Data-driven Classification Approaches for Stability Condition Prediction of Soil Cutting Slopes

Approches de classification de données pour la prévision des conditions de stabilité de pentes d'excavation

Joaquim Tinoco & António Gomes Correia
ISISE - Institute for Sustainability and Innovation in Structural Engineering, School of Engineering, University of Minho, Portugal, jtinoco@civil.uminho.pt

Paulo Cortez
ALGORITMI Research Center/Department of Information Systems, University of Minho, Portugal

David Toll
School of Engineering and Computing Sciences, University of Durham, UK

ABSTRACT: For transportation infrastructures, one of the greatest challenges today is to keep large-scale transportation networks, such as railway networks, operational under all conditions. In this paper we present a tool aimed at helping in management tasks related to maintenance and repair works for a particular component of these infrastructures, the slopes. For that, the high and flexible learning capabilities of artificial neural networks and support vector machines were applied in the development of a tool able to identify the stability condition of soil cutting slopes, keeping in mind the use of information usually collected during routine inspection activities (visual information) to feed the models. This task was addressed following two different strategies: nominal classification and regression. Moreover, to overcome the problem of imbalanced data, three training sampling approaches were explored: no resampling, SMOTE and Oversampling. The achieved results are presented and discussed, comparing both algorithms performance as well as the effect of the sampling approaches. A comparison between nominal classification and regression strategies is also carried out. These achieved results can give a valuable contribution for practical applications at network level.

RÉSUMÉ : Pour les infrastructures de transport, l'un des plus grands défis d'aujourd'hui est de maintenir les réseaux de transport à grande échelle, tels que les réseaux ferroviaires, opérationnels dans toutes les conditions. Dans cet article, on présente un outil d'aide à la gestion liée aux travaux de maintenance et de réparation d'une composante géotechnique de ces infrastructures, les pentes. Pour cela, les capacités d'apprentissage élevées et flexibles des réseaux de neurones artificiels et des machines à vecteurs de support ont été appliquées dans l'élaboration d'un outil capable d'identifier l'état de stabilité des pentes d'excavation du sol en gardant à l'esprit l'utilisation des informations recueillies habituellement lors des activités d'inspection de routine visuelles pour alimenter les modèles. Cette tâche a été abordée selon deux stratégies différentes: la classification nominale et la régression. De plus, pour surmonter le problème des données déséquilibrées, trois méthodes d'échantillonnage ont été explorées: non ré-échantillonnage, SMOTE et sur-échantillonnage. Les résultats obtenus sont présentés et discutés, en comparant les performances des algorithmes ainsi que l'effet des approches d'échantillonnage. Une comparaison entre les stratégies de classification nominale et de régression est également réalisée. Les résultats obtenus peuvent apporter une contribution précieuse aux applications pratiques au niveau du réseau.

KEYWORDS: slope stability condition, soil cutting slopes, railway, soft computing, data mining, imbalanced data.

1 INTRODUCTION

For a good optimization of the available budgets it is important to have a set of tools to help decision makers to take the best decisions. In the framework of transportations networks, in particular for a railway, slopes are perhaps the element for which their failure can have the strongest impact at several levels. Therefore, it is important to develop ways to identify potential problems before they result in failures.

Although there are some models and systems to detect slope failures, most of them were developed for natural slopes, presenting some constraints when applied to engineered (human-made) slopes. They have limited applicability as most of the existing systems were developed based on particular case studies or using small databases. Furthermore, another aspect that can limit its applicability is related with the information required to feed them, such as data taken from complex tests or from expensive monitoring systems. Some approaches found in the literature for slope failure detection are identified below. Pourkhosravani and Kalantari (2011) summarize the current methods for slope stability evaluation, which were grouped into Limit Equilibrium (LE) methods, Numerical Analysis methods, Artificial Neural Networks and Limit Analysis methods. There are also approaches based on finite elements methods (Suchomel et al., 2010), reliability analysis (Husein Malkawi et al., 2000), as well as some methods making use of data mining (DM) algorithms (Cheng and Hoang, 2014; Ahangar-Asr et al., 2010; Yao et al., 2008). More recently, a new flexible statistical system was proposed by Pinheiro et al. (2015), based on the assessment of different factors that affect the behavior of a given slope. By weighting the different factors, a final indicator of the slope stability condition is calculated.

