An exploratory neural network model for predicting disability severity from road traffic accidents in Thailand

Jaratsri Rungrattanaubol[1], Anamai Na-udom[2] and Antony Harfield[1*]

[1]Department of Computer Science and IT, E-mail: jaratsrir@nu.ac.th, antonyh@nu.ac.th
[2]Department of Mathematics, E-mail: anamain@nu.ac.th
Faculty of Science, Naresuan University, Phitsanulok, Thailand

[*]Corresponding author: antonyh@nu.ac.th

Abstract

This paper introduces a computer-based model for predicting the severity of injuries in road traffic accidents. Using accident data from surveys at hospitals in Thailand, standard data mining techniques were applied to train and test a multilayer perceptron neural network. The resulting neural network specification was loaded into an interactive environment called EDEN that enables further exploration of the computer-based model. Although the model can be used for the classification of accident data in terms of injury severity (in a similar way to other data mining tools), the EDEN tool enables deeper exploration of the underlying factors that might affect injury severity in road traffic accidents. The aim of this paper is to describe the development of the computer-based model and to demonstrate the potential of EDEN as an interactive tool for knowledge discovery.

Key Words: Disability, Artificial Neural Network, Interactive Environments, Data Mining.

1. Introduction

Nowadays medical institutions are looking to find ways to reduce fatalities, disabilities and injuries caused by road traffic accidents, including the recommendation of precautionary measures and laws to prevent accidents. However, road traffic injury is still common, and it affects public health, quality of life and social quality, especially when the impact is death or disability. Many institutions have accumulatively recorded details of incidence of death or disability in order to find the causes of road traffic accidents, and the amount of collected data is very large. Many researchers have applied both statistics methods and data mining techniques to investigate the behaviour, characteristics and risks of road traffic accidents and their effect. Alcohol consumption is proposed to be one significant factor of accidents and fatality [1][2] and not-wearing seat belts and helmets is also a factor leading to severe injury [3].

In Thailand road traffic accidents are the second highest cause of death (the highest cause of death is cancer). In 1998 there were up to 7,986 deaths and in 2002 there were 13,438 deaths, increasing almost 70% within 4 years [4]. In this paper we are concerned with disabilities resulting from road traffic accidents, since disability can lead to the loss of personnel, resources, and especially long-term medical expenses. The incidence of disability and impact from road traffic injury has been surveyed and analysed by the Sirindhorn National Medical Rehabilitation Centre (SNMRC), Thailand in 2006 [5]. In this dataset, there are two levels of disability, 'disability inclusion criteria' and 'disability non-inclusion criteria', as stated by the Rehabilitation for Disabled Person Act B.E. 2534 (1991) [6]. The 'disability inclusion criteria' level is considered more severe than the 'disability non-inclusion criteria' level. This paper focuses on using the dataset from the survey to construct a computer-based model that can predict the level of disability and that can be used to explore the factors that lead to disability following road traffic accidents.

2. Statistical data analysis

The dataset used in this research is a secondary injury surveillance data collected from 26 December 2005 to 25 June 2006 in 8 hospitals chosen from 28 sentinel sites (different provinces in various regions) around Thailand. The data is split into 2 groups, which are non-severely injured (i.e. not admitted) and severely injured (i.e. admitted). The surveillance period for non-severe cases is 3 weeks after patients left the hospital and for severe

cases is 6 months. The total number of non-severely injured cases is 14,698 and the number of severely injured cases is 9,737. About 4.6% of the severely injured cases are diagnosed as disability. In order to predict the level of disability caused from road traffic injury, the cases that caused disability are selected for the statistical data analysis step.

The dataset consists of more than 150 variables, e.g. date and time of accident, sex, occupation, medical and alcohol taken, medical diagnosis, etc. In order to simplify the classification of the records, only statistically relevant variables are included in the model. The spearman rank correlation coefficient (r) was used as the criterion for selecting the variables for the model. The input variables that are related to the output variables (level of disability) that are statistically significant at 0.05 level of significance are presented in Table 1. Only these eight variables are considered in the development of the model.

