ANALYSIS OF DEMINING PROJECT PROPOSALS

Matjaž Bevk, Dunja Mladenić
Department of Knowledge Technologies
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel: 386 1 4773127
e-mail: matjaz.bevk@ijs.si

ABSTRACT

This paper is an analysis of project proposals for mine clearance of mine-affected areas. Specifically we are studying the relationship between proposals and their evaluation scores influencing their funding possibilities. First we investigate the possibility of building a reliable predictive model for automatic evaluation score prediction from the attributes of the proposal, and second we analyze individual attributes and their trends with respect to the evaluation score of the proposals.

1 INTRODUCTION

The goal of this study is to investigate, how can machine learning methods help in a time consuming task of project proposal evaluation. Our problem domain are project proposals for mine clearance of mine-affected areas. The proposal evaluation procedure consists of three phases: administrative (checking the formal appropriateness), technical and financial. This study considers the technical evaluation only.

Technical evaluation involves analyzing several parameters which describe project team, animals (dogs), execution plan and equipment to be employed on a contract. The task is to evaluate these parameters in accordance with call for projects and rate the proposal with a numerical score from 0 to 100. The proposals with satisfactory technical score (above a predefined threshold) are further evaluated from financial perspective, which is not the subject of this paper.

The paper is organized in the following manner. In section 2 we present our solution plan and the procedure of data acquisition. Section 3 describes experiments and their results and in the last section we give the conclusions of this study.

2 SOLUTION PLAN AND DATA ACQUISITION

Our data base consists of 10 calls for projects and 39 proposals. The number of proposals per call varies from 3 to 6. Together with domain expert we defined a set of 23 attributes, which are important for technical evaluation. To obtain a data set as accurate as possible, we engaged a human operator to extract the attributes from the proposals which are originally unstructured text documents. We also added two attributes, which were extracted from the corresponding call for projects. These two attributes describe the contract value and the size of the area to be cleared. And finally, we also added the target concept, namely the attribute representing the evaluation score of the proposal assigned by an evaluation committee.

With data base constructed, we wanted to see if we can get some insights in the relation of evaluation score and the attributes describing content of the proposal. We are also interested in a predictive model and in attributes which are the most effective for prediction of the proposal technical evaluation score.

3 DATA ANALYSIS

Our data consists of 39 cases each described by 26 attributes. The target attribute (score) is a real value between 0 and 100, which means that we are dealing with a regression problem. Actually the score ranges from 69.2 to 80.6, which can be seen in Figure 1.

3.1 Target attribute transformation

Since we have small amount of cases (proposals), we decide to binarize the real-valued score attribute by dividing proposals that scored first and the ones that scored below the first. This gives us a binary target attribute, with majority class probability of 0.74.

3.2 Analysis with machine learning

Let us now apply machine learning algorithms to the data set to check how well can the binary score be predicted from the extracted attributes. For this purpose we used three algorithms SMO [3], j48 [4] and Naïve Bayes [5] from the WEKA machine learning software [6]. Since we have a relatively small data set, we use leave-one-out validation method to evaluate the obtained models.

