Annals of the University of Craiova, Mathematics and Computer Science Series
Volume 48(2), 2021, Pages 358–373
ISSN: 1223-6934

New approaches to processing GIS Data using Artificial Neural Networks models

Dana Mihai

Abstract. Spatial data mining is a special type of data mining. The main difference between data mining and spatial data mining is that in spatial data mining tasks we use not only non-spatial attributes but also spatial attributes. Spatial data mining techniques have strong relationship with GIS (Geographical Information System) and are widely used in GIS for inferring association among spatial attributes, clustering and classifying information with respect to spatial attributes. In this paper we use the statistical package Weka on two models, which consist of two parcels plans from the Olt area of Romania. In our experimentation, we compare the results of the vector models depending on the values of the training datasets. Using these models with GIS data from the domain of Cadaster we analyze the performance of the Artificial Neural Networks in context of spatial data mining.

2010 Mathematics Subject Classification. Primary 60J05; Secondary 60J20.
Key words and phrases. Spatial data mining, Classification, GIS, Artificial Neural Networks, Weka.

1. Introduction

The explosive growth of spatial data and extensive use of spatial databases emphasize the need for the computerized discovery of spatial knowledge. Spatial data mining is the process of discovering motivating and previously unknown, but potentially useful patterns from spatial databases. The difficulty of spatial data and essential spatial relationships restrict the convenience of conventional data mining techniques for extracting spatial patterns [2].

The spatial data mining is a newly arisen edge course when computer technique, database applied technique and management decision support technique develop the certain stage. The spatial data mining gather productions that come from machine learning, pattern recognition, database, statistics, artificial intelligence and management information systems. According to different theories, the different methods of spatial data mining, such as methods in statistics, proof theories, rule inductive, association rules, cluster analysis, spatial analysis, fuzzy sets, cloud theories, rough sets, neural network, decision tree and spatial data mining technique based on information entropy are brought forward [2].

Spatial data mining techniques were collectively used with GIS and satellite imagery in various studies to mine interesting facts associated in diverse domains' applications such as traffic risk analysis, fire accident analysis, analysis of forest extent

Received September 30, 2021. Accepted December 5, 2021.

changes, grading of agriculture land, analysis of railways, farming and forestry, warehouse, transport, tourism, military, geology, soil quality monitoring, water resource monitoring, and deforestation, land allocation, meteorology [3].

Among to the basic tasks of spatial data mining we mention: the *Spatial classification* - finds a set of rules which determine the class of the classified object according to its attribute or classification of a pixel into one of classes, the *Spatial clustering* - groups the object from database into clusters in such a way that object is one cluster are similar and objects from different clusters are dissimilar, the *Spatial association rules* - find (spatially related) rules from the database and describe patterns which are often in the database, the *Spatial characteristic rules* - describe some part of database, the *Spatial discriminant rules* - describe differences between two parts of database, the *Spatial trend detection* - finds trends in database, a trend is a temporal pattern in some time series data and a spatial trend (global trends and local trends) is defined as a pattern of change of a non-spatial attribute in the neighborhood of a spatial object or we can find changes of pixel classification of a given area in the last five years and the *Spatial exceptions* - outliers that are isolated from common rules or derivate from other data observations very much [1], [4], [6].

Spatial classification is about grouping data items into classes (categories) according to their properties (attribute values). It is also called supervised classification, as opposed to the unsupervised classification (clustering). The supervised classification needs a training dataset to train (or configure) the classification model, a validation dataset to validate (or optimize) the configuration, and a test dataset to evaluate the performance of the trained model [5]. It is used to partition sets of spatial objects. Spatial objects could be classified using non-spatial attributes, spatial predicated, or spatial and non-spatial attributes. The spatial object has been classified by using its attributes. Each classified object is assigned a class. It is the method of finding a set of rules to decide the class of a spatial object. It classifies attributes of the object along with neighboring objects with their spatial relation. The spatial classification methods such as decision trees (C4.5), artificial neural network, remote sensing, spatial autoregressive regression are used to find the group of the spatial objects together [3].

The study in this paper focuses to show the statistical results obtained by applying the Artificial Neural Networks on the two spatial datasets. The main purpose of these results presented in the current research is to contribute to the literature in order to develop classification techniques based on spatial data with very good results through applicability, with very small errors which represent the accuracy of the characteristics of objects on the earth's surface and also emphasize the practical importance of the spatial data mining domain.

The paper proposes the applications of spatial classification and it is organized as follows: in Section 1 we present the spatial data mining domain with the fundamentals rules and the description of the spatial classification, in Section 2 the related work develops ones of the most important researchers for Artificial Neural Networks, in Section 3 we expose the Artificial Neural Networks in the spatial data mining domain, Section 4 describes the two GIS data models, in Section 5 we present and analyse the results based on the experimental steps and finally we come to the conclusions and future work in Section 6.

2. Related work

In the next part some important researchers for Artificial Neural Networks are developed.

P. K. S. C. Jayasinghe and M. Joshida [7] described in their work an application of the ANN modeling with GIS technology to predict the potential area for beetroot in Sri Lanka, based on nine crop requirement factors. Soil properties, meteorological data, current land use and slope accessibility were considered as important factors to identify potential lands for beetroot. Average annual temperature and precipitation (1961 - 1990 data), topographic, soil and land use maps of the study area were used for the study. The ANN model based on LM back - propagation algorithm was developed, which has a 9-22-4 network structure. The result suggests that the neural network based - GIS modeling can be powerful alternative approach toward automated spatial decision making. Also the study has demonstrated the capabilities of using neural network, GIS and field data for identifying potential sites for beetroot.

