Use of data mining tools for cut soil slope condition state identification

Joaquim Tinoco[1,2,*], António Gomes Correia[1], Paulo Cortez[2], and David Toll[3]
[1]*ISISE, School of Engineering, University of Minho, Campus de Azurém, Guimarães, Portugal*
[2]*ALGORITMI Research Centre/Department of Information Systems, University of Minho, Guimarães, Portugal*
[3]*School of Engineering and Computing Science, University of Durham, Durham, UK*
jtinoco@civil.uminho.pt, agc@civil.uminho.pt, pcortez@dsi.uminho.pt, d.g.toll@durham.ac.uk

1 Introduction

Transportation systems play a fundamental rule in nowadays society. Indeed, every developed or in development country had invested and keep investing to build a complete, safe and functional transportation network. Now, the main concern, particularly for developed countries, is to keep it operational under all security conditions. However, due to the network extension and increased budget constraints, such task is difficult to accomplish. In the framework of transportations networks, particularly for highway and railway, slopes are perhaps the element for which its failure can have a strongest impact at several levels. Although there are some models and systems to detect slop failures, most of them were developed for natural slopes, presenting some constrains when applied to man-made slopes. Moreover, most of the existent systems were developed based on particular case studies or require information gathered from complex/expensive tests, which can represent an important applicability limitation.

Aiming to overcome this drawback, we are taking advantage of the learning capabilities of flexible DM algorithms, such Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs), which can model complex nonlinear mappings. Both algorithms were fitted to predict the condition state of a given slope according to a pre-defined classification scale contemplating four levels (classes). One of the premises of this work is to try to identify the real condition state of a given slop using information collected during routine inspections complemented with geometric, geologic and geographic data.

2 Data and Models

In this work, a model based on DM tools is proposed to identify the condition state, from this point referred as EHC (Earthwork Hazard Category), for cut soil slopes, based on 4 classes (“A”, “B”, “C” and “DE”). For that, a database with 14978 records was compiled, containing information collected during routine inspections and complemented with geometric, geologic and geographic data of each slope. All information was supplied by *NetworRail* and is concerning to the railway network of UK. Analyzing the distribution of EHC classes, it was possible to observe a high asymmetric distribution (unbalanced data). Indeed, around 60% of the slopes are classified as “A” and only 1% as “DE”. Although this type of asymmetric distribution, where most of the slopes present a low probability of failure (class “A”), is normal and desired from the safety point of view, it can represent an important challenge for DM models learning. The proposed model is feed by 101 variables (such as: high, slope, animal activity, boulders present, ground cover, etc.) usually collected during routine inspections and complemented with geometric, geographic and geologic information.

To model EHC prediction of cut soil slopes two of the most flexible DM algorithms, namely ANNs and SVMs were applied. For ANNs (Kenig et al., 2001) we adopted the multilayer perceptron with feedforward connections and one hidden layer containing H processing units. To find the best H value a grid search of {0; 2; 4; 6; 8} was adopted. The neural function of the hidden nodes was set to the popular logistic function $1/(1+e^{-x})$. For SVMs (Cortes and Vapnik, 1995), the popular Gaussian kernel was adopted. In this context, its performance is affected by three parameters: γ, the parameter of the kernel; C, a penalty parameter; and ε (only for regression), the width of a ε-insensitive zone. The heuristics proposed by Cherkassky and Ma (2004) were used to define the first two parameter values. A grid search of {1; 3; 7; 9} was adopted to optimize the kernel parameter γ.

