ISSN 1561-5359. Штучний інтелект, 2016, № 3

УДК 519.7: 621.8

R.M. Trokhymchuk
Kyiv National Taras Shevchenko University, Ukraine
pr. Academician Glushkov, 4d, Kyiv, 03680

RESULTS OF APPLICATION OF MODULAR ARTIFICIAL NEURAL NETWORKS FOR INTELLIGENT DATA ANALYSIS (DATA MINING) AND FORECASTING PROCESSES IN THE FIELD OF ECOLOGY AND ENVIRONMENT PROTECTION

Р.М. Трохимчук
Київський національний університет імені Тараса Шевченка, Україна
пр. Академіка Глушкова, 4д, Київ, 03680

РЕЗУЛЬТАТИ ЗАСТОСУВАННЯ МОДУЛЬНИХ ШТУЧНИХ НЕЙРОННИХ МЕРЕЖ ДЛЯ АНАЛІЗУ ІНТЕЛЕКТУАЛЬНИХ ДАНИХ ТА ПРОГНОЗУВАННЯ ПРОЦЕСІВ У СФЕРІ ЗАХИСТУ НАВКОЛИШНЬОГО СЕРЕДОВИЩА

The aim of this work is the use of modular artificial neural networks (ANN) for data mining (Data Mining) and forecasting of various processes in the field of ecology and environmental protection, as well as the comparison of the results of the proposed model with the results of other data analysis methods (the methods of mathematical modeling and mathematical statistics).
Keywords: municipal solid waste (MSW), Data Mining, artificial neural networks (ANN), modular ANN, forecasting processes.
Метою даної роботи є використання модульних штучних нейронних мереж для виведення даних та прогнозування різних процесів у галузі екології та охорони навколишнього середовища, а також порівняння результатів запропонованої моделі з результатами інших методів аналізу даних (методи математичного моделювання та математичної статистики).
Ключові слова: муніципальні тверді відходи, штучні нейронні мережі, процеси прогнозування.

Introduction. At present, the constant growth in population and the natural growth of the consumption of renewable and non-renewable resources is accompanied by a steady growth of the volume of waste, in particular, the so-called municipal solid waste (MSW). Therefore, the tasks of analysis and forecasting the volume and composition of solid waste in separate territories, regardless of the methods of their processing (burial in landfills, composting, recycling, thermal treatment, and others) are extremely relevant. Such forecasting is necessary to make informed and effective planning of transportation and all types of solid waste recycling. For efficient management and long-term planning in this area it is necessary to make predictions for 10-15 years, which would allow to estimate the volume and composition of MSW in the future, plan the location, construction and structure of enterprises for waste recycling and improve the overall ecological and economic level of the region. Unfortunately, the experience of long-term forecasting of the composition and volume of formation of MSW is practically absent in domestic practice.

There are balance, factor and statistical models used to analyze the process of formation and forecasting the volume and composition of MSW [1].

In the balance model the formation and the forecasting of composition of the waste are estimated on the basis of information on production, sales, consumption of products which generate specific waste streams in the analyzed area. Factor models are based on an analysis of the factors (parameters), which directly affect the processes of waste. Various socio-economic and demographic characteristics of a given region may be examples of these parameters. Namely, the size and composition of the population in the analyzed territory, GDP per capita in the country, the annual or monthly income of the individual family, the minimum hourly wage, and so on. Seasonal variations of parameter values and a big number of them greatly complicate the construction and use of adequate formal methods for this model.

Statistical models identify statistical regularities of changes in the composition and volume of solid waste formation. In some studies, researchers complement factor model by mathematical and statistical methods that can significantly improve the accuracy and quality of the forecast [2,3].

Other models that take into account a wide range of quantitative and qualitative parameters and performing the analysis of these processes, taking into account environmental, economic and social aspects, are also offered today in order to solve this problem.

Today, other models are also available to solve this problem. These models use a set of quantitative and qualitative parameters and perform data analysis processes, taking into account environmental, economic and social aspects.

At the same time, the following requirements apply to systems analysis and forecasting of solid waste composition.