In the paper [8] the authors J. Kashyap, A. Bansal and A. K. Sao include the modeling characteristic of artificial neural networks based on spatial feature. The estimated model was initiated and tested with Indian solar horizontal irradiation (GHI) metrological data and the results was evaluated with different statistical error. This works certifies the ability of ANN to accurately reproduce hour's global radiation forecast. ANN model is a well-organized technique to estimate the radiation using different meteorological database. In the paper nine spatial neighbor locations and 10 years of data for assessment of neural network were used.

J. Liu and H. Lu [9] presented an outlier detection algorithm based on SOM (Self Organizing Maps) neural network for the spatial series dataset. Outlier detection for spatial series dataset is a very meaningful and challenging task. The experimental results show that the outlier detection algorithm based on SOM neural network performs well in multidimensional spatial series dataset containing both isolated and assembled outliers.

The study of the L. Akil and H. A. Ahmad [10] showed a significant correlation between socioeconomic status and the increased rates of Salmonella especially in Mississippi which had higher rates than other neighboring states and some of the northern states. The principle objectives of this article are to determine the extension of Salmonella and Escherichia coli infections using geographical information systems (GIS) and neural network models. In this study, the model was created using four input variables and one output. NN models accounting for non-linearity predicted better association than regression models. GIS mapping was also shown to be a very useful instrument to map and visualize the areas and districts of highest Salmonella outbreaks in addition to the socioeconomic status. The results showed that Northeast and Tombigbee regions of Mississippi had the highest rates of Salmonella outbreaks. The northern region also had the highest rate of unemployment, and primary care provider rate was shown to be the lowest in the northwest and east-central. Also with the geographical and economic relation with infections diseases, the authors helped to determine effective methods to reduce outbreaks within these communities.

A. Mallabo, L. Mao and P. Rashidi [11] presented the machine learning techniques in spatial modeling can be applied to Tuberculosis (TB) incidence rate across the continental US. The authors collected 278 exploratory variables including environmental

and a broad range of socioeconomic features for modeling the disease across the continental US. They investigated the applicability of multi layer perceptron (MLP) ANN for predicting the disease incidence. Predictive performance of the MLP was compared with linear regression for test dataset using root mean square error, mean absolute error and correlation between model output and ground truth. Results of clusters analysis showed that there is a significant spatial clustering of smoothed TB incidence rate ($p < 0.05$) and the hotspots were mainly located in the southern and southeastern parts of the country. Among the developed models, single hidden layer MLP had the best test accuracy. Sensitivity analysis of the MLP model showed that immigrant population, and minimum temperature were among the factors with the strongest contributions.

In the article [12] the authors S. N. Rajan and A. K. Sinha studied the rate socioeconomic factors like migration rate [MGRT], ratio of female to male literacy [LTFM], average distance traveled by migrant/bridge population [AIMD], human development index [HDI], gender development index [GDI], in populating the disease in India. An Artificial Neural Network based model has been developed which has correlated these spatial and nonspatial factors with the various spread pattern of the disease. The machine learning process using back propagation algorithm of Artificial Neural Network has been successfully implemented with PL/SQL procedure developed on ORACLE database. The result of the model reveals an interesting pattern which in agreement with the report published by the government on the basis of the physical survey of various geographical locations.

The study [13] focused on the development of a methodology based on artificial neural networks (ANN) that is able to spatially predict soil units. Within a test area Rhineland - Palatinate (Germany), covering an area of about 600 km^2, a digital soil map was predicted. Based on feed-forward ANN with the resilient backpropagation learning algorithm, the optimal network topology was determined with one hidden layer an 15 to 30 cells depending on the soil unit to be predicted. To describe the occurrence of a soil unit and to train the ANN, 69 different terrain attributes, 53 geologic-petrographic units, and 3 types of land use were extracted from existing maps and databases. 80% of the predicted soil units (n=33) showed training errors (mean square error) of the ANN below 0.1, 43% were eve below 0.05 validation returned a mean accuracy of over 92% for the trained network outputs. Altogether, the presented methodology based on ANN and an extended digital terrain-analysis approach is time-saving and cost effective and provides remarkable results.

The performance of the artificial neural network was demonstrated in the study [14]. The authors presented an artificial neural network (ANN) model predicting values of sodium adsorption ratio (SAR), residual sodium carbonate (RSC), magnesium adsorption ratio (MAR), kellys ratio (KR) and percent sodium (% Na) in the groundwater of India. A neural network model consisting 13 input neurons, 7 hidden neurons and 5 output variables was used for computing ground water suitability for irrigation in study area. The SAR, MAR and KR values increases due to excessive use of chemical fertilizers. The RSC and % Na were found to be more precise as compared to SAR, MAR and KR values. The proposed model is developed through R programming and compared with MS - Excel software, gave satisfactory fit to the experimentally obtained dataset in 50 observation wells. The spatial distributions maps of measured and predicted values of irrigation indices were prepared using ARCGIS software. The

results confirm that the ANN model is an applied tool to predict the groundwater suitability for irrigation purpose in India.

