

Using Ordinal Support Vector Machines to Model the Risks Associated with the Transportation of Hazardous Goods

J. M. Matías¹, C. Ordóñez², J. Taboada²

¹ Dep. of Statistics, University of Vigo, Spain, e-mail: jmmatias@uvigo.es

² Dep. of Environmental Management, University of Vigo, Spain.

^{1,2}E.T.S. de Ingenieros de Minas. Campus Lagoas-Marcosende 9. 33200 Vigo. Spain.

Abstract

The transportation of hazardous goods by road implies a risk for both humans and the environment in that an accident involving a vehicle transporting this kind of material may cause extensive material and environmental damage and might even endanger lives.

This study presents a model designed to assess the human, material and environmental consequences of an accident on a particular stretch of road on the basis of different variables associated with the characteristics of each segment.

Given the number of factors determining the state of a particular road segment, the only viable statistical methods for implementing a non-linear risk model are machine learning techniques, such as multilayer perceptron networks (MLPs), regression trees and support vector machines (SVMs). The best results were obtained using SVMs for ordinal data following the approach described by [Herbrich, Graepel and Obermayer, 2000].

Introduction

The transportation of hazardous goods by road implies a risk for both humans and the environment, in that an accident may cause extensive material damage and may even endanger lives. Recently, Martínez-Alegría et al. (2003) proposed a macroscopic conceptual model for identifying the roads within a network with the greatest accident risk. This conceptual model responded to the following mathematical model for each kind of road:

$$R = G * P$$

where R is road accident risk, G is the gravity of the potential damage and P is the probability of the adverse event occurring.

An absence of statistical data at the level of the stretch of roadway, however, obliges us to shift the focus of the model, to base it solely on factors determined by experts. More specifically, the historical statistical factors that make up P must be replaced by new factors representative of the construction morphology of each stretch of roadway, its design and condition, as well as certain geographic accidents of the location.

Given the dimensions of the problem (we have used a total of 33 factors), the only viable statistical methods for implementing a non-linear risk model are machine learning techniques such as classification and regression trees (CARTs), multilayer perceptron networks (MLPs) and support vector machines (SVMs).

Definition of the model

Our model takes the following form:

$$R = f(X_1, \dots, X_d)$$

where X_i , $i = 1, \dots, d$ are the 33 factors used to model the risk R associated with an accident on a particular stretch of roadway and f reflects expert knowledge on the influence on risk of the above-mentioned factors.

Thirty-three impact factors were identified and subsequently subdivided into two main groups, as follows:

- Factors that affect the probability of the occurrence of an accident (total 21) related to design, constructive morphology, signalling and signposting, type of road works, visibility threshold and condition of the road.
- Environmental vulnerability factors (total 12) related to land use, natural morphology of the land and surface and subterranean hydrology.

The variables X_i , $i = 1, \dots, 33$ were coded on ordinal scales of 0 to 10, with the lower values representing greater risk. Risk, in turn, was codified on an ordinal scale of 1 to 3 to indicate the hazard represented by a stretch of road (the higher the rating, the greater the number of corrective actions required).

Our approach (Matías et al. 2004) involved the use of linear discriminant analysis, neural networks, multilayer perceptrons (MLPs), classification trees (CART) and support vector machines (SVMs).

Assuming a sample of independent observations $\{(X_i, Y_i)\}_{i=1}^n$ where $X_i \in \Omega \subset \mathbb{R}^d$, $Y_i \in \Theta$ are random variables and where $\Theta = \{\tau_1, \dots, \tau_c\}$ is a set of ordered ranks $\tau_i > \tau_j$ if $i > j$ such that $\tau_c \succ \tau_{c-1} \succ \dots \succ \tau_1$ where \succ is a preference relation with strict order properties (irreflexive, asymmetric and transitive). Rather than a classical loss function $\ell_{0-1}(y, \hat{y}) = I_{\{y \neq \hat{y}\}}$ that just penalizes classification errors, with a view to penalizing violations in the order produced by an ordering rule $g : \Omega \rightarrow \Theta$ with $\hat{y} = g(\mathbf{x})$, we define the following loss function (Herbrich et al. 2000):

$$\ell_{\text{pref}}(y_i, y_j, \hat{y}_i, \hat{y}_j) = \begin{cases} 1 & \text{if } y_i \prec y_j \text{ and not } \hat{y}_i \prec \hat{y}_j \\ 1 & \text{if } y_j \prec y_i \text{ and not } \hat{y}_j \prec \hat{y}_i \\ 0 & \text{otherwise} \end{cases}$$

Results and discussion

With a view to construct the knowledge base represented by the model, 28.6 km of roadway located between the Spanish regions of Castilla-León and Galicia were selected for modeling.

Two hundred and eighty-six stretches of 100 m were marked out. For each stretch the factors X_i , $i = 1, \dots, 33$ defined above were evaluated, as also the level R of corrective measures necessary for adaptation of the stretch to the transportation of hazardous goods. Obtained as a result were 286 records of the form (X_1, \dots, X_{33}, R) . Of these 150 were used for the estimation of the model and 136 were used as a test sample to evaluate the behavior of the different techniques.

The results are depicted in Table 1, which also includes for reference purposes, the results obtained using different classification methods.

The error percentages for the different techniques were, respectively: 26.47% (linear discriminant analysis), 15.44% (CART), 14.71% (MLP), 14.71% (SVM) and 13.24% (ordinal SVM).

Table 1. Results of the classification

		Linear Disc.			Class. Tree			MLP			SVM			Ord. SVM		
Risk		L	M	H	L	M	H	L	M	H	L	M	H	L	M	H
expert	L	17	8	0	21	4	0	19	6	0	20	5	0	24	1	0
	M	19	12	3	11	23	0	8	26	0	7	22	5	11	23	0
	H	0	6	71	0	6	71	0	6	71	0	3	74	0	6	71

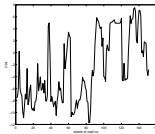


Figure 1: Utility function for the stretches of road in the training sample numbered sequentially.

Conclusions

In this paper we have constructed a model to determine the risk of a hazardous goods accident that incorporates expert knowledge. The model can be applied on a large scale to other roads without the direct intervention of an expert, although subsequent supervision would be necessary.

To estimate the model, a support vector machine approach applied to ordinal data was compared to linear discriminant analysis and other machine learning classification techniques. The ordinal SVMs performed more satisfactorily than the other techniques, and without increasing the computational burden to any significant extent. Moreover, they provided an estimation of both the expert latent utility function and the decision rule used to determine the level of risk for each stretch of roadway.

The positive results would demonstrate the benefits of tackling such problems as ordinal regression problems rather than as mere classification problems, which focus on classification errors and fail to penalize inversion in the order of the ex-amples.

References

- Fishburn PC (1985) Interval orders and interval graphs. John Wiley & Sons.
- Herbrich R, Graepel T, Obermayer K (2000) Large margin rank boundaries for ordinal re-gression. In: Advances in Large Margin Classifiers. MIT Press, pp 115-132.
- Martínez-Alegría R, Ordóñez C, Taboada J (2003) A conceptual model for analyzing the risks involved in the transportation of hazardous goods: Implementation in a geo-graphic information system. Human and Ecological Risk Assessment 9: 857-873.
- Matías J, Saavedra A, Taboada J, Ordóñez C (2004) SVM and neural networks for model-ling the risks involved in the transportation of hazardous goods. Technical Report Uni-versidad de Vigo, Spain.
- Schölkopf B, Smola AJ (2002) Learning with kernels. The MIT Press.
- Vapnik V (1998) Statistical learning theory. John Wiley & Sons.