1. Generalization level should correspond to the level of the forecast. The parameters used in the model must take into account the peculiarities of the region. Balance models using only averaged data are insufficient to explain the regional dynamics. In this case, preference should be given to factor models that use socio-economic and demographic characteristics of the region.

2. Predictability of parameters. The parameters that can be predicted with reasonable accuracy over a long period must be selected for the forecasting.

3. Ease of use. The technique should provide output that can be easy to obtain and easy to interpret.

The classical methods of mathematical statistics and systems analysis, expert systems, fuzzy models, etc. certainly meet the specified requirements.

In addition, to solve this kind of problems in recent years often is used so-called intellectual data analysis, which corresponds to the actively used term – Data Mining. The term Data Mining is interpreted as extraction of data, in-depth analysis data, digging (receiving, finding) of knowledge in databases. Interdisciplinary field of Data Mining also uses the methods of mathematical statistics. Furthermore, it involves for the study of the larger or smaller amounts of data significantly different methods, in particular methods of pattern recognition, artificial intelligence algorithms, artificial neural networks, genetic algorithms, methods of evolutionary programming, an associative memory, a fuzzy logic theory databases etc. [4].

Data Mining can be characterized as a technology that is designed to find non-obvious, objective and useful in the practice laws when dealing with the large data sets. We say non-obvious, as extracted patterns are often not detected by conventional methods of information processing and by the expert way._We say objective, as identified patterns correspond to reality, as opposed to expert opinion, which is always subjective. We say practically useful, because the conclusions allow to predict the course of the analyzed processes, which always has a particular practical application.

The model used. In this paper, the parameters used in the factorial model, are analyzed and predicted by means of a new generation of artificial neural networks (ANN), namely – a modular ANN [5,6]. Modular ANN is a logical continuation of the ideas of classical ANN with specific architecture, the main feature of which is the availability of tools and techniques for the construction of stages or systems composed of individual neural networks. Modular ANNs are promising model of ANN, as they provide an opportunity to combine at the stage of learning a variety of traditional architecture and learning algorithms of classical neural networks to optimize, improve efficiency and adapt the ANN model for this situation.

There are two approaches for the definition of a separate module in the modular ANN. In both cases the module is considered to be a specific group of neurons in a network. But in the first approach, all modules of the modular ANN have the same architecture and in the second – each of the network modules may have the original architecture.

It is this second approach that makes it possible to note the main advantage of modular ANN compared with classical neural network of a particular type. We know that the success and effectiveness of ANN in solving some problems greatly depends on an adequate selection of the type or network architecture. Therefore, breaking a complex task into subtasks, and using the appropriate modules of different architecture solutions for individual subtasks, makes it possible to increase significantly the quality of the solution of the whole problem.

In this work the second option of constructing a modular ANN was used. Corresponding modules had the following architecture: Hopfield network, multilayer perceptron of Rosenblatt and neocognitron.

The central problem in the construction of any ANN is a training procedure. In accordance with the definitions of the module and the modular ANN, modules are divided into two categories:

1) pre-trained, or deterministic modules. Such a module is built into the structure as a pre-trained neural network and during training of the whole network it is not changed. This module has a non-standard number of I/O and is identical to the conventional neuron;

2) untrained or non-deterministic modules. In this case, only the structure and type of input/output (the number of inputs and outputs, their types, etc) are predefined for the network to be used as a module. Training of this module takes place in the process of training of the external neural network.

Training of the neural network, which comprises only deterministic modules, practically does not differ from the training of classical neural networks, while training procedure for individual modules in the case of non-deterministic modules is a creative process and rests solely with the developer of modular ANN. Here it is worth noting that the problem of choosing the types of modules, a way to combine them into a single network, training algorithms for such modular ANN today are still poorly understood. These problems certainly are of great interest for a separate study.

Formulation of the problem. In this paper we investigate the problem of analyzing and forecasting the volume and composition of MSW for certain regions. In particular, we study the dependence of the volume and composition of MSW on various socio-economic indicators in the region. This problem has been divided into two sub-tasks:

1) Prediction of volume and composition of solid waste for a certain period of time.

2) Analysis and forecasting of indicators measuring the efficiency of solid waste recycling.

