American Journal of Applied Science 3 (2): 1698-1702, 2006
ISSN 1546-9239

The Artificial Neural Network Based Approach for Mortality Structure Analysis

[1]Sokolova M. Kursk, [2]Rashad J. Rasras, and [2]Skopin D.
[1]State Technical University, Russia
[2]Faculty of Engineering Technology, Al -Balqa' Applied University, Jordan

Abstract: The theoretical outcomes and experimental results of applying of Artificial Neural Networks towards mortality structure examination in Russian Federation are presented in the paper. The most influencing classes of diseases are selected by sensitivity analysis. It is proved that Neural Network Models appear to be an adequate tool for mortality structure examination. Some interesting results for males and females total mortality are discussed. Proposed technology can be used for support in decision making in social sphere.

Key words: Neural Networks, decision support systems, modular Neural Networks, optimization, mortality structure, diseases, sensitivity analysis

INTRODUCTION

Rapid technology growth and its numerous real-life applications lead to some contradictory tendencies, which appear in economy and sociology. On the one hand, economic development programs arrear to be more expensive and more complex that is why requirements to social and economical forecasting increase with the same proportion. Adequate demographical forecasts are to be of a great importance, as they are necessary as a foundation for economical and urban planning and as a core of decision-making support systems, which are used now in economy, management, finance, medicine, welfare, logistics, constriction and others areas and issues.

Statistical methods and special indexes represent an *ordinary* tool for demographic forecasting and include such methods as micromodelling, regression analysis, component and variance analysis and different indicators. These methods are being used for a long time, thought they cannot be used as an adequate tool for demographic parameters forecasting from the system approach positions.

In accordance with system approach, system of social and ecological parameters (including demographical parameters) represents a complex system. It is known that a complex system cannot be studied with statistical methods as we do not obtain full information about it, its structure and all its connections with upper *mega system (environment*). Statistical methods construct internal model structure and in case we do not take into account all the factors of such a model, we will not receive an adequate one. Eliminating or adding factors of a model also *cardinally* changes it, which will lead to wrong forecasting.

That is why a "black box" principle is applied for analysis of complex systems. In accordance with this principle, we do not know an internal model structure, but we create a model of systems reaction to different input factors or its reaction to changing environment. Such approach let us to take into account even influence of the factors we do not know anything about as their influence may be presented as a part of studies signals (processes).

One of the techniques used here is Artificial Neural Networks (ANN), which was successfully used for demographical and social information analysis[1] and became adequate and reliable tool for complex systems dynamics research. In this work Werbos created population forecasts and proved that ANN approach showed better results than ARIMA methods. In work[2] was made a comparative analysis of ANN and Frailty Index to Assessment of Individual Risk of Death, which stated, that ANN improved accuracy of survival classification compared with an un-weighted frailty index.

In this paper we apply ANN and analyze their ability to simulate complex demographic processes. Some ANN-based models for medical and demographical parameters such as mortality and morbidity were created and influence of different diseases into total males and female's mortality indicator was studied. On the one hand, it was important to create adequate models of the processes and, on the other hand, to evaluate an influence of every disease. The models created with ANN could help to discover some hidden interconnections between into total mortality and morbidity in order to find out a mechanism, which can be used in decision support systems or in social medicine and welfare organizations. Statistical analysis of the ANN-based models is also represented.

Data structure: An important step in data preprocessing is development of a database, which include all the classes of diseases. The database, which

Corresponding Author: Sokolova M.; Kursk State Technical University; Russia

was used in the work, includes 176 variables: total mortality and morbidity (175 variables with respect to 1988 year's classification of mortality from diseases, which includes mortality from blood and cardiovascular systems, respiration and digestion system diseases, cancer cases and traumatic cases, etc.) since 1965 till 2005 years. Data were represented in standardized death rate (per 1 million).