As mentioned above, the main limitations of almost approaches so far proposed are related with its applicability domain or dependency on information that is difficult to obtain. Indeed, the prediction of whether a slope will fail or not is a multi-variable problem characterized by a high dimensionality.

In this work we take advantage of the learning capabilities of flexible data mining classification algorithms, such as the Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs). These data mining algorithms were used to fit a large database of soil cutting slopes in order to predict the stability condition of a given slope according to a pre-defined

classification scale based on four levels (classes). One of the underlying premises of this work is to identify the real stability condition of a given slop based on information that can be easily obtained through visual routine inspections. For that, more than fifty variables related with data collected during routine inspections as well as geometric, geological and geographic data were used to feed the models. This type of visual information is sufficient from the point of view of the network management, allowing the identification of critical zones for which more detailed information can then be obtained in order to perform more detailed stability analysis, which is out of the scope of this study.

2 DATA AND METHODOLOGY

2 .1 *Data characterization*

In this work a model is proposed to identify the stability condition, from this point referred to as EHC (Earthwork Hazard Category, adapted from Power et al. (2016)), of soil cutting slopes using data mining tools.

The EHC system comprises 4 classes ("A", "B", "C" and "D") where "A" represents a good stability condition and "D" a bad stability condition. In other words, the expected probability of failure is higher for class "D" and lower for class "A". To fit the models for EHC prediction, a database was compiled containing information collected during routine inspections and complemented with geometric, geological and geographic data of each slope. The database was gathered by Network Rail workers and is concerned with the railway network of the UK. For each slope a class of the EHC system was defined by the Network Rail Engineers based on their experience/algorithm (Power et al., 2016), which will be assumed as a proxy for the real stability condition of the slope for the year 2015.

Figure 1 depicts the distribution of the 10928 records by each EHC class. From this analysis, it is possible to observe an asymmetric distribution (imbalanced data). Indeed, more than 57% of the soil cutting slopes are classified as "A" and only about 1% belongs to class "D". Although this type of asymmetric distribution, where most of the slopes present a low probability of failure (class "A"), is normal and desirable from the safety point of view and slope network management, it can represent an important challenge for data mining models learning, as detailed in next section.

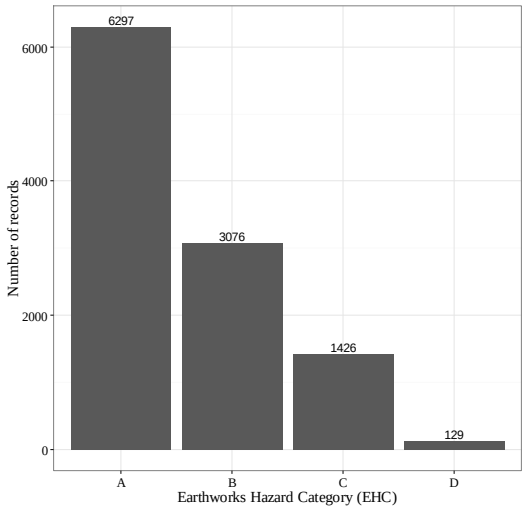

Figure 1. Soil cutting slopes data distribution by EHC classes.

2 .2 *Modeling*

To model EHC prediction of soil cutting slopes two of the most flexible data mining algorithms, namely ANNs and SVMs were applied. Both algorithms had already been successful applied in different knowledge domains (Liao et al., 2012) including in civil engineering (Tinoco et al., 2014a,b). There are also some examples of ANN and SVM applications in slope stability analysis (Yao et al., 2008; Cheng et al., 2012).