The main indicators were chosen as follows: 1) the composition of solid waste per capita; 2) the area of the territory in landfills, which is occupied by each category of solid waste; 3) calorie (energy output) of each of the categories of solid waste; 4) the results of the recycling: energy output and materials produced after processing.

Review of methods and studies. Analysis of seasonal fluctuations in the composition of MSW is one of the main aspects in its research and forecasting. Unfortunately, there are currently no sufficiently accurate data and methods needed to solve this problem, as one of the main conditions for this is the ability to obtain data on the exact content of the composition of solid waste, and therefore – separate collection of MSW. In countries where these studies were conducted (Ukraine, Russia, Georgia, Lithuania), the percentage of separate collection of MSW is an average of 30-40% depending on the category. Therefore, only that portion of the MSW was analyzed, which was made for the separate collection.

Seasonal fluctuations in the composition of MSW for a long time were completely ignored, which led to significant errors in the forecasts. Despite the fact that in recent years they have begun to pay attention to them, at the moment there are no adequate means to predict these fluctuations, or any model of relationships between them and their causes (weather conditions, seasonal changes in the size of the population, changes in consumption patterns, etc).

One of the most popular methods for solving this problem is the use of methods of regression analysis and time series analysis. These methods allow you to check availability of related data, to determine the degree of this dependence, and by approximation to build a simplified model of the monitoring process. An example of such a study are the results presented in [3]. Here, analysis and forecasting were carried out in two ways: by analyzing the relevant time series, based on the collected statistics on the composition of MSW, and by regression analysis based on the volume of solid waste, that is depending on the socio-economic indicators in the region. The result confirmed a high enough depends of the volume of MSW on the size of population and GDP per capita in the monitored region, and predicted values obtained for a certain period practically coincided with the real ones. However, this approach proved to be ineffective when trying to predict the differentiation of waste by categories, and the data only on the total volume of MSW do not allow to justifiably plan construction of processing enterprises in the territory and expect benefits from their activities, both for the economy and for the environment.

A number of papers predicting the composition of MSW was carried out with the help of ANN. The paper [7] contains a survey of these methods. However, most often quite simple ANN model of small size were used. As a result, projections obtained did not differ from the results of the approximation and further input data extrapolation by classical methods of mathematical statistics or computational algorithms. Obtained by these methods projected (extrapolated) indicators beyond a given training sample grow quickly and indefinitely, and do not reflect the presence of seasonal fluctuations, indicating a poor-quality forecast.

In this work, data analysis was performed using three ANN: the multilayer perceptron with three hidden layers, the cognitron with one hidden simple layer and one hidden composite layer, and finally, the modular neural network with three aforementioned modules.

Input data, considered as training sample, were divided in several ways in a constructive (training) and a control part for evaluating the quality of a proven model. The most high-quality results for each of the methods of division of the training sample showed a modular ANN. It allowed to reflect the forecast seasonal cyclical fluctuations of the analyzed parameters [8]. In addition, analysis of the results allowed to formulate recommendations for environmental specialists: what data, how often and how much should be collected for the high-quality forecasting.

Various other traditional methods of data analysis (regression and correlation analysis, interpolation, mathematical (analytical) modeling) were tested for comparison and evaluation of the results. The comparative analysis has shown that the classic methods are ineffective in this case. Firstly, due to small amounts of available data sets for each analyzed period, when there is a large number of different indicators. This fact makes it very difficult to determine interdependencies and data sorting indicators by the degree of their influence. Secondly, the proposed training sample for the analysis is not complete or sufficiently representative.

Data Analysis (Data Mining). In the first stage of the study, the regression and correlation analysis of the original data in order to verify the representativeness of the sample and the presence of interdependencies between individual indicators, was carried out. It was determined that the initial sample does not contain significantly interrelated data and is quite suitable for the detection of seasonal variations and forecasting. Given the rather subjective, incomplete and chaotic nature of the data provided for the analysis by environmental experts, pre-processing of the data was carried out. Some of the input data were either non-existent at certain time intervals or were represented by only a few values for the entire period. Missing data were approximated by smooth polynomial, which is consistent with their real continuity and absence of sudden changes.