ANN model: The origin of neural networks dates back to the 1940s. First works dedicated to simple computing devices that could model neurological activity and learning within these networks belonged to McCulloch and Pitts (1943) and Hebb (1949). In 1962 Rosenblatt described a perceptron and its computational ability in single-layer feed-forward network. Later, Rumelhart, Hinton and Williams (1986) introduced the generalized delta rule for learning by back propagation (which was described before by Werbos in his Ph.D. thesis), which is the most commonly used training algorithm for multi-layer networks today. More complex network types, alternative training algorithms involving network growth and pruning and an increasing number of application areas characterize the nowadays state of this area of science. In our work we used:

* Multi-layer ANNs, which were learnt with backpropagation, rule with momentum.
* Multi-layer ANNs which were learnt with genetic algorithm.

We use multi-layer ANN as the most common type of network and the type which shows good results in function approximation and gives a possibility to make sensitivity analysis of inputs. We choose genetics optimization of ANN weights as it is a technique of directed stochastic search and can simultaneously possesses a large amount of candidate solutions to a problem, called population.

In genetic approach, candidate solutions are perceived as individuals often called chromosomes. Initial solution population could be generated randomly. Selecting individuals playing the role of parents and creating children from the parents by genetic operators change the population: mutation and crossover. Selecting the survivors is again done using a fitness-based random mechanism, which is a survival-of-the-fittest principle. The genetic operators: mutation, crossover and selection continue until meet the stopping criterion (minimal of an error function or given number of epoch).

ANN structure: Artificial neural networks provide a robust approach to approximate real-valued, discrete-valued and vector-valued target functions from some very complex (input, output) data pairs. Many attribute-valued pairs represent input data.

The most common neural network architecture is a fully connected node web. Every edge on the web has a weight associated with it. Each input node accepts an input data $X=\{X_1, X_2, \ldots X_N\}$. All other nodes on hidden layers and output layer compute the linear combinations of the inputs from their immediate backward layers, apply a threshold on the results and send the outputs to their immediate forward layers. Often the threshold is a sigmoid function.

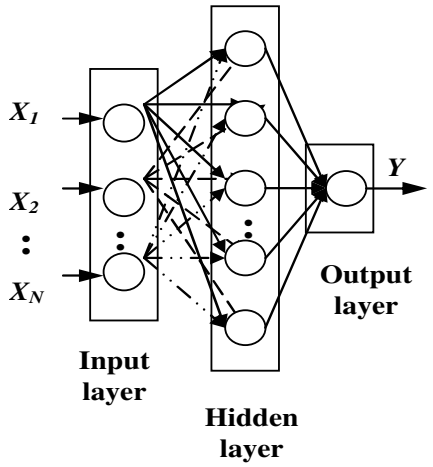


Fig. 1: Graph of a feed-forward ANN with one hidden layer

Data processing: Data sets were divided into training and testing non-overlapping subsets in proportion 95% to 5%. We decided to leave 95% of data for training in order to give ANN possibility to teach more examples. As it was mentioned above, there were 175 inputs and 1 output. In case of scarce data sets it was not possible to examine dependence between inputs and output with one ANN that is why we decided to divide input variables into groups which will represent inputs for separate ANN. Data sets included 79 values for every variable and training data sets included 75 values.

We have to state such a number of input variables, which let an ANN reach generalization while being taught with given training data sets. Generalization is influenced by three factors: the size of training set, the architecture of ANN and the physical complexity of the problem at hand[3]. As the training set is fixed, we may change only the architecture of ANN and will use a rule derived by specialists of Ward corp., which stated that number of hidden neurons can be calculated as:

$$N_{hidden} = \frac{1}{2}(N_{input} + N_{output}) + \sqrt{DS},$$

Where DS-is number of examples in data set; N_{input}-number of input neurons; N_{output}-number of input neurons. According to that rule, number of hidden neurons in case of 5 inputs and 75 examples in data sets will be equal or less than 11. If we use Modular ANN, we divide the total number of hidden neurons between layers.

So, every ANN had the same architecture, consisted from 5 inputs, four hidden layers, consisting from 2 neurons each and 1 output.

On the first stage we created 70 ANN (35 for males and 35 for females total mortality), which was dedicated to find out which inputs influence outputs most. We evaluated influence by sensitivity analysis and noticed one o most important inputs for every ANN. Then, we created ANN, which consisted from selected inputs and repeated sensitivity analysis.