ANNs try to simulate basic aspects of the human brain (Kenig et al., 2001). The information is processed using iteration among several neurons. This technique is capable of modeling complex non-linear mappings and is robust in exploration of data with noise. In this study we adopt the multi-layer perceptron that contains only feedforward connections, with one hidden layer containing *H* processing units. Because the network's performance is sensitive to *H* (a trade-off between fitting accuracy and generalisation capability), we adopt a grid search of {0; 2; 4; 6; 8} during the learning phase to find the best *H* value. The neural function of the hidden nodes was set to the popular logistic function $1/(1+e^{-x})$.

SVMs was initially proposed for classification tasks (Cortes and Vapnik, 1995). Then it became possible to apply SVM to regression tasks after the introduction of the ε-insensitive loss function (Smola and Scholkopf, 2004). The main purpose of the SVM is to transform input data into a high-dimensional feature space using non-linear mapping. The SVM then finds the best linear separating hyperplane, related to a set of support vector points, in the feature space. This transformation depends on a kernel function. In this work 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 an ε-insensitive zone (Safarzadegan Gilan et al., 2012). The heuristics proposed by Cherkassky and Ma (2004) were used to define the first two parameter values, C=3 (for a standardized output) and $\varepsilon = \hat{\sigma}/\sqrt{N}$, where $\hat{\sigma} = 1.5/N \cdot \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$, y_i is the measured value, $\hat{y}_i$ is the value predicted by a 3-nearest neighbour algorithm and *N* is the number of examples. A grid search of {1; 3; 7; 9} was adopted to optimize the kernel parameter γ.

The problem of EHC prediction of soil cutting slopes was initially approached following a nominal classification strategy. However, aiming to improve the models performance, the problem was also addressed following a regression strategy, adopting a regression scale where "A" = 1, "B" = 2, "C" = 4, "D" = 10. Moreover, in order to minimize the effect of the imbalanced data (see Figure 1), Oversampling (Ling and Li, 1998) and SMOTE (Chawla et al., 2002) approaches were applied over the training data before fitting the models. When approaching imbalanced classification tasks, where there is at least one target class label with a smaller number of training samples when compared with other target class labels, the simple use of a data mining training algorithm will lead to data-driven models with better prediction accuracies for the majority classes and worst classification accuracies for the minority classes. Thus, techniques that adjust the training data in order to balance the output class labels, such as Oversampling and SMOTE, are commonly used with imbalanced datasets. In particular, Oversampling is a simple technique that randomly adds samples (with repetition) of the minority classes to the training data, such that the final training set is balanced. SMOTE is a more sophisticated technique that creates "new data" by looking at nearest neighbours to establish a neighbourhood and then sampling from within that neighbourhood. It operates on the assumptions that the original data is similar because of proximity. More recently, Torgo et al. (2015) adapted the SMOTE method for regression tasks.

All experiments were conducted using the R statistical environment (Team, 2009) and supported through the rminer

package (Cortez, 2010), which facilitates the implementation of ANNs and SVMs algorithms, as well as different validation approaches such as cross-validation.

The distinct data mining models will be evaluated and compared using three classification metrics: recall, precision and F1-score. The recall measures the ratio of how many cases of a certain class were properly captured by the model. In other words, the recall of a certain class is given by *TruePositives*/(*TruePositives*+*FalseNegatives*). On the other hand, the precision measures the correctness of the model when it predicts a certain class. More specifically, the precision of a certain class is given by *TruePositives*/(*TruePositives* + *FalsePositives*). The F1-score was also calculated, which represents a trade-off between the recall and precision of a class. The F1-score corresponds to the harmonic mean of precision and recall. For all four metrics, the higher the value, the better are the predictions, and their values can range from 0% to 100%. The generalization capacity of the models was accessed through a 5-fold cross-validation approach under 20 runs (Hastie et al., 2009). This means that each modeling setup is trained $5 \times 20 = 100$ times. Also, the four prediction metrics are always computed on test unseen data (as provided by the 5-fold validation procedure).

