of large forest surfaces (dependent variables) are included. This set of data has 581,012 instances of which: the first 11,340 records are taken for view the data from a subset, the next 3,780 records are taken for data verification in subset, and the last 565,892 records are taken for test the data in the subsets [1].

Table 1. Basic characteristics of *Covertype DataSet* qualification problem

Characteristics of data:	More variants	**Number of instances:**	581012
Characteristics of attributes:	Cartographic, integers	**Number of attributes:**	54
Assigned tasks:	Division, classification	**Missing variables?**	No

Managers of national parks responsible for management of for strategy of eco-system require basic information, including a list of earth reforestation inventories to make it easier for the decision-making process. One way of obtaining this information is model prediction.
In the papers [2,3,4], two predictive models were examined: a model of a neural network and a traditional statistical model based on a discriminant analysis. The overall objectives of these studies are to develop these two predictive models [5], to compare and estimate their precision in the division of types of wood cover in unexplored (uninhabited) forests.
Several sub-sets of these variables have been tested for determining of the best predictive model [6,7,8,9]. For each subset of twelve cartographic variables, which were examined in studies, the relative classification indicates that the approach to the application of neural networks exceeds the traditional method of discriminatory analysis in predicting of the forest cover type. The final neural network model was more precise in the classification (70.58%) than the linear regression model for prediction (58.38%). In support of these results, there are thirty more networks with randomly selected initial results. The total mean value of the precision in split for the neural network model is 70.52%.
Therefore, national park managers can use an alternative method in predicting of the forest cover type that is superior to the traditional method and adequate to support their decision-making process for the eco-system management strategy.

2. Attributes and Classes for *Covertype DataSet* qualification problem

The dataset consists of 54 attributes, 10 attribute values are from numerical type, while the *Wilderness_Area* type attributes values consists of 4 binary variables, and the *Soil_Type* values consist of 40 binary values. For Attribute information (*name, type, unit of measurement, description)* see UCI Machine Learning Repository: *Covertype DataSet* [1].
All examples are associated with one of the *Class_Type* qualification attributes, whose values can be numeric in the range from 1 to 7, where each of numbers represents one of the classes[1]:

1. *Spruce / Fir*
2. *Lodgepole Pine*
3. *Ponderosa Pine*
4. *Cottonwood/Willow*
5. *Aspen*
6. *Douglas-fir*
7. *Krummholz.*

Below are given the values of the first two samples:

- 2596,51,3,258,0,510,221,232,148,6279,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0, 0,0,0,0,0,0,0,0,0,0,0,5
- 2590,56,2,212,6,390,220,235,151,6225,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0, 0,0,0,0,0,0,0,0,0,0,0,5

[1] Standard names of classes from the given Covertype dataset which is used in other test cases and case studies

Since *Wilderness_Area* and *Soil_Type* consist of mutually exclusive binary values, we can combine them, so from 4-binary attribute for *Wilderness_Area* we will obtain one attribute with a nominal value. In the same way, the *Soil_Type* with 40 binary attributes, we expire on one attribute with a nominal value. After this preprocessing, the attribute definitions in *.arff* format are given below (see *Chart* 1):

@relation covertype
@attribute elevation numeric
@attribute aspect numeric
@attribute slope numeric
@attribute horz_dist_hydro numeric
@attribute vert_dist_hydro numeric
@attribute horiz_dist_road numeric
@attribute hillshade_9am numeric
@attribute hillshade_noon numeric
@attribute hillshade_3pm numeric
@attribute horiz_dist_fire numeric
@attribute wilderness_area {1,2,3,4}
@attribute soil_type {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31, 32,33,34,35,36,37,38,39,40} and *@attribute class {1,2,3,4,5,6,7}*

The values of the first two samples from the data set in the *.arff* format are given below:
- 2596,51,3,258,0,510,221,232,148,6279,1,29,5
- 2590,56,2,212,-6,390,220,235,151,6225,1,29,5.