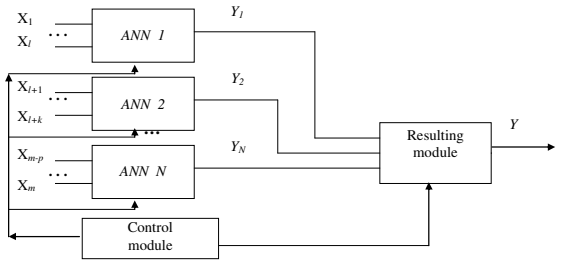


Fig. 2: Structure of a modular ANN

We continued this procedure until we have several most influencing inputs.

Input variables were named after their classification numbers-from 1 to 175, so, names of input variables were from *r1* till *r175*.

Modular ANN structure: ANN is nonparametric models and like all nonparametric models requires a large number of observations to build and evaluate the model. But data is always finite and most often scarce in real world applications, especially, in medical demography [4].

As experimental retrospective data were taken from statistical yearbooks, we could obtain only 2 evaluations per year that is why data sets appeared to be scarce. This problem can be solved with using modular ANN.

Modular ANN (MANN) realize an idea that committees of solving rules or ANNs[5] will represent better results that every solving rule or ANN separately.

In accordance with a Fig. 1, the Resulting module calculates and output function Y as:

$$Y = \frac{\sum_{i=1}^{N} Y_i \cdot R_i^2}{\sum_{i=1}^{N} R_i^2},$$

Where *N*-number of neural networks in a committee, y_i-output of ANN_i, R_i^2-R square of ANN_i, *Y*-a resulting output.

Sensitivity analysis: Sensitivity analysis provides a measure of the relative importance among the inputs of the neural model and illustrates how the model output varies in response to variation of an input. The first input is varied between its mean +/- a defined number of standard deviations while all other inputs are fixed at their respective means. The output of every MANN was computed for a defined number of steps above and below the mean. This process was repeated for each input. Sensitivity analysis let us select most influencing inputs.

Ann software: As software were used NeuroSolutions 4.24 and its module for Excel. During the training

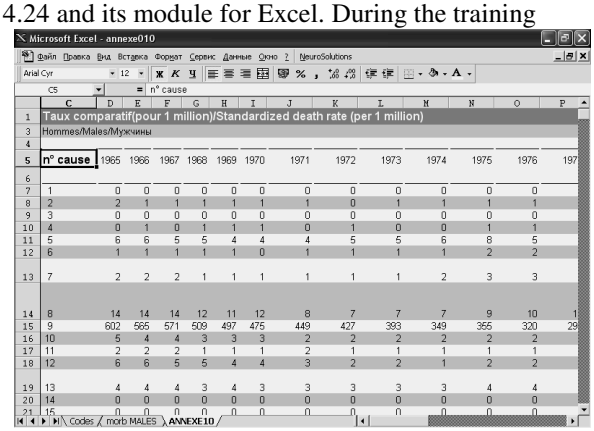


Fig. 3: Table with statistical data sets

sessions the networks parameters were calculated according to the backpropagation algorithm with parameters of learning rate = 0.1, momentum=0.7, number of epochs=1000, activation function-sigmoid, initial threshold =0.3. When using genetic training, we used the following parameters: Population = 50 individuals, selection-of Roulette type in accordance with rank of the every chromosome, probability of one-pointed crossover and mutation is proportional to chromosome's fitness, number of generations is 100.

We created MANN for every set of input variables. As inputs were taken several (4 or 5) variables representing morbidity cases and total mortality (for males in one case and for females in other) was taken as output. It was created and tested more that 600 MANNs and finally selected 70 (for both males and females total mortality), which simulated dependence between mortality and morbidity classes and 23 MANNs, which simulated dependence between mortality and mostly influencing morbidity classes. So, finally, we receives 12 MANN for males total mortality and 11 for females total mortality.