3 RESULTS AND DISCUSSION

As described above, two different data mining algorithms (ANN and SVM) were applied for EHC prediction under two distinct modeling strategies: nominal classification and regression. Moreover, in order to overcome the problem of unbalanced data, three training sampling approaches were explored: Normal (no resampling), OVERed (Oversampling) and SMOTEd (SMOTE). In case of regression, we compared two sampling approaches: Normal (no resampling) and SMOTEd (SMOTE for regression). We note that the different sampling approaches were applied only to training data, used to fit the data-driven models, and the test data (as provided by the 5-fold procedure) was kept without any change.

Figures 2 and 3 show and compare models performance in EHC prediction of soil cutting slopes based on metrics recall, precision and F1-score, following a nominal classification and regression strategies respectively. SMOTE and Oversampling approaches are also compared.

Following a nominal classification strategy, Figure 2 shows that soil cutting slopes of class “A” can be correctly identified, particularly by ANN model, with or without sampling. Also for classes “B” and “C” a promising performance is observed, with an F1-score around 55%, in particular by the ANN algorithm. Concerning the class “D”, although an F1-score lower than 36% was achieved, the obtained value for recall metric around 57% shows a promising performance for class “D” prediction according to ANN algorithm.

Based on a regression strategy (Figure 3), the achieved results are very similar to those obtained from a nominal classification strategy. The main differences are related with the effect of the sampling approaches, which is not so relevant following a regression strategy, particularly for the minority classes. Comparing ANN and SVM algorithms, ANN works better (as observed previously), particularly in the prediction of class “C” and “D” classes.

Comparing both strategies, we can observe that better results are achieved following a nominal classification strategy. Moreover, analyzing Figures 4 and 5 that show the relation between observed and predicted EHC values according to the best fits, following a nominal classification and regression strategies respectively, we can see that the models performance is very promising. Indeed, according to a nominal classification strategy and sampling the database with the SMOTE approach (see Figure 4), the ANN algorithm is able to predict correctly around 57% of soil cutting slopes of class “D”, which represents a very promising performance if we take into account that this is the minority class. For class “C”, around 40% of the records are correctly predicted. Moreover, when not predicted as “C” they are classified as belonging to the closest class, that is, “B” or “D”. This type of misclassification is also observed for classes “A”, “B” and “D”, which can be interpreted as an advantage. Concerning to classes “A” and “B”, the ANN model was also able to identify it very accurately.

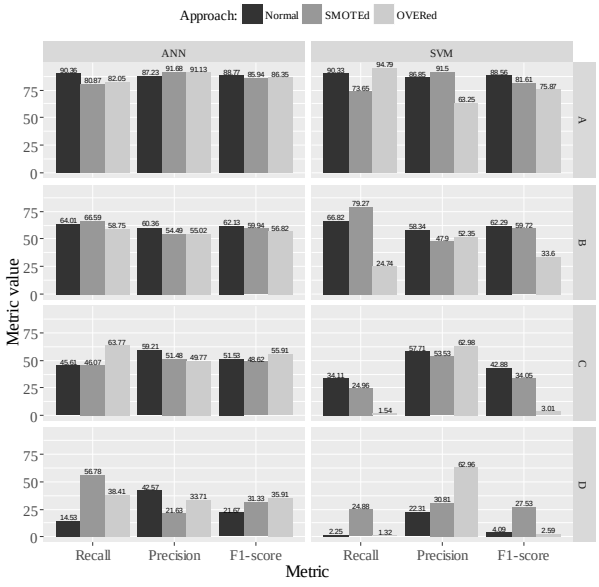

Figure 2. Model comparison based on recall, precision and F1-score, according to a nominal classification strategy in EHC prediction of soil cutting slopes.