The training set contained data collected for during one year data from different regions of Eastern Europe, which have significantly different main socio-economic indicators: Georgia (Kutaisi), Lithuania (Kaunas), Russia (St. Petersburg), Ukraine (Boryspil).

The statistics have been divided into several categories, depending on the methods and results of the processing of MSW. Processing results include materials, energy and heat generated directly during the processing of MSW and after it.

In turn, MSW have been divided into separate categories in order to improve efficiency of the system operation, as the studied parameters for such categories differ from each other and accordingly, change independently for each category.

Note the possible increase in the efficiency of the analysis in case of the further division of these categories into subcategories. However, it takes much more complete sample for such analysis. In our case, such data are not available for most regions. Furthermore, it should be taken into account that such expansion will considerably increase the amount of predicted parameters and may reduce the accuracy of the forecast.

In constructing the model, the results of previous studies have also been taken into account, namely – the analysis of the interdependence of the composition and amount of waste and the socio-economic indicators in a particular region [8]. As a result, two modules for forecasting was actually allocated for each indicator: the first one for the prediction based on the values of previous periods, the second one for the prediction on the basis of the values of social and economic indicators for the current period.

Forecasting Model. Four main groups of indicators have been proposed and researched: 1) the composition of solid waste; 2) the filling of garbage bins; 3) the total calorie content; 4) the processing result.

Three main areas for the prediction were identified:

1. Based on the values of parameters for the current period to determine the values of these parameters in the following period.

2. On the basis of the known socio-economic indicators in the subsequent period of time, to determine the composition of MSW during this period.

3. Based on the composition of MSW in the next period of time to evaluate the efficiency of their processing in that time period.

Thus, for example, for the result of processing in the output we have two: first, the value obtained by prediction based on the values in previous periods, and “associated” values obtained from the composition of MSW at the period of forecasting based on socio-economic and ecological status of the region. The total value is calculated as a kind of average of these two values and the corresponding coefficients are selected in the training process.

Thus, this approach allows us to combine the main currently existing approaches to predict the composition of MSW and the results of their processing: namely, classical forecasting based on time series approach, and the forecasting that takes into account the dependence of the composition of MSW on the socio-economic and ecological state of the region. Implementation of the balance between these two approaches is performed automatically during neural network training. As a result, we get a much more accurate forecast, since both approaches are not without drawbacks, and the balance between them allows to compensate for these shortcomings.

The results and conclusions. The modular ANN used in this work, showed high efficiency and accuracy compared to traditional methods and approaches applied to solve this problem, mainly due to the use of data on the relationship and the nature of changes in the composition of MSW and the results of their processing.

The data indicate that the modular ANN is able to identify fluctuations in the initial sample with sufficient accuracy and to transfer these fluctuations to the data obtained from the prediction [8].

However, for the long-term forecasting the sample should be extended so that the presented data reflect the information on long-term fluctuations. This extension of the sample will not only get a forecast for the period of time sufficient to use the results in order to optimize the storage and processing of MSW, but also will increase the accuracy of forecasting cyclical fluctuations (eg seasonal).

References

1. Beigl P., Wassermann G., Schneider F., Salhofer S. Forecasting municipal solid waste generation in major European cities. – // IEMSs International Conference «Complexity and Integrated Resources Management». 2004.
2. Bandara NJGJ, Hettiaratchi JPA, Wirasinghe SC, Pilapiiya S. Relation of waste generation and composition to socio-economic factors: a case study. – // Environmental Monitoring and Assessment, 2007, 135(1–3): 31–39.
3. Rimaityte I., Ruzgas T., Denafas G., Racys V., Martuzevicius D. Application and evaluation of forecasting methods for municipal solid waste generation in an eastern-European city. – // Waste Management & Research, 2012, 30(1) 89–98.
4. Макленнен Дж., Танг Ч., Криват Б. Microsoft SQL Server 2008. Data Mining – интеллектуальный анализ данных. – // СПб.: БХВ-Петербург, 2009.
5. Schmidt A. A modular neural network architecture with additional generalization abilities for high dimensional input vectors. – // Manchester Metropolitan University, Department of Computing, September 1996. – 113 p.
6. Резник А.М., Куссуль М.Э., Сычов А.С., Садовая Е.Г., Калина Е.А. Система автоматизированного проектирования модульных нейронных сетей CAD MNN. – // Математичні машини і системи.– 2002, № 3.– С.28-36.
7. Chaudhari S.R., Dhawale C.A. A survey on application of ANNs in solid waste prediction. – // International Journal of Advanced Research in Computer and Communication Engineering, vol. 5, issue 10, October 2016, 110-113.
8. Trokhymchuk R.M., Kozlov K.E. Application of modular neural networks for analysis and prediction of socio-economical and ecological indicators of the region. – // Вісник КНУ імені Тараса Шевченка. Кібернетика. Вип.1(15), 2015, с.63–72.