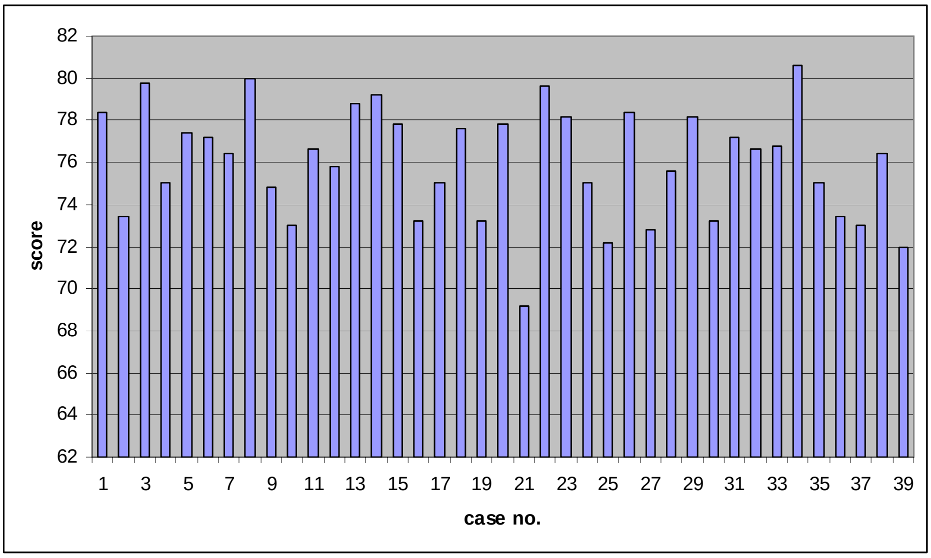

Figure 1: Graph showing values of the score attribute.

algorithm	accuracy	st. deviation
Baseline	74.4	44.2
J48	79.5	40.9
Naïve bayes	66.7	47.8
SMO	76.9	42.7

Table 1: Classification accuracies and standard deviations of j48, Naïve Bayes and SMO algorithm on our data set.

From the Table 1, we can see that neither of the three algorithms managed to improve the accuracy significantly over the baseline. Thus we conclude, that building a usable prediction model is not possible with the present data set. However, we can still take a peek at the models, for instance a decision tree built by j48 and investigate the most important attributes in that tree (Figure 2).

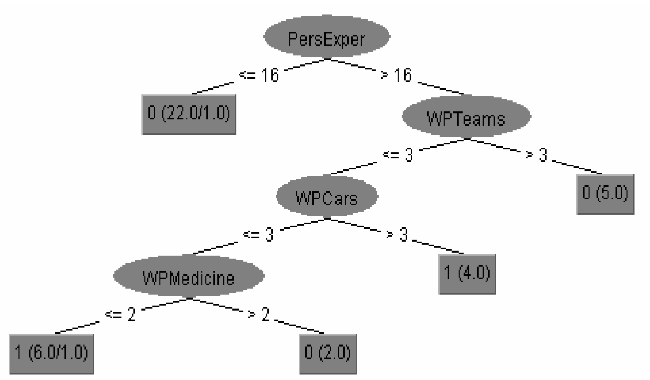

Figure 2: Decision tree built by j48.

In Figure 2, we see that the most important attribute is the experience of the personnel (PersExper). Figure 3 shows a graph containing the rank of the proposal against average experience of the personnel.

.

Graph in Figure 3 shows a trend which means that better personnel experience yields better rank in the evaluation of the proposal, which is also reasonable.

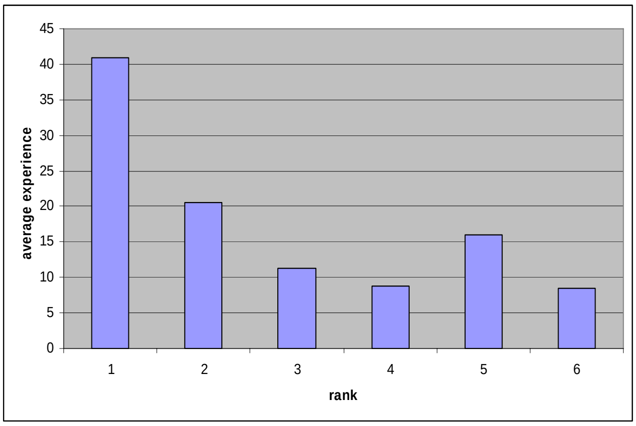

Figure 3: Graph showing the rank of the proposal against average experience of the personnel

3.3 Analysis of important attributes

In section 3.2 we tested several prediction models generated automatically from the available data, but using the available limited amount of data we did not get much further from the baseline prediction. However, by analysis of the generated models we found an interesting and reasonable trend in the data. Thus, we continue in that manner and analyze the most important attributes. In order to identify important attributes, we use one of the measures for attribute evaluation ReliefF [1,2]. Then, we plot the five most important attributes on graphs showing rank of the proposal against one of the selected attributes. Table 2 shows the top five attributes. Figure 4 through Figure 8 plot each of the top attributes against the average value of rank.

rank	Attribute
1	Dogs (number of dogs)
2	PersExper (experience of the personnel)
3	Helmets (number of helmets)
4	Vests (number of vests)
5	Cars (number of cars)

Table 2: Top five attributes as ordered by ReliefF.

