7th INTERNATIONAL CONFERENCE
''TOWARDS A HUMANE CITY''
Environmentally Friendly Mobility
Novi Sad 6th and 7th December 2019

Sandra Nemet, MSc[1]
E-mail: sandra.nemet@rt-rk.com
Dragan Kukolj, PhD[2]
E-mail: dragan.kukolj@rt-rk.uns.ac.rs
Dragan Jovanović, PhD[2]
E-mail: draganj@uns.ac.rs
Gordana Ostojić, PhD[2]
E-mail: goca@uns.ac.rs
Stevan Stankovski, PhD[2]
E-mail: stevan@uns.ac.rs
[1] RT-RK Institute for Computer Based Systems
Narodnog fronta 23A, 21000 Novi Sad, Serbia
[2] Faculty of Technical Sciences, University of Novi Sad
Trg D. Obradovića 6, 21000 Novi Sad, Serbia

IDENTIFICATION OF INFLUENCING ATTRIBUTES FOR THE TRAFFIC ACCIDENT SEVERITY PREDICTION MODEL

Key Words

Traffic accident prediction; Feature selection; Swarm optimization; Neural network

Abstract

This paper deals with identification of the most influencing input attributes related to the accuracy of the prediction model. It is assumed that the prediction model may be represented by any machine learning-based models, including artificial neural networks, fuzzy models, etc. Selection of influencing attributes is based on particle swarm optimization (PSO) combined with neural networks. The role of neural networks is to estimate fitness of each particle during the search procedure implemented using a PSO algorithm. The presented feature selection method represents the first step in the prediction model design which is applied on the accident severity prediction. The method is applied on a dataset characterized with weak correlation between the model's inputs and output representing an accident severity. The proposed approach has shown the ability to identify subsets of factors that have an influence on accident severity prediction. Selection of appropriate inputs improves the prediction model accuracy.

1. INTRODUCTION

Growth in road traffic intensity as consequence has a rapid increase in road accidents. Identification of the most influencing factors related to traffic accidents represents a basic phase in analysis of accident causes and their consequences.

Artificial neural networks (ANN) are the most frequent machine learning techniques employed in traffic accident severity prediction. Back-propagation neural networks were applied in [1, 2]. Ogrenci [3] predicted accident severity with Multilayer Perceptron (MLP) neural networks. Moreover, Support Vector Machines (SVM) were used for crash severity prediction [4], and traffic fatalities prediction was improved with an SVM model parameters optimization using hybrid particle swarm optimization [5]. In order to design more accurately, a robust and interpretable prediction model with reduction of input data by proper selection of key variables, should be performed [6].

Feature selection is a process of taking a subset of features from the initial set without transformation. Generally, feature selection methods are classified in three groups of approaches: wrappers, filter approaches and their combination. The wrapper approach uses a machine learning algorithm, repeatedly called by a data re-sampling technique, while the filter approach looks for features which have a stronger measure of association to the output variable.

This work describes the problem of finding fetures that are most influential as a problem of searching feature space. As a search technique, the Particle Swarm Optimization (PSO) algorithm was used [7]. The PSO algorithm uses a particle swarm made out of a vast number of particles in a candidate population that represent feature subsets. In each PSO iteration the particles „learn“ from one another and refresh their knowledge in relation to the current and best solution, as well as the best solution for the whole swarm. Every particle in the swarm is presented with its position and speed. Fitness of every particle in this PSO algorithm implementation presents an estimation of the crash severity estimated using an MLP type of ANN.

In the second section, the attributes selection method is described using a hybrid PSO–ANN approach. The third section contains results of the attributes selection method applied on the traffic accident dataset. In the last section, concluding remarks are given.

2. FEATURE SELECTION USING PSO+ANN ALGORITHM

The feature selection algorithm presented here is based on a Particle Swarm Optimization (PSO) method, which is combined with an artificial neural network (hereinafter referred to as the PSO+ANN algorithm). Input of the ANN is a candidate subset of features, and it is trained to predict crash severity. The calculated prediction error is presented as fitness of the particle and it has an influence on the current position of the particle (feature subset), the direction of the particle, and preferably also on the entire swarm. Therefore, the used hybrid method of feature selection using the PSO+ANN algorithm is characterized with the following:

- Number of particles, i.e. population size, defines how many neural networks will be used during the search for the best feature combination.
- Every neural network has a different combination of input variables in one generation, but their number is constant.
- During the search of the solution space, synaptic weight coefficients of all neural networks are iteratively adapted with an aim to achieve a minimal error.
- The search method using the hybrid PSO+ANN algorithm has ended when the given number of generations is achieved.

Particle swarm optimization algorithm was first described by J. Kennedy and R. C. Eberhart in 1995 [7]. The main idea of the method resulted from an association with social behavior of an individual from a flock of birds or fishes. PSO is a stochastic search algorithm based on a population of adaptive solutions by simulating the behavior of individuals in the swarm. The PSO approach has become popular because of the simplicity of the basic PSO algorithm implementation [8] and because it is characterized by low computational complexity [9].