M. B. Kia, S. Pirasteh, B. Pradhan, A. R. Mahmud, W. N. A. Sulaiman and A. Moradi [15] investigated the integration of GIS and neural network techniques in the field of water resource which opened various new approaches in hydrological modeling, improved our ability to create more accurate flood models, and helped to present the results in a spatial environment. Floods are affected by several factors such as rainfall, initial soil moisture, geology, land use, evaporation, watershed infiltration, geomorphology. There exists a very complicated relationship between these factors and the interaction between them are necessary for hydrological modeling. The specific objective of this article is to develop a flood model causative factors using ANN techniques and geographic information system (GIS) to modeling and simulate flood-prone areas in the southern part of Peninsular Malaysia. The ANN model was developed in MATLAB using seven flood causative factors. Relevant thematic layers (including rainfall, slope, elevation, flow accumulation, soil, land use and geology) are generated using GIS, remote sensing data and field surveys. In the context of objective weight assignments, the ANN is used to directly produce water levels and then the flood map is constructed in GIS. The verification results showed satisfactory agreement between the predicted and the real hydrological records. Also the results presented in this study could be used to help local and national government plan for the future and develop appropriate new infrastructure to protect the lives and property of the people of Johor, Malaysia.

3. Artificial neural networks in context of spatial data mining

The term *neural network* has its origins in attempts to find mathematical representations of information processing in the study of natural neural systems [16]. A neural network is a system composed of many simple processing elements operating in parallel whose function is determined by network structure, connection strengths and the processing performed at computing elements or nodes [17].

The term has been very broadly used to include a wide range of different model structures, many of which have been the subject of exaggerated claims to mimic neurobiological reality. As rich as neural networks are, they still ignore a host of biologically relevant features. From the perspective of applications in spatial data analysis, neurobiological realism is not necessary. In contrast, it would impose entirely unnecessary constraints [16].

Artificial Neural Network is a data refining and processing prototype, stimulated via the manner organic apprehensive structures, inclusive of the mind, process information. The important thing of this paradigm is the unconventional structure of the information processing systems. They can give reasonable answers for issues, which are for the most part described by non-linear ties, high dimensional, noisy, complex, loose, and imperfect or mistake inclined sensor information [20]. Similar to biological neural networks, it is a system composed of many interconnected neurons (nerve cells), which receive process and transfer information. After reaching a certain threshold, nerve cells are activated and forward the information to other connected neurons. During a learning process the interconnections are adapted. The simulation

of these biochemical processes in an ANN is realized by artificial neurons. The connections are realized by directed interconnection weights. Artificial neural networks are usually organized in layers. The network topology describes the number of layers, the number of neurons in layers and the way of their interconnection. Important parameters are the direction of signal propagation (forward/backward) and the type and level of connection (completely connected/with shortcuts) [19].

The Artificial Neural Network can be classified based on the learning paradigm, architecture and function. In case of supervised learning, the ANN is supplied with a sequence of both input data and desired output data. In the ANN methodology, the sample data is often subdivided into training, validation and test sets. In unsupervised learning, the training scheme only consists of input data. The ANN discovers some of the properties of the dataset and learns to reflect these properties in its output. In spatial mapping obtaining labeled datasets is expensive and sometimes impossible. A solution is the development of a new learning paradigm called semisupervised learning [18].

Artificial Neural Networks can be used to address a wide variety of real-world problems. They have the ability to learn from experience in order to improve their performance and dynamically adapt themselves to changes in the environment. In addition, they are able to deal with fuzzy or incomplete information and noisy data and can be very effective, especially in situations where it is not possible to define the rules or steps that lead to the solution of a problem. Hence they are fault tolerant. In addition, the ANN information - processing model is inherently parallel. Today artificial neural networks are used in a variety of disciplines including engineering, finance, artificial perception, control and simulation. The wide use of ANN in geospatial science stem from their roles in spatial data processing and analysis. Satellites orbiting and imaging the Earth produce massive amounts of geospatial data, on the order of tera-to peta-bytes. ANN, programmed on parallel neurally inspired hardware architectures, can analyze and classify this vast amount of data quickly and draw meaningful insights via mapping or modeling. ANN's generalization capability in dealing with classification across multiple spatial scales and resolutions is a significant advantage [18].

Moreover, many physical processes modeled in the geospatial sciences require accurate knowledge of process dynamics and such knowledge is often unknown. In this case, ANN can be used in function approximation. The learning that occurs in ANN is not affected by the integration of multisource data. The learning process is robust and fault tolerant. Never learning paradigms such as semisupervised learning or self-supervised learning can deal with incomplete data, overcoming the difficulties and expenses involved in gathering training labels for geospatial sensor data. Recent studies have attempted to make ANN more spatially explicit, either by introducing fundamental spatial principles such as spatial autocorrelation directly into the neural network structure or during post-processing and labeling using spatial neighborhood relationships. When using an ensemble approach, ANN can assist in characterizing the spatial heterogeneity of the Earth's surface and spatial uncertainty in labeling [18].