Figure 4 is a graph showing a relationship between rank of the proposal against average number of dogs per rank. The trend here is clear: More dogs employed on a contract means more chances of getting a funding.

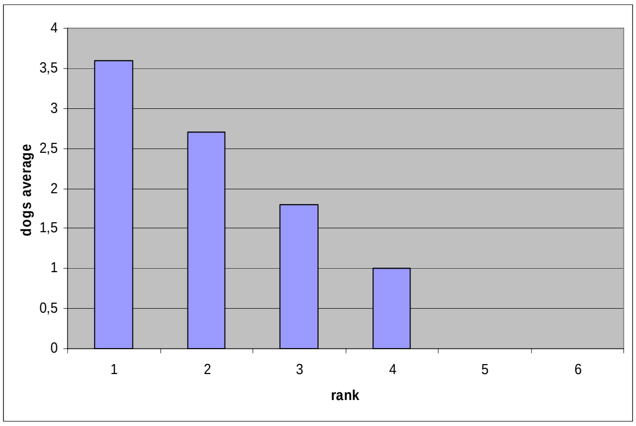

Figure 4: Graph showing the rank of the proposal against average number of dogs per rank.

The second attribute PersExper was already analyzed in section 3.2 and plotted in Figure 3. The trend shows that more experience of personnel means more chances of successfully getting funding for the proposal.

The third most important attribute is the number of helmets to be used on a contract. This attribute is plotted in graph in Figure 5. The trend indicates that, less helmets mean more chances of a success, which is a bit unusual. This fact could be explained away with a conclusion that excessive use of resources means less chances of a success in getting funding.

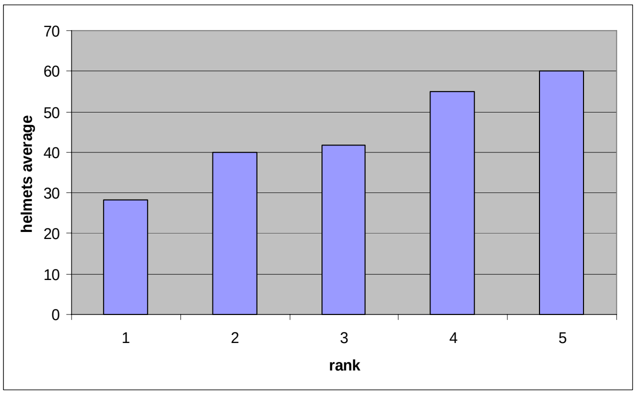

Figure 5: Graph showing the rank of the proposal against average number of helmets per rank.

The fourth attribute is the number of protective vests to be used on a contract. This attribute is plotted in a graph in Figure 6. The trend is similar to that for the helmets attribute. It shows that less protective vests means more chances of a success.

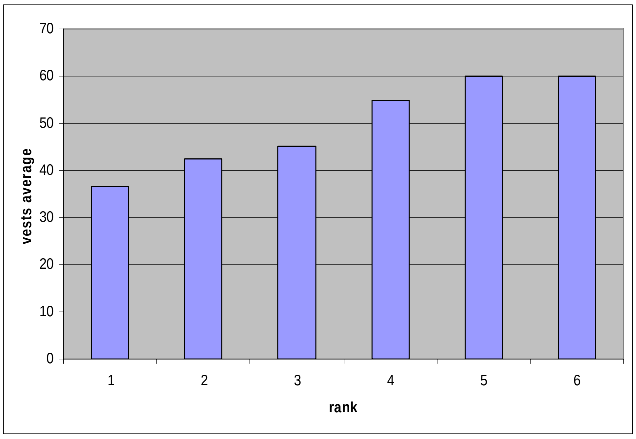

Figure 6: Graph showing the rank of the proposal against average number of protective vests per rank.