Table 1: The selected significant variables

Variable Group	Variable name	r	P-value
Personal/ behavioural characteristics	Age (age)	-0.272	<0.001
	Alcohol usage (alcohol)	0.112	0.021
	Drug usage (drug)	0.144	0.003
	Belt/Helmet (belt)	0.127	0.010
	Position of patient (position)	0.122	0.012
	Cause of injury (cause)	0.170	<0.001
Environmental factors	Time of accident (atime)	0.464	<0.001
First aid	Ventricular assist device usage (airway)	0.291	<0.001

The input variables cover the personal and behaviour characteristics of patients as well as the environmental factors and first aid involved in the road accident.

Age is categorized into 8 categories (0-4, 5-14, 15-29, 30-44, 45-59, 60-69, 70-79, over 80).

Alcohol usage is 'yes', 'no' and 'don't know'.

Drug usage is 'yes', 'no' and 'don't know'.

Belt/Helmet is 'yes', 'no' and 'don't know'.

Position of patient is 'pedestrian', 'driver', 'passenger' and 'don't know'.

Cause of injury is 'motorcycle fall or vehicle overturned', 'crash or collision' and 'others'.

Time of accident is categorized into 8 intervals (6.01-9.00, 9.01-12.00, 12.01-15.00, 15.01-18.00, 18.01-21.00, 21.01-0.00, 0.01-3.00, 3.01-6.00).

Ventricular assist device usage is 'used well', 'used but not well', 'not used' and 'don't know'

There is one output variable which is Level of disability.

Level of disability is 'disability inclusion criteria' and 'disability non-inclusion criteria'.

A disabled patient is defined as 'disability inclusion criteria' if he/she is a person with physical, intellectual or physical abnormality or impairment as categorized and prescribed in the Ministerial Regulation as stated in Rehabilitation for Disabled Person Act, B.E. 2534 (1991) [6]. The other type of disability is 'disability non-inclusion criteria' which is considered less severe (e.g. a person with unilateral blindness).

These eight input variables and one output variable are used to construct a neural network as described in the next section.

3. Using neural networks for data mining

To construct a neural network for predicting the severity of injuries in road traffic accidents, the following steps were performed:

1. Data preparation and cleaning
2. Training the neural network using Weka
3. Presenting the neural network using Eden

The original data set was prepared in SPSS. There were a total of 423 records of accidents resulting in disability. Some records had missing data and they were removed in order to simplify the training process. Therefore the final dataset contained only 387 records for training and testing the neural network.

Two continuous input variables are categorised: *Age* and *Time of accident* were each divided into 8 distinct groups as mentioned in section 2.

The output variable *Level of disability* has 2 discrete values, either 0 or 1. The value 0 means 'disability inclusion criteria' and the value 1 means 'disability non-inclusion criteria'. Of the 387 records, 273 are categorised 'disability inclusion criteria' and 114 are categorized as 'disability non-inclusion criteria'.

After cleaning the data, it was loaded into the Weka data mining tool [7]. Weka was configured to perform the training of a multilayer perceptron (MLP) (for background on multilayer perceptrons and neural networks see [14]). There were eight inputs (as chosen in section 2), and one output that represents the severity of the injury. Using Weka's suggested formula, the multilayer perceptron contained one hidden layer with five nodes. All of the nodes in the multilayer perceptron use a standard sigmoid function ($f(x) = (1+e^{-x})^{-1}$). The multilayer perceptron was trained using back-propagation, with a learning rate of 0.1, a momentum of 0.1, and a training time of 1000. The resulting model was tested with two different methods. Initially a cross-validation method was used to determine the accuracy of the MLP.