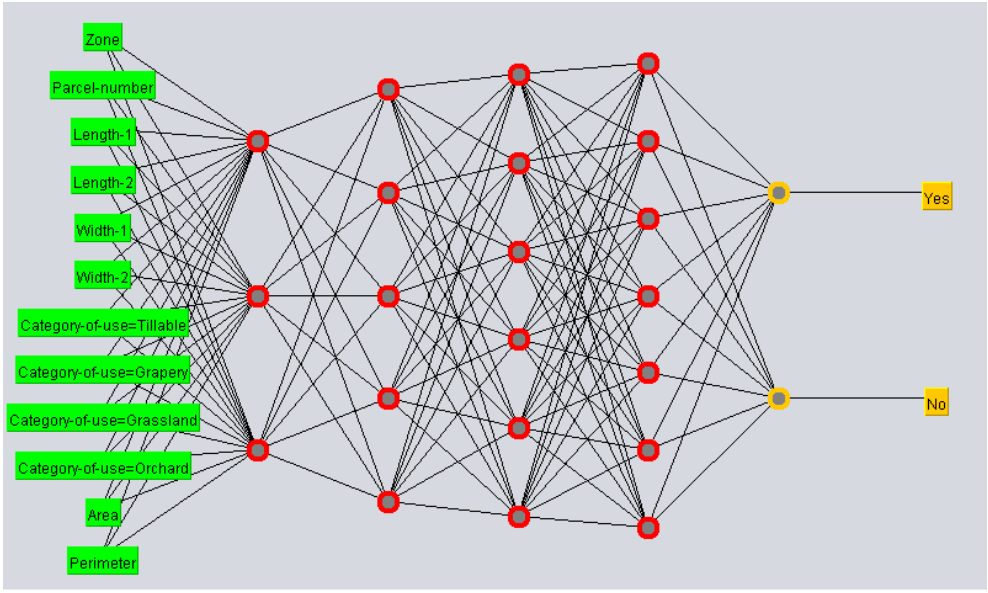

FIGURE 1. Artificial Neural Networks Pictorial Representation

Figure 1 presents an example of neural network diagram for Model I in the current research. The input, hidden and output variables are represented by nodes and the arrow denotes the direction of information flow through the network during forward propagation. For our figure the input layer contains 12 nodes which are represented by the attributes of the model: Zone, Parcel-number, Length-1, Length-2, Width-1, Width-2, Category-of-use with the four values: *Tillable*, *Grapery*, *Grassland*, *Orchard*, Area and Perimeter. The four hidden layers are formed by 3, 5, 6 and 7 nodes. The output layer contains the 2 output values of the nominal attribute Cadastral-number in the same model, attribute according to which the prediction was made.

4. The Artificial Neural Networks models based on GIS data

Extracting implicit information from geographical databases appears, in comparison to traditional non-spatial databases, to be more challenging. Together with non-spatial attributes, spatially referenced objects also carry information concerning their representation in space by geometrical and topological properties. Topology covers the geographical properties which are not closely connected to the actual position of objects, it represents the spatial relationships among objects. The topology is a branch of geometry that deals with those properties of a figure (object) that remain unchanged even when the figure is transformed. On the other side, geometric characteristics of data concerns information related to the actual location of the object in space. The location is usually described by Euclidian coordinates or Latitude and Longitude. Besides the core spatial characteristics dealing with geometry and topology, geographical data also contains information about the behavior of a phenomenon the data represents [21].

GIS (Geographic Information System) has advanced as a new emerging field as enhancement of communication technologies. Today's GIS has become indispensable and used in multidisciplinary areas to access information with respect to position. Enormous amount of GIS data is collected in numerical, text, graphics and analogues forms from satellite imagery sensors and other devices which represent the spatial and temporal situation [3].

GIS represents how the spatial entities are in the real world, via binary digits in the form of zeros and ones to approach them. A spatial entity may be interpreted as the cases, states, processes and phenomena in the real world or the natural and artificial objects with geometric features of points, lines, areas and volume. The attribute is drawn from the facts that can be known about a location. The valued attribute may be quantitative or qualitative and attribute data are the qualitative or quantitative description of points, lines, areas, polygons and cube with the forms of vector and raster [22].

A GIS data model mainly includes an object model and a field model. When they are used to describe a spatial entity with spatial data, the object model is for a discrete entity with vector data, and the field model is for a continuous entity with raster data. The object model assumes that the spatial entities may be precisely described via points with the exactly known coordinates, lines linking a series of crisply known points and areas bounded by sharply defined lines. The field model depicts spatial entities via giving each unit field an attribute value, instead of extracting objects or describing their topological relationships and it is more suitable to fuzzy, ambiguous spatial entity. A field model is in contrast to an object model. When it is used to study the attribute uncertainty, the field model has different characteristics from the object model. Many geographical phenomena, air pollution, population distribution, are studied as fields. When the images are used to investigate environments and resources or produce thematic maps, in GIS, a field model is suitable to discuss the attribute uncertainty than an object model [22].

4.1. The Datasets - Model I and Model II. In this paper the models with which we conducted the research include the description of two parcels plans of a sectoral index of a specific area of Romania in Olt, in the field of cadaster. The spatial data are real and have been processed with AutoCAD software, a software used mainly in topographic engineering for GIS data transformation, based on GPS (Global Positioning Systems) technology. The data that we used in the research are represented as vector data type and the datasets that represent the models were called *Model I* (*Cadaster_Romania_Parcel_I*) and *Model II* (*Cadaster_Romania_Parcel_II*).