The position of a particle denoted by vector $\boldsymbol{x}_i$ (in iteration t) is changed by adding the speed vector $\boldsymbol{v}_i$ in the current position:

$$\boldsymbol{x}_i(t) = \boldsymbol{x}_i(t-1) + \boldsymbol{v}_i(t). \tag{1}$$

Speed vector $\boldsymbol{v}_i$ defines where will the i-th particle move in the next iteration, considering a previous best positions, as well as the „experience" of the most successful particle from the swarm:

$$\boldsymbol{v}_i(t) = W\boldsymbol{v}_i(t-1) + C_1 r_1 (x_{pbest_i} - \boldsymbol{x}_i(t)) + C_2 r_2 (x_{gbest} - \boldsymbol{x}_i(t)), \tag{2}$$

where r_1 and r_2 are random values in the interval [0,1], W is the inertion coefficient, while C_1 and C_2 represent constants for controlling the influence of the best local solution for a particle $\boldsymbol{x}_{pbest}$ and global best solution for particle $\boldsymbol{x}_{gbest}$, respectively. In each iteration of the PSO algorithm, particles "learn" from each other and find new solutions on a better location in the solution space. The local best position (*pbest*) and the global best position of all particles (*gbest*) have an influence on every particle in the PSO set of solutions, besides the current moving direction of every particle.

Feature selection is ranking features based on importance using the PSO+ANN method defined by a 100 particles (individual solutions) and 50 iterations. The complete algorithm is repeated 20 times, in order to reduce the stochastic nature of the search method. Inertia coefficient W =0.754 was adopted, while a value of C_1=C_2=1.5457 was adopted for the personal (C_1) and global (C_2) learning coefficient. Values of the best particles, i.e. particles with the best combination of input variables are used in forming the importance index of the following form:

$$\sum_i^m \sum_j^n \left(1 - \frac{c_{ij} - c_{min}}{c_{max} - c_{min}}\right), \tag{3}$$

with the meaning: m – number of repetitions; n – number of variables; c_{ij} – value of the solution for the i-th repetition in which the j-th variable participated; c_{max} – the weakest achieved result; and c_{min} – the best achieved result.

3. RESULTS OF FEATURE SELECTION OBTAINED WITH THE PSO+ANN

The accident dataset used for this study contained traffic accident records from year 2008 to 2011 in the urban area of the city of Novi Sad [10]. The dataset comprised of the official statistics on traffic accidents, with a total of 42 features and 752 traffic accidents, from which 798 were pedestrian casualties. All factors involved in accidents are classified and labeled, including consequences which consist of four classes: fatalities, major injuries, minor injuries and without injuries.

A neural network with one hidden layer and fitness function in the form of the mean absolute error between the real and estimated output was embedded in the search method using the PSO algorithm. In case of five input variables, neural network has eleven neurons on the hidden layer and one output neuron with the meaning of variable *Consequences*.

Different cases for a variable set of input variables were considered. When considering the choice of the best five features, the PSO+ANN algorithm has selected variables shown on Figure 1. The initial set with 25 input variables corresponds to the variable list given on the ordinate in Figure 1. Bars shown above the number of the corresponding variable represent a variable's importance index (3). Based on the importance index, a subset of variables [9, 11, 20, 24, 25] can be selected, i.e. [*Gender, Behavior, Age group, Pedestrian crossing, Road crossing*]. Interestingly, a combination of features [11, 17, 20, 24, 25] was achieved using the PSO+ANN algorithm, i.e. [*Behavior, Gender group, Age group, Pedestrian crossing, Road crossing*], with a minimal fitness value of 0.026.

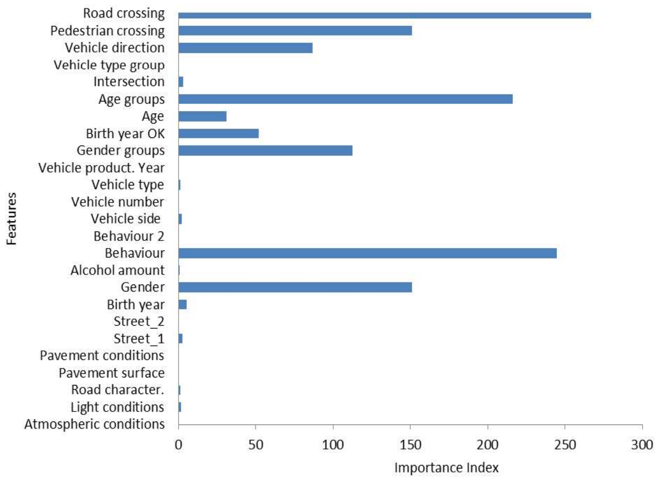

Figure 1. Selection of five features using the PSO+ANN algorithm

As an illustration of PSO+ANN algorithm convergence, Figure 2 shows 50 iterations of a randomly selected single run of PSO execution. The algorithm converged towards the best combination of features [9, 11, 20, 24, 25] in a first few iterations, i.e. [*Gender, Behavior, Age group, Pedestrian crossing, Road crossing*], where after the seventh iteration it still minimized the fitness, but with the same feature combination. The achieved solution represents one of the best solutions.