Literatura

1. Beigl P., Wassermann G., Schneider F., Salhofer S. Forecasting municipal solid waste generation in major European cities. – // IEMSs International Conference «Complexity and Integrated Resources Management». 2004.
2. Bandara NJGJ, Hettiaratchi JPA, Wirasinghe SC, Pilapiiya S. Relation of waste generation and composition to socio-economic factors: a case study. – // Environmental Monitoring and Assessment, 2007, 135(1–3): 31–39.
3. Rimaityte I., Ruzgas T., Denafas G., Racys V., Martuzevicius D. Application and evaluation of forecasting methods for municipal solid waste generation in an eastern-European city. – // Waste Management & Research, 2012, 30(1) 89–98.
4. Maklennen Dzh., Tang Ch., Krivat B. Microsoft SQL Server 2008. Data Mining – intellektualnyiy analiz dannyih. – // SPb.: BHV-Peterburg, 2009.
5. Schmidt A. A modular neural network architecture with additional generalization abilities for high dimensional input vectors. – // Manchester Metropolitan University, Department of Computing, September 1996. – 113 p.
6. Reznik A.M., Kussul M.E., Syichov A.S., Sadovaya E.G., Kalina E.A. Sistema avtomatizirovannogo proektirovaniya modulnyih neyronnyih setey CAD MNN. – // MatematichnI mashini I sistemi.– 2002, # 3.– C.28-36.
7. Chaudhari S.R., Dhawale C.A. A survey on application of ANNs in solid waste prediction. – // International Journal of Advanced Research in Computer and Communication Engineering, vol. 5, issue 10, October 2016, 110-113.
8. Trokhymchuk R.M., Kozlov K.E. Application of modular neural networks for analysis and prediction of socio-economical and ecological indicators of the region. – // Вісник КНУ імені Тараса Шевченка. Кібернетика. Вип.1(15), 2015, с.63–72.

RESUME

Р.М. Трохимчук

Результати застосування модульних штучних нейронних мереж для аналізу інтелектуальних даних та прогнозування процесів у сфері захисту навколишнього середовища

У цій роботі подано результати застосування модульної штучної нейронної мережі для інтелектуального аналізу даних і прогнозування процесів в області екології та охорони довкілля. Побудована модель була апробована для аналізу та прогнозування об’ємів і складу твердих побутових відходів на певних територіях чотирьох країн з істотно різними соціально-економічними та демографічними показниками.

Для порівняння та оцінки отриманих результатів було випробувано різні інші традиційні методи аналізу даних: регресійний і кореляційний аналіз, інтерполювання, математичне (аналітичне) моделювання та звичайні штучні нейронні мережі (багатошаровий персептрон і когнітрон). Порівняльний аналіз показав суттєві переваги модульної штучної нейронної мережі, що складалась з трьох модулів (мережа Хопфілда, багатошаровий персептрон Розенблатта і неокогнітрон), в ефективності, якості та точності прогнозування, зокрема, у передбаченні сезонних коливань значень аналізованих параметрів.

Отримані результати дали змогу сформулювати рекомендації фахівцям-екологам, які саме дані, з якою періодичністю і в якому обсязі слід визначати для підвищення якості прогнозування досліджуваних процесів.

Надійшла до редакції 11.09.2016