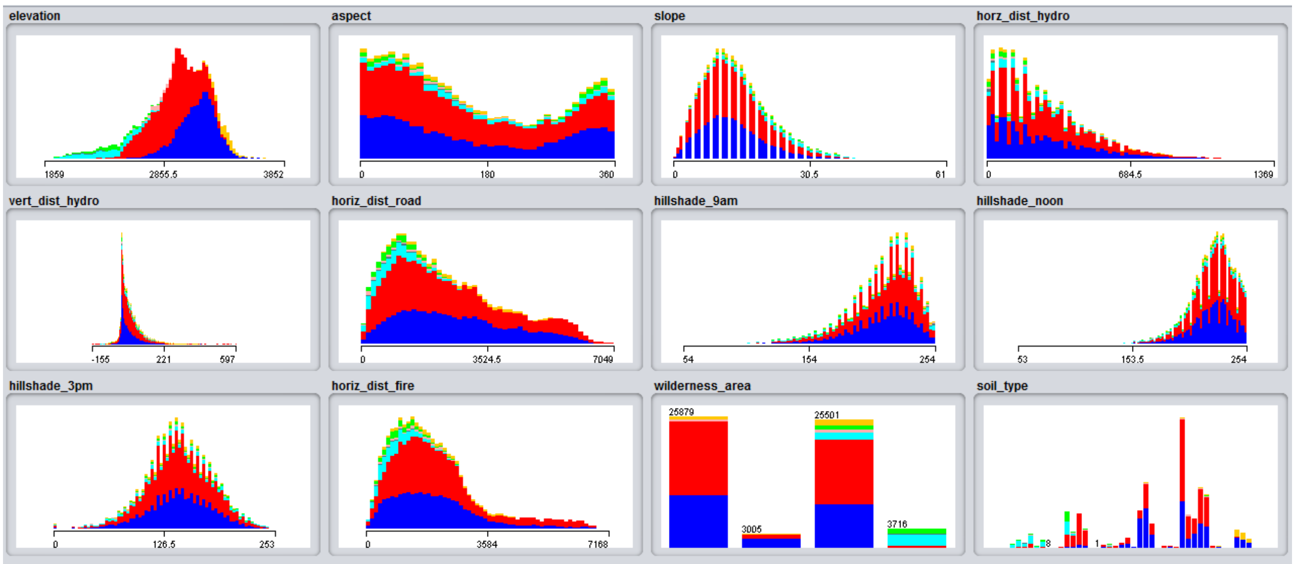

Chart 1. Visual representation of attribute distribution after preprocessing

Redistribution of examples by classes (see Table 2):

Table 2. Without "Resample" filter

Classes	No. of samples	
Spruce-Fir	211840	
Lodgepole Pine	283301	
Ponderosa Pine	35754	
Cottonwood/Willow	2747	
Aspen	9493	
Douglas-fir	17367	

Krummholz	20510	
Total instances	**581012**	

Redistribution after the application of unsupervised instances of the *"Resample"* filter, taking 10% (see Table 3):

Table 3. "Resample" filter

Classes	No. of samples	
Spruce-Fir	20885	
Lodgepole Pine	28618	
Ponderosa Pine	3611	
Cottonwood/Willow	280	
Aspen	922	
Douglas-fir	1730	
Krummholz	2055	
Total instances	**58101**	

Next chart shows redistribution based on Table 3.

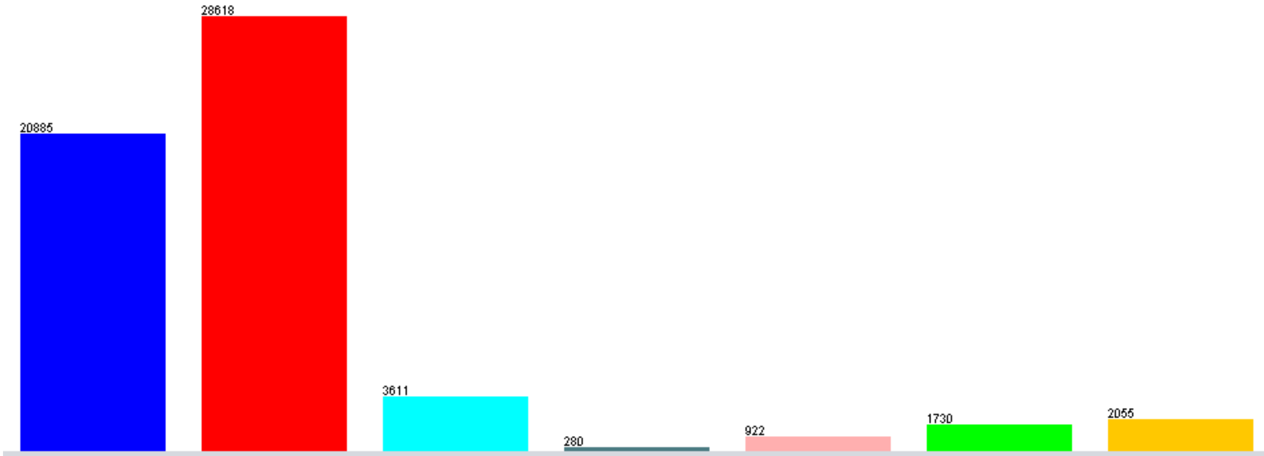

Chart 2. Redistribution after the application of unsupervised instances of the "Resample" filter, taking 10%