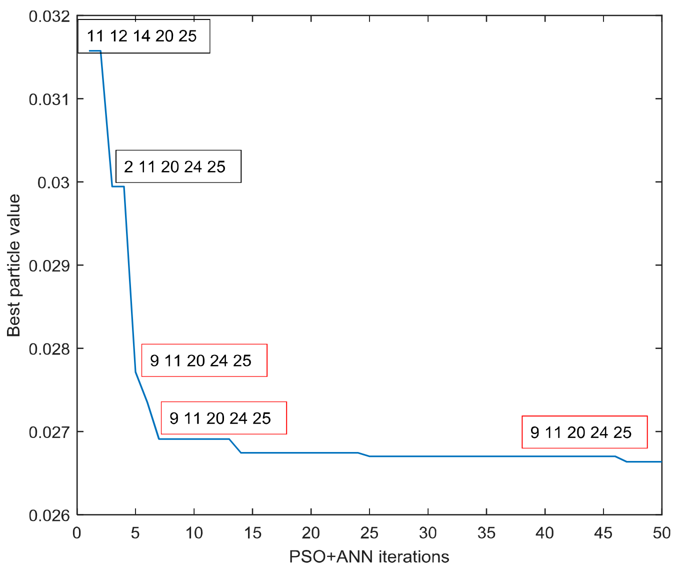

Figure 2. Convergence of the PSO+ANN algorithm during selection of five features

4. CONCLUSION

The feature selection method in the form of a specific realization of the hybrid algorithm for optimizing the swarm of solutions with artificial neural networks (PSO+ANN) has shown that it enables choosing independent input variables necessary for generating the prediction model. Specificity of the PSO+ANN algorithm realization is defined by: (a) the method is composed of an environment which is based on the PSO search rules and generates a large number of variable subsets; (b) it uses neural networks for fitness functions estimation; and (c) the entire method is repeatedly executed in order to reduce the stohastic characteristics in the PSO+ANN algorithm and easily identify the most frequently selected variables. The analysis of the feature selection PSO+ANN algorithm's results has shown that the most suitable subset consists of the following five variables: *Gender*, *Behavior*, *Age group*, *Pedestrian crossing*, and *Road crossing*.

ACKNOWLEDGMENT

This research is partially supported by the project "Development of a system for detection and pre-diagnosis of participants' injuries in road traffic accidents" by a grant 142-451-2545/2019-02 from the Secretary of Science of AP Vojvodina and by a grant TR32034 from the Ministry of Education, Science and Technology Development of the Republic of Serbia.

REFERENCES

[1] X. Meng, Q. Hou and Y. Shi, "Research on Accident Prediction Models for Freeways in Mountainous and Rolling Areas," in *7th Int. Conf. on Measuring Techn. & Mechatronics Automation*, 2015, pp. 838-838.

[2] S. Li, and D. Zhao, „Prediction of Road Traffic Accidents Loss Using Improved Wavelet Neural Network," in *Proceedings of IEEE TENCON'02*, 2002, pp. 1526-1529.

[3] Ogrenci, A.S., "Complex Event Post Processing for Traffic Accidents," in *13th IEEE International Symposium on Computational Intelligence and Informatics - CINTI*, 20–22 Nov., Budapest, 2012, pp. 341-345.

[4] A. Iranitalaba and A. Khattak, "Comparison of four statistical and machine learning methods for crash severity prediction," *Accident Analysis & Prevention*, vol. 108, pp. 27-36, 2017, doi:10.1016/j.aap.2017.08.008.

[5] X. Gu, T. Li, Y. Wang, L. Zhang, Y. Wang and J. Yao, "Traffic fatalities prediction using support vector machine with hybrid particle swarm optimization," Journal of Algorithms & Computational Technology, vol. 12(1), pp. 20-29, 2017.

[6] S. Nemet, D. Kukolj, G. Ostojić, S. Stankovski, D. Jovanović, "Aggregation Framework for TSK Fuzzy and Association Rules: Interpretability Improvement on a Traffic Accidents Case," *Applied Intelligence*, 2019, DOI: https://doi.org/10.1007/s10489-019-01485-6.

[7] J. Kennedy and R. C. Eberhart, "Particle swarm optimization," *Proc. IEEE Conf. on Neural Networks*, Piscataway, New Jersey, 1995, pp. 1942–1948.

[8] M. Reyes-Sierra and C. A. Coello, "Multi-objective particle swarm optimizers: A survey of the state-of-the-art," *International Journal of Computational Intelligence Research (IJCIR)*, vol. 2, pp.287–308, 2006.

[9] J. Kennedy and R. C. Eberhart, *Swarm Intelligence*, Morgan Kaufmann Publishers, San Francisco, California, 2001.

[10] A. Bulajić, D. Jovanović, B. Matović, S.D. Bačkalić, "Identification of High-Density Locations with Homogenous Attributes of Pedestrian Accident in the Urban Area of Novi Sad," *Proc. Road Accidents Preventions,* 2014, pp. 89-98.