Development of C# Application for Neural Network Based Precipitation Data Mining

Miljan Jeremić, Milan Gocić, Miljana Milić, Jelena Milojković

Abstract – The impact of climate changes, and especially its negative effects to human society, represent potentially devastating problems of today and the future that cannot be easily confronted, defined, and solved. In order to successfully combat climate changes' negative effects, it is necessary to assess their actual beginning in time as well as its duration and the final effects, but it is also difficult to determine the intensity and points of observation. Due to the extensive damage caused over a long period of time, it is necessary to find reliable and standardized indicators in space and time that would define phenomena such as rainfall, droughts and floods. The paper presents an application in C# programming language that, using data from multiple measuring stations in Niš, Serbia recorded in a database, and a machine learning technique, such as Data Mining, creates a neural network that models and analyzes a rainfall index for a period of 20 years.

Keywords – Climate changes, Data mining, neural networks, machine learning.

I. Introduction

Learning is a process that humans perform almost continuously. Since the ability to learn is a basic characteristic of intelligent beings it is not surprising that machine learning represents the central area of research in artificial intelligence. Artificial systems that are capable of learning, evidently improve their performance, while the intelligent biological systems increase the probability of their survival and extension of the species.

Machine learning [1, 2] is an area that studies the processes that underline learning in humans as well as in artificial systems. In order to cover all relevant aspects, it relies on a number of other disciplines, including artificial intelligence, probability and statistics, information theory, psychology, neurobiology, and control theory

Inductive learning or inductive empirical learning has the greatest importance in the field of computing and artificial intelligence. The idea of this type of learning is to learn from the available examples or to learn from the experience of others. Considering the object of learning, the most general learning methodology is the functional learning, i.ee input-output mappings.

Measuring the quality of a trained system is one of the key problems in machine learning. It can easily happen that the chosen hypothesis represents the most reliable one for a given training set, but also very unreliable on new examples that did not participate in the training i.e. the test set. The ability to generalize or to correctly deal with instances that were not the part of the training set data, is a measure of the quality of any machine learning algorithm. The synthesis is based on two sets, a training set and a test set. The test set data should not be used in the training stage. A common statistical measure of system performance is the mean square error that is calculated using the system's output values and target values.

In practical applications there are many different machine learning models, and their structure depends on the information contained in the training sets, how the training process is controlled, and whether the training system can affect the surrounding.

The aim of this paper is to describe the entire creation process of a Data Mining architecture for a system that, based on a neural network type machine learning model, shows the visual result of rainfall estimation in the territory of the city of Niš during one year. It begins with data acquisition from the relational database, and ends with the visualization of the machine learning algorithm's conclusions. In the next section few most important machine learning algorithms and techniques will be explained briefly. Then, the basic information about the initial precipitation data will be listed. It is followed by the applied Data Mining methodology [3], and its results. The paper ends with some concluding remarks.

II. Machine Learning Algorithms and Techniques

There are many available data science techniques today, such as linear regression, logical regression, SVM (Support Vector Machine), Random Forest algorithm, K-means clustering, neural networks and convolutional neural networks, and they are most commonly applied on data in the form of an Excel CSV file or in JSON format.

One of the best-known techniques of machine training are neural networks and decision trees [4, 5].

A. Decision making systems

M. Jeremić and M. Gocić are with the Facultyof Civil Engineering and Architecture, University of Niš, Aleksandra Medvedeva 14, Niš, Serbia E-mail miljan.jeremic@gmail.com, milan.gocic@gaf.ni.ac.rs.

M. Milić and J. Milojković are with the Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14, Niš, Serbia E-mail: miljana.milic@elfak.ni.ac.rs, jelena.milojkovic@elfak.ni.ac.rs.

There are three large groups of machine learning techniques [2]: supervised learning, unsupervised learning and reinforcement learning.

In the supervised learning training set is in the form $\{x_i, f(x_i)\}$, where $f(x_i)$ denotes a target value for the instance x_i. The task of the training is to find the approximation for the function f, while the measure of performance is the quality of the approximation for points that are outside of the training set.

For the unsupervised learning, the training set contains only the input instances $\{x_i\}$ A typical problem for the unsupervised learning is the clustering problem, i.e. the classification of the available data into a smaller number of groups. Since the clustering is done on the basis of data similarities, it means that one cluster contains data with similar properties. The reinforcement learning is more complex learning concept in compare with the supervised and the unsupervised one. It originates from the control theory in which a dynamic environment can be described with a condition, action and a reward. Within this type of learning technique, it is necessary to learn how to map the situations into actions, while achieving the maximal reward. The training algorithm is not familiar with types of actions that should be performed for the particular situation. An example of such a learning process is learning how to play chess.