3. Learning outcomes and learned rules

As can be seen from the previous section of this paper, after filtering with the unsupervised instance *Resample* filter, the number of instances is reduced to 58101, which is 10% of the total number of samples in the entire dataset. With using of the PART method, with the default parameters to the preprocessed dataset, we get 1760 learned rules.

The precision is checked by a 10-fold cross-validation and we get that 84.29% is correctly classified. However, the number of rules is very high in order to be manually translated into a knowledge base for the expert system, so we must try to "tree pruning" by changing the parameters in the PART method:

- By increasing the parameter M (the minimum number of instances as a rule), we reduce the tree or the number of rules, since the data set is relatively large, we take M = 1000.
- We will also reduce the value of C (Confidence Factor) whose reduction we achieve a greater "tree pruning", we take the value C = 0.15 (default is 0.25).
- With these parameters we get 143 learned rules, and precision is 75.36%, which is again a great number for manual translation into the knowledge base.
- By adjusting the parameters we will try to get a reasonable number of rules that we can manually translate into the knowledge base.
- After several attempts for different values of M and C, we have come to an optimal solution where for the parameter values M = 400 and C = 0.15 we obtain a tree of 28 rules and a precision of 71.07%.

3.1 Obtained learning rules

Using the chosen method for learning of the production rules, PART from the WEKA data research system [10], is inductively learned set of the production rules. The accuracy and comprehensiveness of the learned knowledge is optimized with the available **M** and **C** parameters of the selected learning methods.

Table 4. PART decision list (only the first 17 rules)

elevation > 2714 AND elevation <= 3049 AND elevation > 2907 AND horz_dist_hydro > 95: 2 (10430.0/2639.0) elevation > 2908 AND elevation <= 3294 AND elevation <= 3145 AND horiz_dist_road <= 5608 AND hillshade_noon > 242 AND vert_dist_hydro > 25: 2 (614.0/144.0) elevation > 2908 AND elevation <= 3294 AND elevation <= 3181 AND horiz_dist_road <= 5624 AND elevation > 2983 AND horiz_dist_fire <= 3158 AND soil_type = 23: 1 (2003.0/604.0) elevation > 2908 AND elevation <= 3294 AND elevation <= 3145 AND horiz_dist_road <= 5608 AND elevation > 2983 AND horiz_dist_fire > 3158: 1 (941.0/205.0) elevation > 2938 AND elevation <= 3328 AND elevation <= 3124 AND horiz_dist_road > 1584: 1 (664.0/284.0)	elevation > 2908 AND elevation <= 3294 AND elevation <= 3139 AND horiz_dist_road <= 5624 AND soil_type = 32: 2 (1164.0/413.0) elevation > 2908 AND elevation <= 3294 AND horiz_dist_road > 5433 AND horiz_dist_fire > 1090: 2 (1277.0/353.0) elevation > 2908 AND elevation <= 3294 AND elevation <= 3181 AND elevation > 3067: 1 (5534.0/1819.0) elevation > 2909 AND elevation <= 3328 AND elevation <= 3124 AND horiz_dist_road <= 3543 AND hillshade_noon <= 221: 1 (800.0/265.0) elevation > 2909 AND elevation <= 3328 AND elevation <= 3124 AND elevation > 2955 AND horiz_dist_road > 1734 AND horiz_dist_road > 3632: 2 (465.0/191.0)	elevation > 3063 AND elevation <= 3328 AND soil_type = 23: 1 (1394.0/350.0) elevation > 3063 AND elevation <= 3328 AND soil_type = 22: 1 (1257.0/125.0) elevation > 3063 AND elevation > 3329 AND wilderness_area = 1: 1 (886.0/236.0) elevation > 3063 AND elevation <= 3333 AND soil_type = 33: 1 (1006.0/329.0) elevation > 3066 AND elevation <= 3328 AND horiz_dist_road > 1127 AND soil_type = 32: 1 (859.0/292.0) elevation > 3066 AND elevation > 3349 AND horiz_dist_road <= 3517 AND horiz_dist_fire <= 2016 AND hillshade_3pm <= 157: 7 (499.0/130.0) elevation > 2668 AND elevation <= 3066: 2 (13171.0/3182.0) …