The fifth attribute is the number of cars to be used on a contract. The attribute is plotted in a graph in Figure 7. The trend here is not clear.

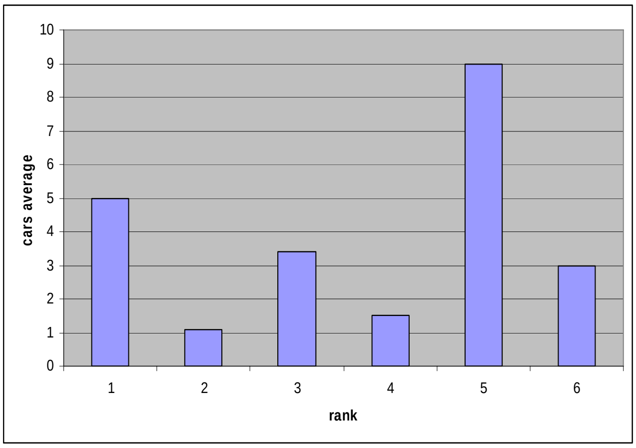

Figure 7: Graph showing the rank of the proposal against average number of cars per rank.

4 CONCLUSION

The subject of this paper is analysis of project proposals for mine clearance of mine-affected areas. We wanted to alleviate the time consuming task of project proposal evaluation.

The intention was to build a classifier which would predict proposal estimates from parameters which were previously extracted from documents. It turned out that, we were unable to construct a reliable classifier from our data set. One of the explanations for this fact could be a relatively small number of learning examples.

Further on, we analyzed five most important attributes according to ReliefF measure. We found out some interesting trends. The most important attribute was the number of dogs employed on a contract. The trend showed that employing more dogs helps achieving a high evaluation score. The second important attribute was experience of the personnel. It turned out that employing a more experienced personnel is also a very important factor of success. The third and fourth important attributes were the number of helmets and protective vests respectively. It turned out that using more equipment (helmets and vests) is not favorable for the success of the proposal. The fifth important attribute was the number of cars to be used on a contract. This attribute showed no apparent trend.

To give a concluding remark: We weren't able to obtain a good predictive model from our data set, but the analysis of the individual attributes showed some interesting trends, which seem to have an important influence on the evaluation score of the proposals.

Acknowledgement
This work was supported by the Slovenian Research Agency and the IST Programme of the European Community under PASCAL Network of Excellence (IST-2002-506778). This publication only reflects the authors' views. The authors would also like to thank The Slovenian International Trust Fund for providing the data and Dalija Sega for help in extracting attributes from the textual proposals.

References

[1] Kira, K. and Rendell, L. A. A practical approach to feature selection. In D. Sleeman and P. Edwards, editors, Proceedings of the International Conference on Machine Learning, pages 249-256, 1992, Morgan Kaufmann.

[2] Kononenko, I. Estimating attributes: analysis and extensions of Relief. In De Raedt, L. and Bergadano, F., editors, Machine Learning: ECML-94, pages 171-182, 1994, Springer Verlag.

[3] Platt, J. Fast Training of Support Vector Machines using Sequential Minimal Optimization. Advances in Kernel Methods - Support Vector Learning, B. Schoelkopf, C. Burges, and A. Smola, eds., 1998, MIT Press.

[4] Quinlan, R. C4.5: Programs for Machine Learning, 1993, Morgan Kaufmann Publishers, San Mateo, CA.

[5] George H. John and Pat Langley Estimating Continuous Distributions in Bayesian Classifiers. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. pp. 338-345, 1995, Morgan Kaufmann, San Mateo.

[6] Witten I. and Frank E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, 1999, Morgan Kaufman.