B. Neural networks

Neural networks [8] are designed to mimic the structure of neurons in the human brain. Each artificial neuron should be connected to other neurons within the system in a proper way. Neural networks consist of layers of neurons. Data is transferred from one layer of neurons to the next - inner layer. Eventually, the signals within the network arrive at the output layer, where the network presents its best troubleshooting assumptions.

The examples of neural networks; applications could be found in a series of different industries, and telecommunications and media as well. They could be used for text or language translations, frauds detections etc. Also, their usage is common in digital filters, function approximations, sample detections, forecasting in many complex processes. One large area that exploits neural networks is gaming

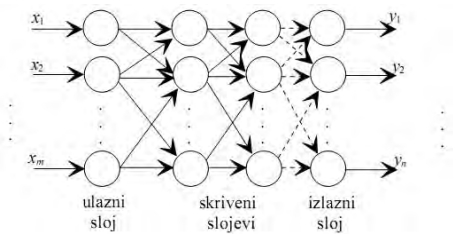

Fig 1. A neural network with its connections

The advantages of neural networks are that they can be used to model complex systems, they are more robust and give more accurate predictions than linear models. The disadvantage of this technique is that it is not suitable for hypothesis testing and is therefore used in combination with statistical techniques [6].

There are many types of artificial neural networks topologies today [7], but the most commonly used is a feedforward back propagation multilayer perceptron. Based on a set of known inputs and outputs for the real problem, training of the network is performed by applying the known inputs of the modeled problem to the inputs of the network, and observing the difference between the known output of the problem and the output of the neural network. The aim of the training is to minimize this error. Artificial Neural Network mimics the functioning of the human brain, where the data processing is performed by connecting a large number of neurons. A typical neural network consists of multiple layers, one input, one or more hidden layers, and one output layer. The number of hidden layers depends on the complexity of the problem being modeled, but for common classes of problems one or two appear to be enough. Figure 1 shows a neural network consisting of three layers: the input, hidden and the output layer.

Neural networks accept multiple inputs in parallel and process the information received in a distributed manner. The information stored in the neural network is distributed across multiple computing units, which is opposite to the conventional information storage in memories where each specific information is stored in its memory space. The property of distributed information storage as well as the redundancy are the most important benefits [8].

A neural network has the ability to learn and adopt, which makes them applicable for processing uncompleted datasets in the unknown or unlearned environment. The network can generalize and conclude how to process data that were not present in its training set. Considering their structure, neural networks represent multivariable systems, which makes them useful for modeling, identification and control of multivariable processes.

C. Data Mining

Data Mining can be defined as the procedure of examining large databases in order to generate some new information about these data and their relations. It is a technique used to discover the hidden, valid, and potentially useful regularities and patterns amongst the data in large datasets. It employs machine learning, statistics, AI and database technology and is therefore a complex multi-disciplinary procedure that can be useful in solving everyday problems using data.

Very often, classic database reports do not provide data required for the end users, or decision makers. Some advanced techniques then need to be employed in order to extract more specific information or even business predictions from the available data.

The idea is to replace the decision makers with an appropriate software that can predict for a client [2]. In order to perform such analyzes the knowledge or support of a particular set of algorithms is required. It should be mentioned that Data Mining also uses some statistical methods for data analysis.

There are few Data Mining algorithms that should be studied first in order to select the one that is most suitable for a particular data analysis problem

Some of the most often Data Mining algorithms are: classification algorithms, regression algorithms (linear or logic regression dedicated for one or two steps ahead forecasting), segmentation/cauterization algorithms (objects are separated into sets with some common properties), association algorithms (dedicated for the correlation search among data) and sequence analysis algorithms (they help in search for the user's path through the web, or the order of inserting items into the cart, etc.) .

Classification is the most common Data Mining methodology, whose goal is to accurately predict the class of the target object that has the unknown class label [9-10]

III. DATA AND THE FIELD OF STUDY

In this section, a step by step procedure of the Data Mining system development will be given. The selected machine learning approach that supports our Data Mining problem are Neural Networks. An *SQL Server Data Tools* toolset, *Visual Studio,* and a *C#* programing language are used for this purpose [11]. The problem to be solved must be defined and presented through Data Mining analysis.