Statistical analysis: It has frequently been argued that statistical error measures do not measure the right thing in case of ANN, therefore many authors developed additional evaluation methods[6]. So, in our work the MANN-based models were estimated using statistical criterions presented in NeuroSolutions

For a multi-layer ANN:

* Mean Squared Error (MSE)
* Normalized MSE (MSE/variance of desired output)
* Mean Absolute Error (MAE)
* Minimal Absolute Error (MinAbsError)
* Maximal Absolute Error (MaxAbsError)
* Linear correlation coefficient (r)

For an ANN trained with genetic algorithms:

* Generation when the best fitness was at its minimum.
* Minimal best fitness criterion (Minimum MSE)
* Best fitness at last generation (Final MSE)

RESULTS

We received MANN-based models for medical and demographical parameters: mortality and morbidity (with respect to its classification from 1988). Data sets were presented in Excel tables: first 40 columns-for value dated for 1 of January of correspondent year and the following 39 columns contained values dated for 1 of July of corresponded years (1965-2005).

We found out the most influencing morbidity classes towards total mortality. Firstly, we created about 300 ANNs for males' mortality and, in accordance with their statistic estimations chouse 35 best ones. Then we did sensitivity analysis and found out that the most influencing to total mortality diseases were 28. So, we decreased input variables number from 175 till 28 most important ones. On the Figure Sensitivity Analysis Diagram is represented. We can see that most changeable variable about the Mean is variable *r84*, it means that *r84* is the most influencing input and corresponding mortality class will impact mostly to Total mortality.

Below there is Fig. 5, which explains how sensitivity analysis was done for variables r12 and r84. So, variable r12 do not react significantly to mean, but variable r84 shows high dependence about the Mean. We calculated that the most influencing morbidity classes for males' mortality are:

r94	Other forms of ischemic heart diseases with essential hypertension
r96	Acute myocardial infarct with essential hypertension
r84	Meningitis
r86	Others central nervous system inflammatory diseases
r90	Acute rheumatism
r142	Skin and subcutaneous cellular tissue diseases
r153	Congenital and aspiration pneumonia
r150	Others congenital anomaly
r122	Alcoholic hepatocirrhosis
r115	Stomach ulcer
r114	Others respiratory organs diseases

On the other hand, from quantitative point of view (if we calculate number of diseases for a million), most numerous morbidity classes for male's mortality are:

* Aterosclerotic cardiosclerosis
* Vascular brain lesions with essential hypertension
* Vascular brain lesions without essential hypertension, diseases of arterioles, capillaries, diseases of arteries
* Vascular brain lesions without essential hypertension, diseases of arterioles, capillaries, diseases of arteries
* Stagnant and hydrostatic diseases of lungs and postinflammatory pulmonary fibrosis
* Accident evoked by firearms

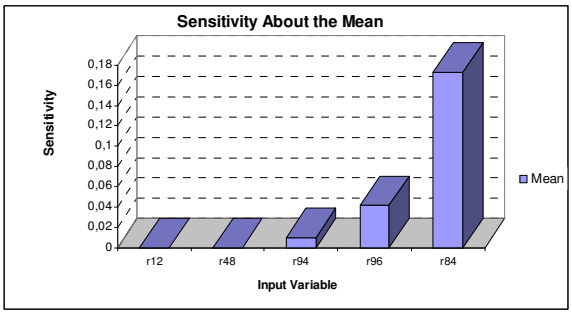
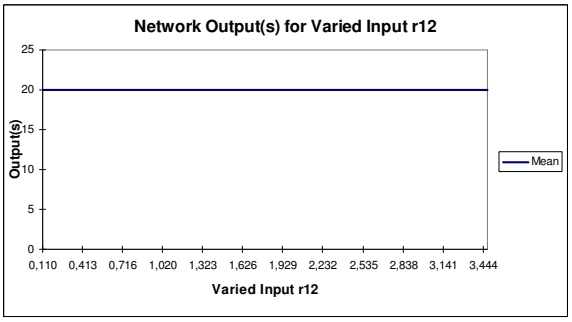


Fig. 4: Sensitivity analysis diagram for inputs of the MANN



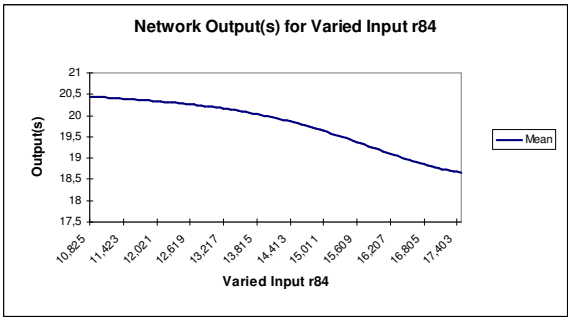


Fig. 5: Charts of sensitivity analysis for input variables *r12* and *r84*

* Accident evoked by electric current.
* Tuberculosis of respiratory organs
* Stomach malignant neoplasms
* Malignant neoplasms of trachea, bronchus and lungs.