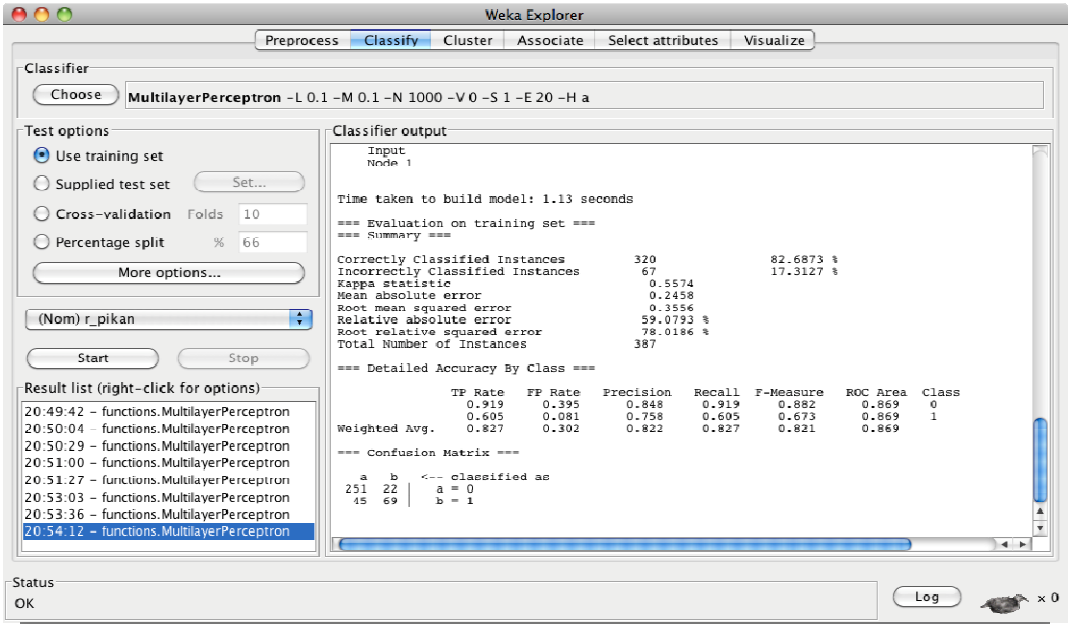


Figure 1: Training and testing the multilayer perceptron using Weka

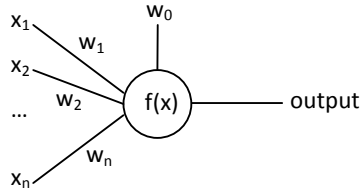
This gave an accuracy value for comparison purposes. Following this, the neural network was re-trained using the full training set (i.e. all the data) to give the most accurate model for prediction purposes and further exploration. Figure 1 shows the configuration and output from Weka, for the training of the MLP and testing using the full training set.

The output from Weka gives the results of the training and the specification of the resulting neural network. The final step was to load the specification into EDEN (Engine for DEfinitive Notations) [8] for further exploration. EDEN is a general purpose interactive environment for creating and engaging with models or artefacts. An EDEN model consists of observables, dependencies and agency [9]. The observables in a model represent the current state, and often correspond to features that have meaning—such as the value of an input to the neural network corresponding to the age of an individual. The dependencies in a model represent the linkages between observables—for example, there is a dependency between the output of one perceptron and the input of another. The agency in a model corresponds to the external interference on state, usually by the user when they change, for example, the input values to the neural network or the configuration of the neural network itself. Models in EDEN therefore offer a highly flexible and open-ended environment for exploration.

A model was constructed in EDEN of the neural network trained using Weka. The constructed multilayer perceptron consists of a set of definitions, capturing the dependency in the model. As an example, the definition for one of the nodes in the hidden layer is:

```
node2 is sigmoid(-24.9 +
                 27.7*atime +
                 -17.8*age +
                 -25.5*position +
                 4.9*cause +
                 45.2*alcohol +
                 4.3*drug +
                 3.7*belt +
                 11.5*airway);
```

The above EDEN definition has a close association with the mathematical description of perceptron, as shown in Figure 2. The variables (atime, age, position, etc) correspond to the inputs (x_1, x_2, x_3, etc) of the perceptron, and the values (27.7, -17.8, -25.5, etc) correspond to the