Table 5. Obtained precision by classes

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.677	0.160	0.704	0.677	0.690	0.521	0.830	0.715	1
0.816	0.285	0.735	0.816	0.773	0.533	0.820	0.778	2
0.759	0.031	0.618	0.759	0.681	0.662	0.969	0.637	3
0.000	0.000	0.000	0.000	0.000	0.000	0.985	0.171	4
0.000	0.000	0.000	0.000	0.000	0.000	0.858	0.058	5
0.119	0.005	0.440	0.119	0.187	0.218	0.953	0.337	6
0.422	0.009	0.638	0.422	0.508	0.505	0.968	0.533	7
0.711	**0.200**	**0.689**	**0.711**	**0.695**	**0.515**	**0.843**	**0.710**	

Precision represented as a confusion matrix:

a	b	c	d	e	f	g	<-- classified as
14137	6280	16	0	0	0	452	a = 1
4792	23344	389	0	0	53	40	b = 2
0	662	2740	0	0	209	0	c = 3
0	0	280	0	0	0	0	d = 4
3	905	14	0	0	0	0	e = 5
0	527	997	0	0	206	0	f = 6
1162	26	0	0	0	0	867	g = 7

4. View the functioning of the Expert system on Android Platforms and Web environment

In creation a *covtype.kb* file with a total of 28 rules, is used the *e2gRuleWriter tool*. The learned set of rules was built into the knowledge base of the expert system *e2gDroid Expert System*. The user interface of the system in Serbian was created using the *Expertise2Go translate* directive. Below are given the demonstration of performance testing of a small expert system on some of the selected examples.
The first case of testing on Android Platforms (Figure 1) and second case of testing in web - HTML environment (Figure 2).

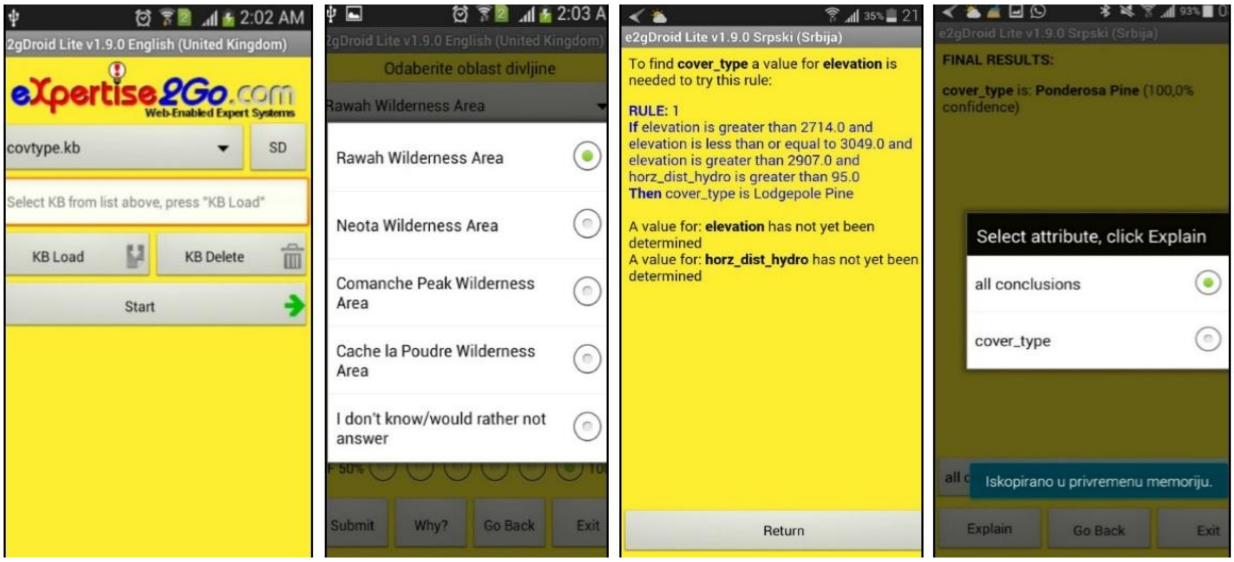

Figure 1. Some parts of an application that has been customized for Android Platforms

Figure 2. Some parts of the application that work in the web - HTML environment (with language customization

5. Conclusion

An efficient way of creating a small Expert Decision Support System for the Android platform is shown without serious programming in the *Java* programming language. The knowledge base of the system, for given area of expertise was generated by inductive learning methods based on examples from the WEKA data research system, and the system was realized using the *Expertise2Go* and *e2gDroid Lite Expert shell* system for mobile devices.