The procedure is exemplified on the analysis of the precipitation value for the city of Niš, using the available application and its neural network algorithm in order to generate a graphical representation of the detected hidden regularities in the available data.

At the beginning of the procedure, one needs to determine what kind of data set we will be applied as the input of the Data Mining algorithm. The data for the selected Data Mining algorithm need to be prepared for further analysis and decision making. Data necessary for the Data Mining analysis are usually preprocessed first. It should be defined what are the input data or variables, and what should be the outputs i.e. for example the forecasts. It is usually not possible to get a direct access to a data storage, a relational database, or an Excel spreadsheet and expect some output to be generated from that vast amount of written data.

The data from different sources should be selected, cleaned, transformed, formatted, anonymized, and constructed (if required).

Data cleaning is a process where noisy data are removed from the series, while the missing data are generated and filled.

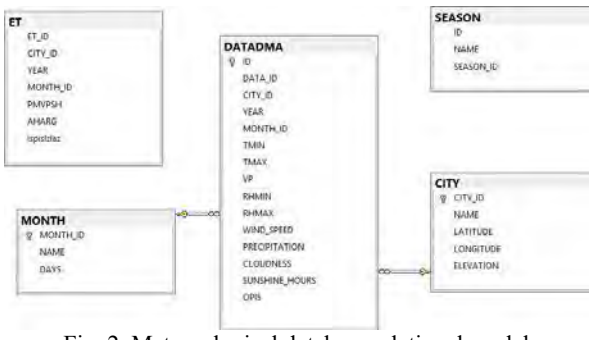

Fig. 2. Meteorological database relational model

The intention here is to analyze the output values for the PRECIPITATION variable recorded and stored in an SQL Server database. The relational model of the meteorological database is shown in Figure 2.

Some data transformations could be applied on data to make it useful for Data Mining. Some of them are: smoothing, aggregation, generalization, normalization or attribute construction.

For the requirements of this application, a query has been generated in the SQL Server database, where a new column is formed that tests each value of the PRECIPITATION and classifies it into three categories: small, medium and large. This table is denoted with DM and it will record all the values of the required variables applied at the input. The second table, DM1, will be created after the SQL Server query is executed. Results of this query will be written in a Predicted column. The table has 348 records in total. This will be a source of data for further Data Mining analysis.

After creating the appropriate source of the required data storage, it is now possible to begin with the development of the application in a specific environment. Here, we have used a Visual Studio, version 2015. For a given project and the available data, it was the most suitable to use Data Tools toolset.

IV. METHODOLOGY

Within the *Data Mining* algorithms, the input data are processed and the corresponding values are generated at the particular outputs. After all necessary preparations over the data, different mathematical models could be applied to determine patterns amongst data. For our problem neural network-based classification modelling technique was selected. Other available techniques are decision trees genetic algorithms and similar.

Within the Data Tools environment, the Business Intelligence tab is used to create the new project. Then the *Analysis Services Multidimensional* and *Data Mining Analysis* were used. These toolboxes as well as the Data Mining mode must be installed and enabled within the SQL Server.

In any Data Mining project, data sources and their views are basic objects used to define tables and queries for Data Mining structures. In turn, a structure defines one or

more tables or columns that are used as input attributes, keys, and prediction outputs for the structure.

Now our particular mining structure can be created, with the defined source of data the desired mining algorithm. In this case the structure's source of data is the database with the values for meteorological parameters, which is shown in Figure 3. After this step, a connection between the mining structure and the database should be established.

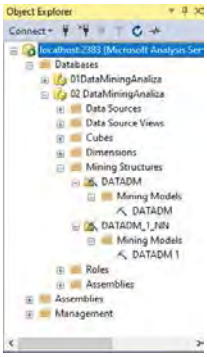

Fig. 3. Creation of the *Mining* model and the corresponding neural network withing the Analysis Services

As mentioned before, the created model has the direct access to the database in order to select the required table. The next step is to create a logical data view, or Data Source View. As a Mining structure the Decision Trees algorithm is selected, which is used to predict or classify one or more discrete variables. In this example, this is the PRECIPITATION variable.

When the algorithm processes the data, it saves some of the data for later testing. In the next step, the percentage of the data that will represent the test set should be selected. In this case this is 20%.

Finally, the Mining Model Viewer tab shows the results of neural network-based Data Mining. This as well as the generated precipitation decision tree is shown in Figures 4 and 5, respectively.