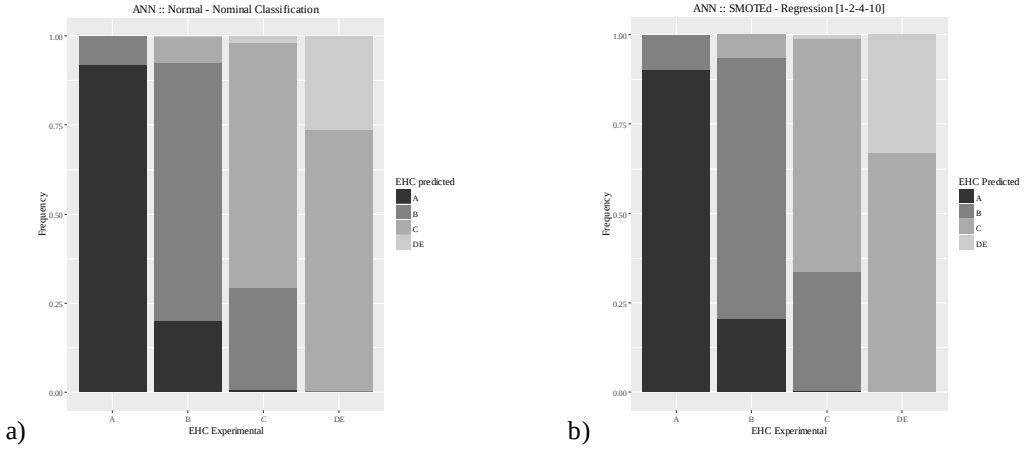
* Corresponding author e-mail: jtinoco@civil.uminho.pt

3 Results

EHC prediction was initially approached following a nominal classification strategy. However, aiming a better performance, it was also addressed following a regression strategy, adopting a regression scale "A" = 1, "B" = 2, "C" = 4, "DE" = 10. Moreover, in order to overcome the problem of unbalanced data, three training sampling approaches were explored: Normal (no resampling), OVERed (Oversampling (Ling and Li, 1998)) and SMOTEd (SMOTE (Chawla et al., 2002)). Although the overall performance shown in Table 1 is not so high, Figure 1 illustrates a very interesting result. In fact, following a regression strategy and applying a SMOTE approach (Figure 1b), ANN model can identify classes "A", "B" and "C" very accurately. Moreover, concerning to cut soil slopes of class "DE", if not correctly identified as it, they are classified as belong to class "C", which is the closest class. This type of misclassification is also observed for classes "A", "B" and "C", which can be seen as a model advantage.

Table 1: Metrics values in EHC prediction of cut soil slopes

	Model	Approach	AUS	Recall				Precision				F1-Score			
				A	B	C	DE	A	B	C	DE	A	B	C	DE
Classification	ANN	Normal	**0.3**	92.14	72.28	68.92	26.30	90.61	**71.81**	**74.09**	47.02	**91.37**	**72.04**	71.41	33.73
		SMOTEd	0.26	84.31	72.49	74.02	**60.85**	**93.80**	63.82	63.85	40.66	88.80	67.88	68.56	48.75
		OVERed	**0.3**	86.24	71.39	**78.09**	61.04	92.83	66.10	66.74	56.57	89.41	68.64	**71.97**	**58.72**
	SVM	Normal	0.07	91.40	74.96	28.97	30.37	89.71	61.83	68.24	**64.31**	90.55	67.76	40.67	41.26
		SMOTEd	0.05	82.74	**84.57**	29.91	26.44	93.01	56.60	71.85	60.46	87.57	67.80	42.24	36.79
		OVERed	-1.17	**92.86**	26.60	0.56	0.33	65.63	48.49	73.75	81.82	76.91	34.35	1.11	0.66
Regression	ANN	Normal	0.23	**92.89**	65.48	**68.15**	48.74	87.85	**70.78**	75.54	65.25	90.30	68.03	**71.65**	**55.80**
		SMOTEd	**0.26**	90.46	73.08	65.40	32.96	90.38	68.32	75.42	**69.37**	90.42	70.62	70.05	44.69
	SVM	Normal	0.19	88.47	82.36	43.05	0.00	92.50	63.17	76.13	NA	90.44	71.50	55	NA
		SMOTEd	0.21	88.32	**83.04**	45.47	0.52	**92.79**	63.68	**77.33**	100	**90.50**	**72.08**	57.27	1.04



a) b)

Figure 1: Models performance comparison in EHC prediction of cut soil slopes.

References

Kenig, S., Ben-David, A., Omer, M., Sadeh, A., 2001. Control of properties in injection molding by neural networks. Engineering Applications of Artificial Intelligence 14, 819 – 823.

Cortes, C., Vapnik, V., 1995. Support vector networks. Machine Learning 327 20, 273 – 297.

Cherkassky, V., Ma, Y., 2004. Practical selection of svm parameters and noise estimation for svm regression. Neural Networks 17, 113 – 126.

Ling, C.X., Li, C., 1998. Data mining for direct marketing: Problems and solutions, in: KDD, pp. 73 – 79.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research , 321 –