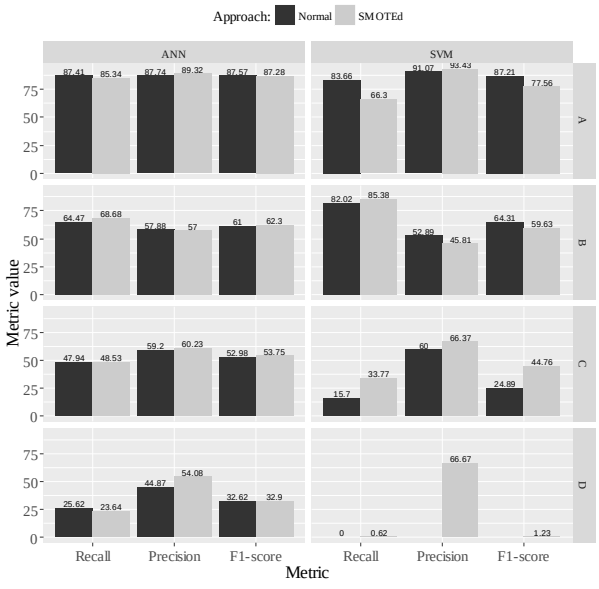

Figure 3. Models comparison based on recall, precision and F1-score, according to a regression strategy in EHC prediction of soil cutting slopes.

As a conclusion, an attempt to predict Earthwork Hazard Classification (EHC) of soil cutting slopes through the application of data mining techniques was presented. Although a very promising performance have been achieved, namely for class “A”, further developments are required aiming to improve

models performance, particularly for a better identification of soil cutting slopes that belong to class "D" (minority class).
Moreover, and considering the overall performance of all models, we would like to stress that data mining algorithms, particularly ANNs, can be seen as a tool with a high potential to identify stability condition of engineered slopes.

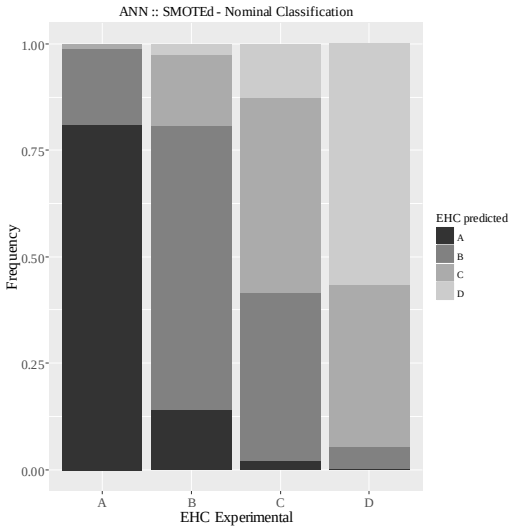

Figure 4. ANN model performance in EHC prediction of soil cutting slopes following a nominal classifications strategy and applying a SMOTE approach.

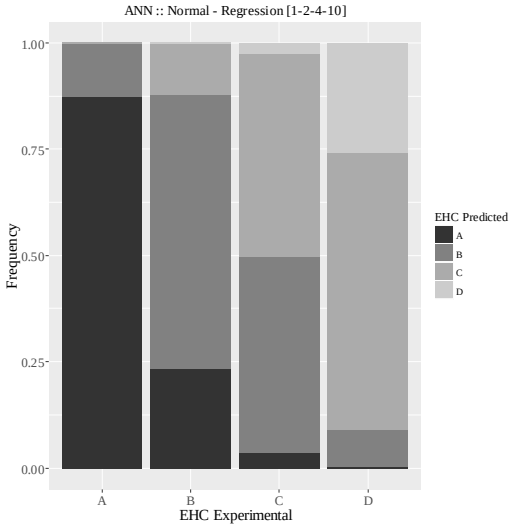

Figure 5. ANN model performance in EHC prediction of soil cutting slopes following a regression strategy and with no resampling.

4 ACKNOWLEDGEMENTS

This work was supported by FCT – "Fundação para a Ciência e a Tecnologia", within ISISE, project UID/ECI/04029/2013 as well Project Scope: UID/CEC/00319/2013 and through the post-doctoral Grant fellowship with reference SFRH/BPD/94792/2013. This work was also partly financed by FEDER funds through the Competitivity Factors Operational Programme - COMPETE and by national funds through FCT within the scope of the project POCI-01-0145-FEDER-007633. This work has been also supported by COMPETE: POCI-01-0145-FEDER-007043. A special thanks goes to Network Rail that kindly make available the data (basic earthworks examination data and the Earthworks Hazard Condition scores) used in this work.