The Model I includes 67 instances representing the parcels from the first parcel plan, the data definition was made from west to east thus respecting the parcel plan orientation from the area index.

The Model II includes 115 instances representing the parcels from the second parcel plan, the data definition was made from north to south also respecting the parcel plan orientation from the area index.

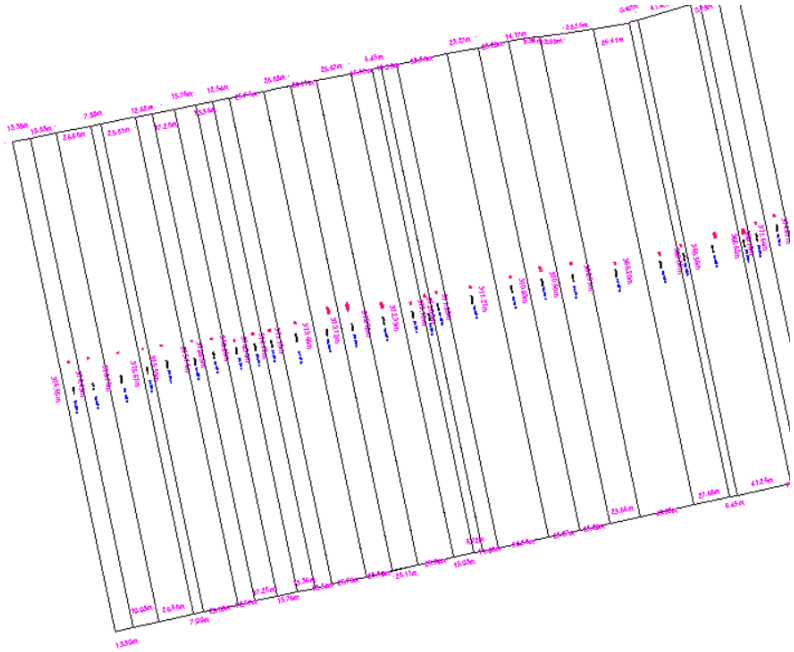

FIGURE 2. Examples of parcels in Model I

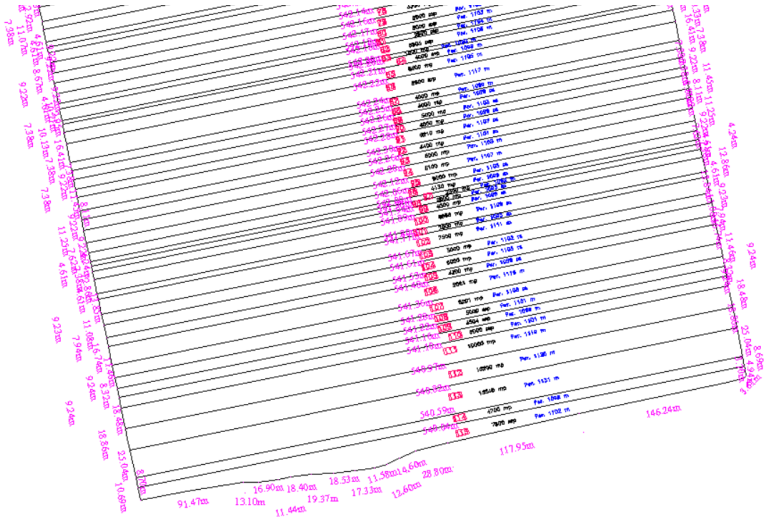

FIGURE 3. Examples of parcels in Model II

In the Figures 2 and 3 we can visualize the representation of the parcels from the two research models. Practically each parcel in a parcel plan represents a delimited

portion of land (land parcel), usually as a vector is represented by a geometric figure (object), used for a particular purpose. Also the two figures reflect the attributes of each parcel and the main spatial relationship that is on the topological type with touches.

In this paper each parcel or instance is defined by the ten attributes, namely:

(1) *Zone* = defined in the models as direct values, represents the area of the parcels plans - Olt;
(2) *Parcel-number* = defined numeric attribute in the models, represents each parcel in the parcel plan;
(3) *Length-1* = defined by the numerical type in the two models and expressed on the pattern in meters.
Length-1 represents the common attribute of two parcels. The representation of the length-1 for each parcel in Model I was done from west to east and in Model II was done from north to south.
(4) *Length-2* = this numeric attribute represents the second length from the geometric figure in the models, expressed also in meters.
(5) *Width-1* = defined numerical type. In Model I represents the line in the south part and in Model II the line from west, expressed in meters.
(6) *Width-2* = this numeric attribute represents the second length from the geometric figure in the models. It is the line in the north part in Model I and the line from west in Model II, also expressed in meters.
(7) *Category-of-use* = the nominal attribute in the models was equal to one of the values: *Tillable*, *Grapery*, *Grassland* and *Orchard*. It represents the category of use of each parcel in both plans.
(8) *Area* = defined numeric attribute in the models, each parcel has an area defined on the pattern in square meters.
(9) *Perimeter* = declared a numeric attribute of each parcel in the two models, expressed in meters.
(10) *Cadastral-number* = declared a nominal attribute and it represents the way to view those parcels are/or not a cadastral number received in the cadastral database.
Actually the cadaster branch is part of geodetic engineering and it is based on another branch of this specialization, namely topographic engineering with land measurements with which we can define the properties of geographical objects in a particular area.
Thus, through the cadaster, each property represented by a geographical object is registered in a cadastral database at the area level, thus leading to the fulfillment of some legal actions imposed by the laws in Romania.