Output = $f(w_0 + w_1x_1 + w_2x_2 + \ldots + w_nx_n)$

Figure 2: Diagram of a perceptron

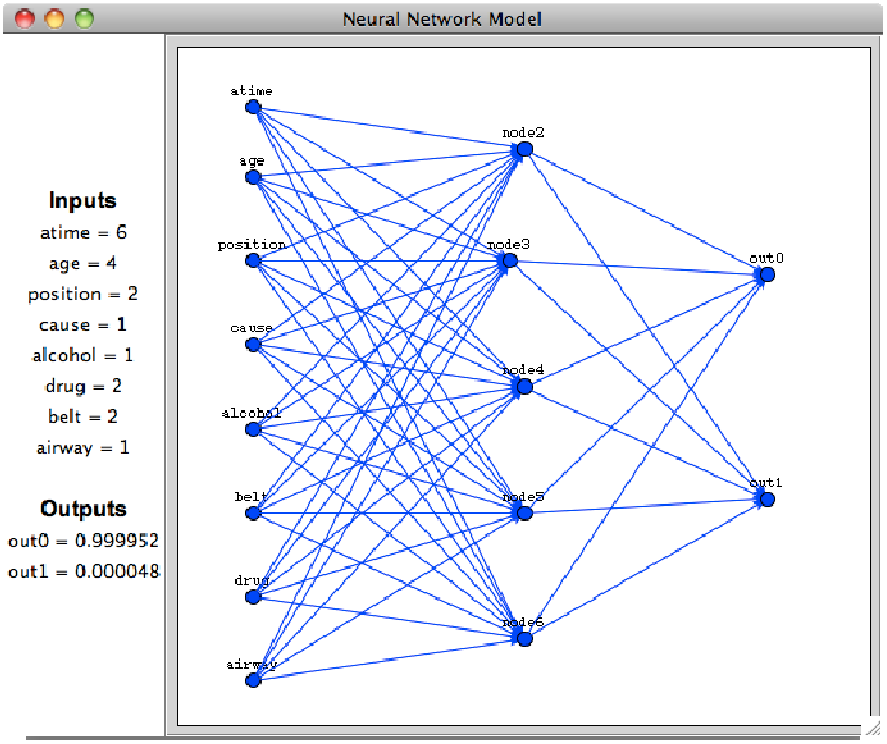


Figure 3: The neural network model in EDEN

weights (w_1, w_2, w_3, etc) of the perceptron. There are similar definitions for all the components of the MLP, and the input for one perceptron is defined in terms of the output of perceptrons from a lower layer. In this respect, an EDEN model consists purely of a set of definitions, each of which corresponds closely to components of the MLP. In contrast, a procedural program representing an MLP consists of a sequence of steps that (when performed in a particular order) calculate the MLP output. However, the individual steps on their own do not have a direct correspondence to the components of the MLP. One of the key benefits of the EDEN environment is that constructing models using observables, dependency and agency allows a semantic correspondence between the software model and the real-world referent (in contrast to traditional software programs [11]). This close, meaningful correspondence is irrelevant when the goal is speed and efficiency of computation (e.g. for data mining large datasets), but it is a desirable quality when concerned with experimentation and knowledge discovery, as will be shown in the next section. A complete discussion of the benefits of model-building in EDEN over procedural programming is beyond the scope of this paper (further evidence is presented in [11]).

On top of the basic model of the MLP (as trained using Weka), a user interface model was constructed to visual the neural network. A screenshot of the user interface (as modelled in EDEN) is shown in Figure 3. The user interface, like the model of the MLP, is a set of definitions. For example, each input/output variable on the left side of the screen is dependent on the underlying MLP. Any redefinition of the MLP could result in changes to the user interface. The flexibility for 'redefinition' is relevant to knowledge discovery because the effect of changes to the configuration of the neural network is immediately reflected in the user interface. This enables the user to explore a wide range of 'what-if' scenarios. For example, if the current configuration of inputs result in 'disability non-inclusion criteria' then a user can experiment by changing the value of an input in order to find out what variables cause the output to change to 'disability inclusion criteria' (more severe disability). Another user might be more interested in understanding the composition of the neural network and could change the connections between nodes in the MLP. As demonstrated in the next section, redefinitions of the interface allow insights into the model that point to the factors behind disability severity in road traffic accidents.