5 REFERENCES

Ahangar-Asr A., Faramarzi A. and Javadi A.A. 2010. A new approach for prediction of the stability of soil and rock slopes. *Engineering Computations* 27, 878-893.

Chawla N.V., Bowyer K.W., Hall L.O. and Kegelmeyer W.P. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16, 321-357.

Cheng M.Y. and Hoang N.D. 2014. Slope collapse prediction using bayesian framework with k-nearest neighbor density estimation: Case study in taiwan. *Journal of Computing in Civil Engineering* 30, 1-8.

Cheng M.Y., Roy A.F. and Chen K.L. 2012. Evolutionary risk preference inference model using fuzzy support vector machine for road slope collapse prediction. *Expert Systems with Applications* 39, 1737-1746.

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

Cortes C. and Vapnik V. 1995. Support vector networks. *Machine Learning* 20, 273-297.

Cortez P. 2010. Data mining with neural networks and support vector machines using the r/rminer tool, in: Perner, P. (Ed.), *Advances in Data Mining: Applications and Theoretical Aspects, 10th Industrial Conference on Data Mining*, LNAI 6171, Springer, Berlin, Germany. pp. 572-583.

Hastie T., Tibshirani R. and Friedman J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag New York. second edition.

Husein Malkawi A.I., Hassan W.F. and Abdulla F.A. 2000. Uncertainty and reliability analysis applied to slope stability. *Structural Safety* 22, 161-187.

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

Liao S., Chu P. and Hsiao P. 2012. Data mining techniques and applications. A decade review from 2000 to 2011. *Expert Systems with Applications* 39, 11303-11311.

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

Pinheiro M., Sanches S., Miranda T., Neves A., Tinoco J., Ferreira, A. and Gomes Correia A. 2015. A new empirical system for rock slope stability analysis in exploitation stage. *International Journal of Rock Mechanics and Mining Sciences* 76, 182-191.

Pourkhosravani A. and Kalantari B. 2011. A review of current methods for slope stability evaluation. *Electronic Journal of Geotechnical Engineering* 16.

Power C., Mian J., Spink T., Abbott S. and Edwards M. 2016. Development of an evidence-based geotechnical asset management policy for network rail, Great Britain. *Procedia Engineering* 143, 726-733.

Safarzadegan Gilan S., Bahrami Jovein H. and Ramezanianpour A. 2012. Hybrid support vector regression-particle swarm optimization for prediction of compressive strength and rcpt of concretes containing metakaolin. *Construction and Building Materials* 34, 321-329.

Smola A. and Scholkopf B. 2004. A tutorial on support vector regression. *Statistics and Computing* 14, 199-222.

Suchomel R. et al. 2010. Comparison of different probabilistic methods for predicting stability of a slope in spatially variable soil. *Computers and Geotechnics* 37, 132-140.

Team R. 2009. R: *A language and environment for statistical computing*. R Foundation for Statistical Computing. Viena, Austria. Web site: http://www.r-project.org/.

Tinoco J., Gomes Correia A. and Cortez P. 2014a. A novel approach to predicting Young's modulus of jet grouting laboratory formulations over time using data mining techniques. *Engineering Geology* 169, 50-60.

Tinoco J., Gomes Correia A. and Cortez P. 2014b. Support vector machines applied to uniaxial compressive strength prediction of jet routing columns. *Computers and Geotechnics* 55, 132-140.

Torgo L., Branco P., Ribeiro R. and Pfahringer B. 2015. Resampling strategies for regression. *Expert Systems* 32, 465-476.

Yao X., Tham L. and Dai F. 2008. Landslide susceptibility mapping based on support vector machine: a case study on natural slopes of hong kong, china. *Geomorphology* 101, 572-582.