Based on the given application area and a set of trained examples, specifically based on the *Covertype DataSet* qualification problem, was developed a support system for the decisions.

6. References

[1] UCI Machine Learning Repository. "Covertype DataSet“. *Center for Machine Learning Intelligent Systems - link: https://archive.ics.uci.edu/ml/datasets/covertype* and (visited 17.12.2017.)

[2] Blackard J.A., Dean D.J. "Comparative Accuracies of Artificial Neural Networks and Discriminant Analysis in Predicting Forest Cover Types from Cartographic Variables", *Computers and Electronics in Agriculture*, Vol. 24. No.3, pp: 131-151, 2000.

[3] Blackard J.A., Dean D.J. "Comparative Accuracies of Neural Networks and Discriminant Analysis in Predicting Forest Cover Types from Cartographic Variables", *Second Southern Forestry GIS Conference.* University of Georgia. Athens, GA. pp. 189-199, 1998.

[4] Blackard J.A., "Comparison of Neural Networks and Discriminant Analysis in Predicting Forest Cover Types (Ph.D. dissertation)", *Department of Forest Sciences. Colorado State University.* Fort Collins, 1998.

[5] Oreški, D., Novosel, T. "Comparison of Feature Selection Techniques in Knowledge Discovery Process", *TEM Journal,* Vol. 3, No. 4, pp. 285-290, 2014.

[6] Gama, J., Rocha, R., Medas P. "Accurate decision trees for mining high-speed data streams“, *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining,* pp.523-528, 2003.

[7] Giannella, C., Sayrafi, B. "An Information Theoretic Histogram for Single Dimensional Selectivity Estimation: Technical Report TR584“. *Department of Computer Science, Indiana University Bloomington,* 2005.

[8] Nikunj, C.O, Stuart J. Russell. "Experimental comparisons of online and batch versions of bagging and boosting“, *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 359-364, 2001.

[9] Obradovic, Z., Vucetic, S. "Challenges in Scientific Data Mining: Heterogeneous, Biased, and Large Samples: Technical Report“, *Center for Information Science and Technology Temple University*, ch. 1, pp. 1–24, 2004.

[10] Miškovic, V., "Sistemi za podršku odlučivanju: Decision support systems“, *Faculty of Informatics and Computing, Singidunum University in Belgrade*, 2013.

BIOGRAPHY OF AUTHORS

Muzafer Saračević (1984) - Associate professor of computer sciences at the University of Novi Pazar and dean of the Department of computer sciences at the same university. He defend bachelor thesis on Faculty of Informatics and Computing in Belgrade, master thesis on Faculty of Technical Sciences (University of Kragujevac) and 2013th year defend doctoral thesis on Faculty of Science and Mathematics (University of Niš). He specialized on Oracle academy in the field of database and programming. He is the author of over 130 professional and scientific papers, one monograph and two practicum. He is the author of one chapter in the scientific monograph of the international publisher. He is a member of the editorial board for five journals.

Aybeyan Selimi (1980) - Currently he is employed as a Teaching assistant at the International University Vision, Faculty of Informatics in Gostivar. Graduated in 2004 at the Institute of Mathematics in the Faculty of Natural Science and Mathematics in Skopje. In 2015 he defended his master's thesis entitled at field od optimization at the Institute of Mathematics, the Faculty of Natural Sciences and Mathematics in Skopje. Also he is PhD Candidate at the University of Novi Pazar, Serbia. Research interests is on the mathematical programming and computational geometry. He is the author of 20 professional and scientific papers.

Mersad Mujević (1965) - Head of Public Procurement Administration in Government of Montenegro. He defend bachelor thesis on Faculty of Traffic Science (Road Traffic and Security) at the University of Sarajevo (BIH). He defend master thesis on Department of Computer Sciences in Novi Pazar and doctoral thesis on University of Novi Pazar. He is the author of 30 professional and scientific papers.