In previous work, EDEN was used for representing and exploring different types of neural network, including self-organising networks [12]. A notation for describing neural networks was developed using the principles of observables, dependency and agency. Such tools support knowledge discovery and learning in the area of neural networks. A closely related application area for EDEN is in decision support systems, as discussed by Beynon et al [13].

4. Results and analysis

There are two parts to the results: firstly we informally assess the accuracy of the model for predicting the level of disability from road traffic accidents; secondly we discuss the use of the model for understanding the factors that lead to disability from road traffic accidents.

In order to judge the accuracy of the neural network, it is relevant to compare the achieved success rate with other data mining techniques. Weka was used to find the classification accuracy by measuring the success rate with cross-validation and full training set techniques. The results showed that the MLP produced a 73% success rate using cross-validation (with 10 folds), and 83% success rate using the full training set. In order to assess the accuracy of this model, the results were compared to three other standard data mining classification techniques. A comparison can be seen in Table 2. In this experiment, the MLP model is more accurate than both the naïve Bayes simple and the naïve Bayes multinomial models. When comparing the MLP model to a random tree model, the random tree outperforms the MLP when tested against the full training set. However, the random tree model performed poorly in the cross-validation test compared to the MLP model. Although the results appear to indicate that the MLP is a suitable data mining technique for the classification of road traffic accident data, it must be noted that the dataset was small and the comparison relatively informal.

Table 2: Classification technique comparison

Classification technique:	Naive Bayes Simple	Naive Bayes Multinomial	Random Tree	MLP
Accuracy (full training)	64%	72%	89%	83%
Accuracy (cross-validation)	-	71%	67%	73%

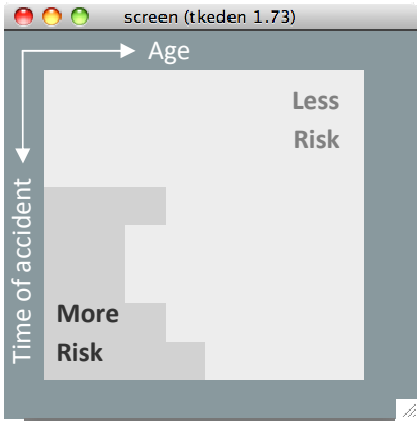


Figure 4: Map of risk according to the MLP

The second part of the results is concerned with what can be learnt about the factors that lead to disability in road traffic accidents. In this paper, we only discuss one such finding. Using the EDEN environment, the factors affecting disability severity were explored by varying the inputs to the neural network. As discussed in the previous section, the visualisation can be redefined to give alternative views on the model. The example given here involves fixing some of the inputs and then observing the outcome of varying the remaining inputs. In one such exploratory exercise, the case is considered where a *pedestrian* is the victim of a *crash* where there was no *alcohol* or *drugs* or *safety devices* involved, and there was no use of a *ventricle assistance device*. If these six variables are fixed then there are two remaining variables that can be explored: the age of the victim and the accident time of day. A new interface was defined on top of the existing model that shows the outputs of the model on an 8x8 grid (see Figure 4), representing all the possible *age* and *accident time of day* values. The constructed visualisation shown in Figure 4 is shaded in a darker colour where the output of the MLP was 0, and a lighter colour where it was 1. According to this model, the group corresponding to the darker shading in the bottom left-hand corner is at risk of more severe disability from road traffic accidents in the specific case described above. This can be interpreted as younger people later in the day are more at risk. This is an example of one result that can be derived by redefining the interface to explore the neural network model. A complete analysis of the model is the subject of a larger study, the purpose here is to demonstrate that through experimentation using the EDEN tool, there is potential to uncover some of the key factors involved in road traffic accidents.