The parcels classification from the two presented models was possible using the artificial neural networks thus defining based on the obtained results their performance in the field of spatial data mining. The Weka files used in the current application was called *Model I.arff* and *Model II.arff*.

5. Experiments and results

In this section we present two tables with results depending on the values of each training dataset from the two models. The tables contain a part of the results at

the prediction of the Artificial Neural Networks using Weka tool according with two random attributes from the models: *Cadastral-number* and *Category-of-use*. Also in this experimental part of this paper we expose in figures the Receiver Operator Characteristic (ROC) curve.

Model I	**Artificial Neural Networks**	
	Category-of-use	**Cadastral-number**
K statistic	0.6968	0.8804
Mean absolute error	0.1722	0.1468
Root mean squared error	0.2698	0.2558
Relative absolute error	50.2773 %	29.4122 %
Root relative squared error	65.3751 %	51.2041 %
TP Rate	0.806	0.940
FP Rate	0.124	0.060
Precision	0.724	0.940
Recall	0.806	0.940
F-Measure	0.759	0.940
MCC	0.691	0.880
ROC Area	0.952	0.955
PRC Area	0.896	0.943

TABLE 1. Test mode=Dataset validation-67 instances for the Artificial Neural Networks

The Table 1, for the dataset validation with 67 instances and for the attributes *Category-of-use* and *Cadastral-number*, contains the results of the implementation of the Artificial Neural Networks regarding to the accuracy of the classes defined by the accuracy measures of the classifier with *True Positives (TP) Rate* values, *False Positives (FP) Rate* values, *Precision* values, *Recall* values, *F-Measure* values, *Matthews Correlation Coefficient (MCC)* values, *Receiver Operating Characteristics (ROC) Area* and *Precision Recall (PRC) Area* values, evaluation on training set with statistical data through the *Kappa statistic* values and the predictor error measures with *Mean absolute error* values, *Root mean squared error* values, *Relative absolute error* values and *Root relative squared error* values.

The Table 2, for the dataset validation with 115 instances and for the attributes *Category-of-use* and *Cadastral-number*, contains the results of the implementation of the Artificial Neural Networks regarding to the accuracy of the classes defined by the accuracy measures of the classifier with *True Positives (TP) Rate* values, *False Positives (FP) Rate* values, *Precision* values, *Recall* values, *F-Measure* values, *Matthews Correlation Coefficient (MCC)* values, *Receiver Operating Characteristics (ROC) Area* and *Precision Recall (PRC) Area* values, evaluation on training set with statistical data through the *Kappa statistic* values and the predictor error measures with *Mean absolute error* values, *Root mean squared error* values, *Relative absolute error* values and *Root relative squared error* values.

Model II	**Artificial Neural Networks**	
	Category-of-use	**Cadastral-number**
K statistic	0.4588	0.4367
Mean absolute error	0.2145	0.3499
Root mean squared error	0.3252	0.4051
Relative absolute error	67.5225 %	73.9929 %
Root relative squared error	81.8337 %	83.3481 %
TP Rate	0.678	0.730
FP Rate	0.228	0.228
Precision	0.682	0.734
Recall	0.678	0.730
F-Measure	0.660	0.732
MCC	0.480	0.437
ROC Area	0.849	0.826
PRC Area	0.721	0.828

TABLE 2. Test mode=Dataset validation-115 instances for the Artificial Neural Networks

Making an analysis of the statistical data from both tables we can see that the values of the parameter *Relative absolute error* for the two attributes *Category-of-use* and *Cadastral-number* are values below the 100% threshold which means that the models have a very good score. Also in case of accuracy values demonstrated by the *ROC Area* parameter indicate the same fact. In Table 1 for Model 1 with the 67 instances the *Category-of-use* attribute has less good values than the *Cadastral-number* attribute and in Table 2 for Model 2 with the 115 instances the *Category-of-use* attribute has better values compared to the *Cadastral-number* attribute.

Thus, in the case of Artificial Neural Networks analyzing the results of the two models we can say that much better statistical values are demonstrated on a smaller number of instances, these values increasing up to the 100% threshold on a larger number of instances, thus leading the model to a score not too well. The results are demonstrated in the same way and by accuracy.

In the experimental section we also present another way to evaluate the accuracy of the models. It is represented in the Figure 4 and Figure 5 which illustrates the measuring of the area under Receiver Operator Characteristic (ROC) curve (AUC).

One of the most important values output by Weka is represented by the ROC area measurement. An optimal classifier will have ROC area values approaching 1, with 0.5 being similar to a Kappa statistic value of 0.

To make the charts we used the dataset of the each model with the class *Tillable* belonging to the attribute *Category-of-use*.