Morbidity for females has a different structure: if we calculate number of diseases for a million, then we notice that number of malignant neoplasms, liver cirrhosis, lung tuberculosis, Accident evoked by fire; suicides; auto transport accident in public roads caused by pedestrian incursion are significantly lower then the same from Males morbidity. When examining females'

mortality we created about 300 ANNs and chouse 35 best ones analyzing their statistic estimations. Then we realized sensitivity analysis and decreases number of most important morbidity classes to 39. We calculated that the most influencing morbidity classes for females' mortality are:

* Malignant neoplasms of lips, oral cavity and throat.
* Diseases of other endocrine glands
* Anemia
* Others diseases of blood and hematopoietic organs
* Unspecified lesions of pericardium, mitral, aortal and pulmonary valves
* Pericardium lesions, lesions mitral, aortal and pulmonary valves
* Vascular brain lesions with essential hypertension
* Others blood circulation diseases.
* Empyema, lung and mediastinum abscesses
* Others diseases of genitals
* Extrauterine pregnancy
* Toxicosis of pregnancy
* Others congenital anomalies of blood circulation
* Accidental drowning and water immersion

DISCUSSION

We tested application of the ANN based on statistical surveys information, in modeling morbidity and mortality in Russia for a period from 1965 till 2005. The networks showed high accuracy, which was evaluated by some statistical criterions: MSE, r, MAE, MinAbsError, MaxAbsError, etc. We created several MANN for every "inputs-output" set and selected the best networks taking their MSE and r as selection criterions. For selected ANN r belonged to interval from 0,69 till 0.97, that shows high correlation between real data and data, approximated with MANN.

So, we may conclude, that that difference in structure of "most numerous" and "most influenced" diseases take place because the latter class of diseases ("most influenced") impact total mortality more and may influence with lag or with some hidden mechanisms, which we will learn in our following research. Interestingly, but ANN-based models selected some morbidity classes, for example, r96-Acute myocardial infarct with essential hypertension, r84-Meningitis as highly influencing, thought their values decrease since 1965 till 2005. This high dependence can be explained as high mortality rate for people suffering from those diseases.

Comparing gender aspects of morbidity, it appeared that in case of males mortality there are 28 diseases which significantly influence into total mortality, but in case of females mortality there are 39 with more death cases connected with pregnancy and childbirth, diseases of blood and cardiovascular systems. Also "Alcoholism" for females' mortality appeared to be mo re influencing than for men. We understand that finding out the most influencing diseases is of great importance as such knowledge can be used as a foundation for decision support systems for administrative, medical and welfare institutions and, of course, of great practical interest. One of limitations of our study is that we do not compare ANN approach with others methods but just noticed it in Introduction, that do not give visible example. Also, still, the MANN models were not tested on separated data set because of scarce amounts of data. Suppose that this is an important area of additional study in this area.

REFERENCES

1. Werbos, P.J., 1974. Beyond regression: New tools for prediction and analysis in the behavioral sciences. Ph.D. in Statistics. Harvard University.
2. Song, X., A. Mitnitski, Ch. MacKnight and K. Rockwood, 2004. Assessment of individual risk of death using self-report data: An artificial neural network compared to a frailty index. J. Am. Geriat. Soc., Iss. 7.
3. Haykin, S., 1994. Neural Networks. A Comprehensive Foundation. New York, NY: Macmillan.
4. Sokolova, M. and O. Hudec, 2003. Neural networks for economical data prediction. 6th Intl. Scientific Conf. Applications of Mathematics and Statistics in Economy. Banska Bystrica, Slovakia.
5. Rastrigin, L. and R. Erenshtein, 1981. The Approach of Collective Identification. M.: Energoizdat, Iss. 7.
6. Steckler, H.O., 1991. Macroeconomic forecast evaluation techniques. Intl. J. Forecasting, 7.