5. Conclusion

This paper has shown that a neural network is an appropriate data mining solution for the purpose of predicting disability severity from road traffic accidents in Thailand. For the selected dataset, a MLP neural network was shown to be more accurate than other chosen data mining classification techniques. An accuracy of 73% was achieved when testing the neural network using 10-fold cross-validation. This compared favourably with other data mining classification techniques. In a similar study [10], different types of neural network were used to build models of the causes of road traffic accidents in Florida, USA. Using a MLP they were able to predict injury severity based on a similar selection and number of input variables, to an accuracy of 73%. In common with our findings, the MLP also proved to be the most accurate type of neural network out of the methods used in their study.

As a consequence of the current study, the resulting neural network can be used as an example of how to predict injury severity in terms of disability. By fixing key variables and experimenting with a subset of the variables, we can observe the factors that contribute to disability severity, as modelled by the neural network trained with the survey from Sirindhorn National Medical Rehabilitation Centre in Thailand. This paper only goes as far as highlighting the potential for understanding the factors behind accidents. Further work is needed, together with larger datasets and better trained neural networks, in order to draw clear inferences regarding the factors that contribute to road traffic accidents.

The final contribution of this paper has been to demonstrate the one potential use of the EDEN environment for constructing and exploring a computer-based model that enables knowledge discovery in the real-world domain of road traffic accidents. Previous work has demonstrated the potential for interactive environments such as EDEN to be utilised for data mining and knowledge discovery—the current study contributes another example of the need for software that encourages exploration and experimentation in the area of intelligent systems.

Acknowledgements

The authors would like to thank Naresuan University for providing research funding and support for this project.

References

[1] S. Zajac, J. Ivan, “Factors influencing injury severity of motor vehicle-crossing pedestrian crashes in rural Connecticut”, Accident Analysis and Prevention. 2003, 35(3), pp. 369-379.

[2] M. Keall, W. Frith, T. Patterson, “The influence of alcohol, age and number of passengers on the night-time risk of driver fatal injury in New Zealand” Accident Analysis and Prevention, 2004, 36(1), pp. 49-61.

[3] F. Valent, F. Schiava, C. Savonitto, T. Gallo, S. Brusaferro, F. Barbone, “Risk factors for fatal road accidents in Udine, Italy” Accident Analysis and Prevention, 2002, 34(1), pp. 71-84.

[4] Thai Health Promotion Foundation: The sustainability of well-being for Thai people. “Annual report 2003 – The Thailand Research Fund (TRF)”. Bangkok, 2004.

[5] D. Suvaphan et al, “Incidence of Disability and Impact from Road Traffic Injury, 2006”, Sirindhorn National Medical Rehabilitation Centre (SNMRC), 2006.

[6] Rehabilitation for Disabled Person Act, B.E. 2534 (1991). URL: http://thailaws.com/law/t_laws/tlaw0245.pdf

[7] Weka 3. Data Mining Software in Java. URL: http://www.cs.waikato.ac.nz/ml/weka/

[8] Empirical Modelling Research Group. EDEN Documentation. URL: http://www2.warwick.ac.uk/fac/sci/dcs/research/em/software/eden/

[9] J. Rungrattanaubol. A treatise on Modelling with definitive scripts. PhD thesis, Department of Computer Science, University of Warwick, April 2002.

[10] M.A. Abdel-Aty1, H.T. Abdelwahab. “Predicting Injury Severity Levels in Traffic Crashes: A Modeling Comparison” Journal of Transportation Engineering, Vol. 130, No. 2, March 1, 2004, 204-2010.

[11] W.M. Beynon, R.C. Boyatt, S.B.Russ. “Rethinking Programming”. In Proceedings IEEE Third International Conference on Information Technology: New Generations (ITNG 2006), April 10-12, 2006, Las Vegas, Nevada, USA 2006, 149-154.

[12] Anonymous. “Neural Networks and Notations”. In The Second Warwick Empirical Modelling Bulletin (WEB-EM-2). URL: http://www2.warwick.ac.uk/fac/sci/dcs/research/em/publications/web-em/02/neural.pdf

[13] W.M. Beynon, S. Rasmequan, S. Russ. “A New Paradigm for Computer-Based Decision Support”. Decision Support Systems, Vol 33, 2002, 127-142.

[14] S. Haykin. Neural Networks: a Comprehensive Foundation, Macmillan, 1994.