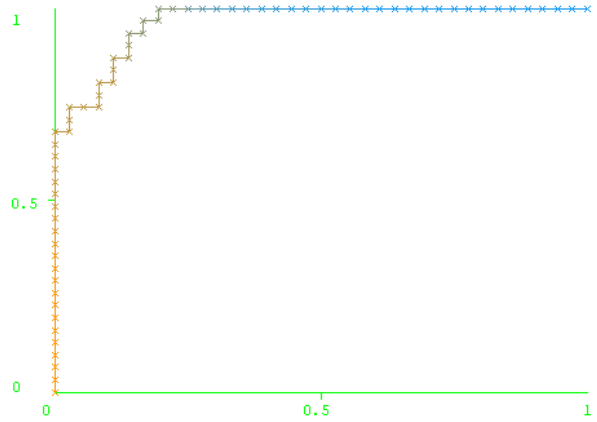

FIGURE 4. Artificial Neural Networks representation under ROC area

Applying Artificial Neural Networks on a dataset with 67 instances belonging to Model I onto the class value *Tillable* we obtained results as the graphical representations in Figure 4 for the model's evaluation. In this case the area under ROC is equal with 0.965 value, a very close to 1 value, and in the graphical representation the Plot Area under ROC appears above the diagonal line, from which it results that in the case of the model with fewer entities presented in this research the classifier based on the ANN is an excellent one.

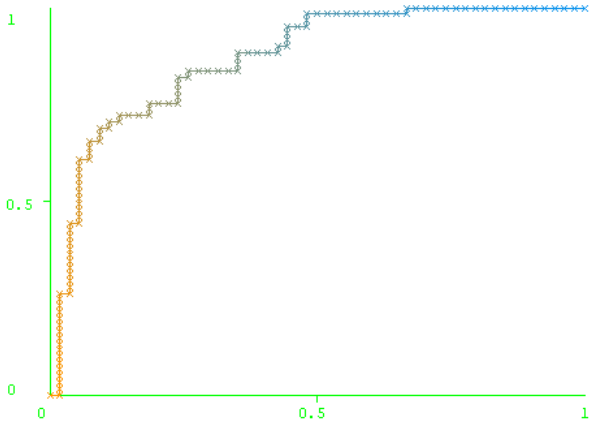

FIGURE 5. Artificial Neural Networks representation under ROC area

Applying Artificial Neural Networks on a dataset with 115 instances belonging to Model II onto the class value *Tillable* we obtained results as the graphical representations in Figure 5 for the model's evaluation. In this case the area under ROC is equal with 0.870 value, a value close to 1, and in the graphical representation the Plot Area under ROC appears above the diagonal line, from which it results that in the case of

the model with several entities presented in this research the classifier based on the ANN is a very good one.

In the case of the first model for the *Category-of-use* attribute the results obtained for the four classes: *Tillable*, *Grapery*, *Grassland* and *Orchard* are in generally excellent, a fact demonstrated by the average value for the ROC Area equal with 0.952 in the Table 1. In the case of the second model for the *Category-of-use* attribute, the results obtained for class *Tillable*, class *Grapery*, class *Grassland* and class *Orchard* are in generally very good, the average value for the ROC Area equal to 0.849 in Table 2, demonstrates this.

Regarding the values obtained for the ROC area, we can highlight the fact that, in addition to the high accuracy and very good results demonstrated by the classifier in this research, we must take into account the significant difference between the results based on the size of the spatial datasets used in the two models.

Thus, based on these results, we can affirm that the major challenges for the spatial data mining domain are represented both by the management of very large datasets and by the complex structure of spatial data, fact which involves efficient and high-performance algorithms as well as efficient visualization approaches for presenting the complex models.

6. Conclusions and future work

The current paper exposes the concepts of spatial data mining, spatial classification and also focuses on the presentation of the Artificial Neural Networks. However, the main purpose of this research is to demonstrate the Artificial Neural Networks performance and efficiency in context of the spatial data mining domain.

Starting from the statistical results and those regarding the accuracy that we obtained in the two tables and in the two charts, after a rigorous analysis, new research can be developed in the future with efficient practical applications of data mining algorithms in the geodetic engineering domain, meaning the cadaster or the topographic engineering. Based on the results from the experimental part of Section 5 we can define some conclusions regarding Artificial Neural Networks in the context of models based on datasets built with GIS data.

Artificial Neural Networks admit a multivariate data analysis of complex problems. They learn from given examples without an explicit programming of problem solution what leads to a simultaneous application. Other important advantages of artificial neural networks are represented by the generation of a mark by using given examples and also by their ability to generalize. Although Artificial Neural Networks have a complex and successful representation in spatial events we can say that in the case of diverse and much larger spatial data the results are not surprising. However, making a comparison with other data mining algorithms, mentioning here the K-nearest neighbor algorithm, the Naïve Bayes algorithm, the Decision tree (C4.5) algorithm, we can affirm that the Artificial Neural Networks are a safe and a successful choice in the spatial data mining domain.

The analysis of this part of the research, presented in this paper, shows that the spatial data mining domain is still an offering domain with very good research results and its potential is still being developed. As further directions, we will approach deep

learning models based on GIS data or we will discuss the role of rough sets in the context of spatial data mining.

References

[1] P. Kuba, Data Structures for Spatial Data Mining, *FIMU Report Series* (2001). Retrieved at https://www.fi.muni.cz/reports/files/2001/FIMU-RS-2001-05.pdf

[2] M. Hemalatha and N.N. Saranya, A Recent Survey on Knowledge Discovery in Spatial Data Mining, *International Journal of Computers Science (IJCSI)* **8** (2011), no. 2, 473–479.