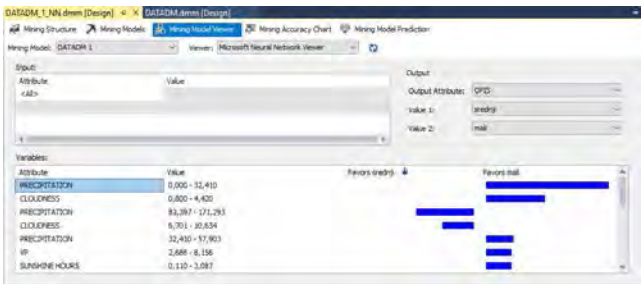

Fig 4. The report window generated by the C# program

The generated data model is uploaded to Analyses Services within the DM database. It should be noted that within the SQL Server one can find a model in the form of the decision tree or a neural network, the same as one shown in the C # application. A prediction query can also be created in this part of the database. Now some reports can be created that are results of data processed through our data model.

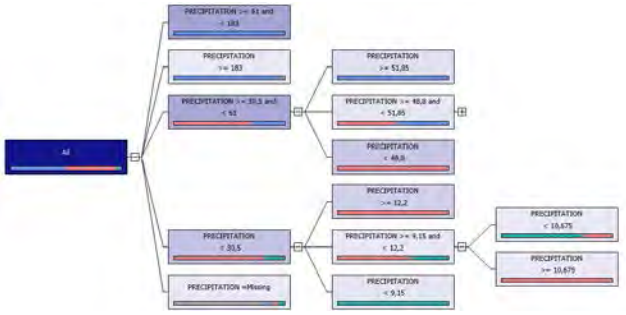

Fig 5. The decision trees for the PRECIPITATION variable

V. Results and Discussion

Queries are generated with a single mouse click, without prior knowledge of the query. The query itself should eventually consider a prediction function to be applied, and make a prediction based on one or more input columns. After running the query, a probability of the occurrence for a particular output value is obtained based on a previously created data model.

In this paper the meteorological data analysis for the period from 1990. to 2018. stored in the SQL Server database is performed. When the PRECIPITATION is taken as an observation parameter, the data Mining model was created over these data, that applied a neural network-based analysis structure. The algorithm offered the following conclusions about the given PRECIPITATION data: the expected average value for a month should be greater than 32.411mm; the most probable future value would be between 83 and 172mm, which can be classified as the medium precipitation value.

VI. Conclusion

The most important benefits of the Data Mining procedure are: the ability to create useful knowledge-based information for the companies, and act accordingly and timely. It is a cheap and efficient methodology that facilitates reasonable and automated forecasting and decision-making, by quickly discovering the hidden correlations amongst data

Creating such an application that works with real data, and generates a neural network, provides the necessary background for the needs of drought analysis and assessment in order to mitigate its unwanted consequences.

References

[1] Tan, P. N., and Steinbach, M., Kumar, V., Introduction to Data Mining, Pearson Education, 2006.

[2] Milosavljevic, M. *Artificial Intelligence (in Serbian)*, Singidunum University, Belgrade, 2015.

[3] Chiwara, A. A., and Gupta, H., *Data Mining, Concepts and Techniques Association Rule Mining*, State University of New York, CSE 634, Chapter 8, 2006.
[4] Bhargava, N., Sharma, G., Bhargava, R., and Mathuria, M. "Decision tree analysis on j48 algorithm for data mining", *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering*, *3*(6), 2013.
[5] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., and Zhou, Z. H. "Top 10 algorithms in data mining", *Knowledge and information systems*, *14*(1), pp. 1-37, 2008.
[6] Kalinić, Z. S., and Marinković, V., "Određivanje relativnog uticaja pojedinih faktora na prihvatanje mobilne trgovine primenom neuronskih mreža", *Poslovna ekonomija*, 10(2), pp. 206-223, 2016.
[7] Lu, H., Setiono, R., and Liu, H.," Effective data mining using neural networks", *IEEE transactions on knowledge and data engineering*, *8*(6), pp. 957-961, 1996.
[8] Craven, M. W., and Shavlik, J. W. "Using neural networks for data mining" *Future generation computer* systems, 13(2-3), pp. 211-229 1997.
[9] Kesavaraj, G. and Sukumaran, S. "A Study on Classification Techniques in Data Mining" *Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT),* pp. 1-7. IEEE, 2013.
[10] Begüm Ç. and Deniz Ü., "Comparison of Data Mining Classification Algorithms Determining the Default Risk", *cientific Programming*, *2019*.
[11] Watson, B., C# 4.0 kako do rešenja, SAMS, Mikro knjiga, 2011.