[3] H. Goyal, C. Sharma, and N. Joshi, An Integrated Approach of GIS and Spatial Data Mining in Big Data, *International Journal of Computer Application* **169** (2017), no. 11, 1–6.

[4] D. Li and S. Wang, Concepts, Principles and Applications of Spatial Data Mining and Knowledge Discovery, *ISSTM 2005*, August 27-29, 2005, 1–13.

[5] D. Guo and J. Mennis, Spatial data mining and geographic knowledge discovery-An indroduction, *Computers, Environment and Urban Systems* **33** (2009), 403–408.

[6] V.R. Kanagavalli and K. Raja, A Study on Application of Spatial Data Mining Techniques for Rural Progress, *Proceedings of International Conference ICICT09* (2009), arXiv:1303.0447 [cs.DB].

[7] P.K.S.C. Jayasinghe and M. Joshida, GIS-Based Neural Network Modeling to Predict Suitable Area for Beetroot in Sri Lanka: Towards Sustainable Agriculture, *Journal of Developments in Sustainable Agriculture* **4** (2009), 165–172.

[8] Y. Kashyap, A. Bansal, and A.K. Sao, Spatial Approach of Artificial Neural Network for Solar Radiation Forecasting: Modeling Issues, *Journal of Solar Energy* **2015**, Article ID 410684, 1–13.

[9] Y. Liu and H. Lu, Outlier Detection Algorithm based on SOM Neural Network for Spatial Series Dataset, *2018 Tenth International Conference on Advanced Computational Intelligence (ICACI)*, March 29-31, 2018, Xiamen, China, 162–168.

[10] L. Akil and H.A. Ahmad, Salmonella infections modelling in Mississippi using neural network and geographical information system (GIS), *BMJ Open* (2016), 1–9.

[11] A. Mollalo, L. Mao, P. Rashidi, and G.E. Glass, A GIS - Based Artificial Neural Network Model for Spatial Distribution of Tuberculosis across the Continental United States, *International Journal of Environmental Research and Public Health* **16** (2019), 1–17.

[12] S.N. Rajan and A.K. Sinha, Modeling the influence of socio-economic factors on HIV prevalence in India using Artificial Neural Network on spatial database, *International Journal of Engineering Research & Technology* **1** (2012), 1–7.

[13] M.B. Kia, S. Pirasteh, B. Pradhan, A.R. Mahmud, W.N.A. Sulaiman, and A. Moradi, An artificial neural network model for flood simulation using GIS: Johor River Basin, Malaysia, *Environmental Earth Sciences* **67** (2012), no. 5, 251–264.

[14] T. Behrens, H. Förster, T. Scholten, U. Steinrücken, E.-D. Spies, and M. Goldschmitt, Digital soil mapping using srtificial neural networks, *J. Plant Nutr. Soil Sci.* **168** (2005), 1–13.

[15] V.M. Wagh, D.B. Panaskar, A.A. Muley, S.V. Mukate, Y.P. Lolage, and M.L. Aamalawar, Prediction of groundwater suitability for irrigation using artificial neural network model: a case study of Nanded tehsil, Maharashtra, India, *Model Earth Syst. Environ.* **2** (2016), 1–10.

[16] M.M. Ficher, Neural Network for Spatial Data Analysis, In: *The SAGE Handbook of Spatial Analysis* (eds. Edited by: A.S. Fotheringham, P.A. Rogerson) (2009), 340–360.

[17] S. Gopal, Unit 188 - Artificial Neural Networks for Spatial Data Analysis. In *Core Curriculum in Geographic Information Science*, UC Santa Barbara: National Center for Geographic Information and Analysis, (2000). Retrieved from https://escholarship.org/uc/item/0c90m7qd

[18] S. Gopal, Artificial Neural Networks in Geospatial Analysis, In: *International Encyclopedia of Geography: People, the Earth, Environment and Technology* (eds. D. Richardson, N. Castree, M.F. Goodchild, A. Kobayashi, W. Liu and R.A. Marston) (2016), 1–7. https://doi.org/10.1002/9781118786352.wbieg0322

[19] S. Noack, A. Knobloch, S. H. Etzold, A. Barth, and E. Kallmeier, Spatial Predictive Mapping using Artificial Neural Networks, *ISPRS Technical Commission II Symposium*, 6-8 October, Toronto, Canada, In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2014, 1–8.

[20] M. Gangappa, C. Mai, and P. Sammulal, Techniques for Machine Learning based Spatial Data Analysis: Research Directions, *International Journal of Computer Applications* **170** (2017), no. 1, 9–13.
[21] D. Muralir, G.H. Shankar and R.S. Kumar, A Survey of Spatial Data Mining Approaches: Algorithms and Architecture, *International Journal of Engineering Trends and Technology*, Special Issue April 2017, 80–86. http://www.ijettjournal.org/Special
[22] S. Wang, W. Shi, H. Yuan, and G. Chen, Attribute Uncertainty in GIS Data, *Springer-Verlag*, 2005, 614–623.

(Dana Mihai) Department of Computers and Information Technologies, Faculty of Automation, Computers and Electronics, University of Craiova, 107 Bvd.Decebal, Craiova, 200440, Romania
E-mail address: dana.mihai@edu.ucv.ro