

Processing GIS Data Using Decision Trees and an Inductive Learning Method

Dana Mihai and Mihai Mocanu

Abstract—This paper extends recent work on spatial data mining, with another application of the classification techniques, namely with the Decision tree classifier algorithm. Spatial data mining represents a various and investigated domain because huge amounts of spatial data have been collected, ranging from remote sensing to geographical information system and computer cartography. In this work we used the Weka tool to implement the C4.5 (Quinlan) Decision tree algorithm on a dataset of Geographic Information System (GIS), data collection called Cadastre formed by a parcel plan from the Dolj district of Romania. The results of the experiments highlight several advantages and also some disadvantages of Decision tree in context of spatial data mining, with a favorable accuracy.

Index Terms—Algorithm, classification, decision tree, C4.5, Weka.

I. INTRODUCTION

Spatial data mining (SDM) or knowledge discovery in spatial databases refers to the extraction of implicit knowledge or other patterns that are not explicitly stored in spatial database. The word spatial refers to the data associated with the geographic location of the earth. A large amount of spatial data have been collected in various applications, ranging from remote sensing to GIS, computer cartography, environmental assessment and planning [1].

The quantity of the collected data is huge, meaning that it is too much human knowledge to analyze it, new and efficient methods are needed to discover knowledge from large spatial databases. Over the last few years, spatial data mining has been often used in many and various applications. Spatial data mining is the analysis of geometric or statistical characteristics and relationships of spatial data [1].

Statistical spatial data mining analysis has been a popular approach to analyzing spatial data and exploring geographic information. Spatial data mining will further develop spatial statistical analysis methods and extend them for huge amounts of spatial data, with more emphasis on efficiency, scalability, cooperation with database and data warehouse systems, improved user interaction and the discovery of new types of knowledge [2].

Spatial classification is the process of finding a set of rules to determine the class of spatial object. Spatial classification methods extend the general-purpose classification methods to consider not only attributes of the object to be classified but also the attributes of neighboring objects and their spatial

relations. The spatial classification techniques such as Decision tree (C4.5) or Artificial Neural Networks (ANN), help to find the group of spatial objects together and to prevent overfitting, how to handle missing attribute values, and other implementation issues [1].

The study in this paper focuses to present the statistical results obtained by applying the Decision tree with C4.5 algorithm on the spatial dataset. On the other hand the results presented in this work can be compared with another statistical research in the field in order to develop or to improve the classification techniques based on a model in the spatial data mining.

The article is organized as follows: in Section I we expose the spatial data mining domain with the spatial classification, in Section II some related work is shown, in Section III we present the C4.5 Decision tree algorithm, Section IV describes the spatial data classification with the model based on GIS data, in Section V which shows the steps of experimental results we analyze and present the results performed and in Section VI we conclude our discussion and also we present further directions.

II. RELATED WORK

In the specialized literature there are many approaches that have studied both the spatial data mining domain and the Decision tree classification technique with C4.5 algorithm.

In paper [3] author M. Vignesh shows that spatial data mining is a promising field of research with wide applications in GIS, medical imaging, robot motion planning. Although, the field is quite recent, a number of algorithms and techniques have been proposed to discover various kinds of knowledge from spatial data. Also this work surveys existing methods for spatial data mining and mentioned their strengths and weaknesses.

G. Zhou, L. Wang, D. Wang, and S. Reichle [4], in their work present a thorough research undertaken to explore the applicability of data mining and knowledge discovery in combination with Geographic Information Systems (GIS) technology to pavement management to better decide maintenance strategies, set rehabilitation priorities, and make in make investment decisions. The main objective of the research is to utilize data mining techniques to find some useful knowledge hidden in the pavement database. Decision trees and association rules, has been used in the analysis.

J. D. Jadhav and H. P. Channe [5] used in the comparative study three of the classification techniques: K-Nearest Neighbor, Naïve Bayes and Decision Trees. The paper focuses on study of various advantages and disadvantages of each algorithm. This study was based on different datasets:

Manuscript received July 25, 2020; revised February 3, 2021.

Both authors are with the University of Craiova, Romania (e-mail: mihai_danam@yahoo.com, mmocanu@software.ucv.ro).

Weather Nominal with 14 instances (small dataset), *Segment Challenge* with 1500 instances (medium dataset) and *Supermarket* with 4627 instances (large dataset). Thus, Decision Trees are very simple and fast. It produces the accurate result, it supports incremental learning and it takes the less memory. The representation is easy to understand and it can also deal with noisy data. This classification technique uses different measures such as Entropy, Gini index, Information gain to find best split attribute. It performs well on large dataset and it has a problem of over fitting.

M. Karim and R. M. Rahman [6] investigated two data mining techniques: the Naïve Bayes and the C4.5 decision tree algorithms. The goal of their work was to predict whether a client will subscribe a term deposit. Also they made comparative study of performance of those two algorithms. Public available UCI data was used to train and test the performance of the algorithm, using WEKA tool.

The study of K. Ganesan [7] proposes a framework with preprocessing techniques namely missing value replacement, discretization, the principal component analysis to extract the key features and then applying C4.5 classifier algorithm to enhance the classification of medicinal data. The C4.5 algorithm was used to construct a decision tree based on the information gained of the training set. The dataset was composed of details of about 303 patients and helps to predict presence or absence of cardio vascular disorder based on 75 attributes. The proposed framework was applied on the dataset and exhibited an accuracy of about 77.73%.

The authors N. Settouti, M. E. A. Bechar, and M. A. Chikh [8] proposed in the statistical comparison study of the 10 top data mining algorithms (including the C4.5 decision trees algorithm) the problem of determining the most suitable classifier to solve a given problem of classification. The choice of the classifier is guided by operational constraints and after that the classifier is configured through a learning basis, the rate of generalization of the classifier or accuracy which is the criterion characterizing its performance. The datasets used in this study are Twelve Medical and Biological datasets mainly selected from UCI Machine Learning Repository. The aim of this statistical comparison is to use the intelligent methods to perform especially on the medical field for detection, diagnosis and classification of certain diseases in a very short time.

N. Dogan and Z. Tanrikulu [9] compare the classification of algorithm accuracies, speed and robustness for various datasets and their implementation techniques with Weka. This work first discusses the application of certain classification models on multiple datasets. The accuracies and the speed of the results are then compared. The datasets were collected from UCI Machine Learning Repository and they are: Acute, Breast Cancer, CPU, Credits, Iris, Letters, Pittsburg, Red Wine, Segment, Wine All and White Wine. Fourteen classification algorithms representing the different types of classification models (decision trees, neural networks, immune systems, probabilistic models, etc.) were selected from the many existing classification algorithms that fell within the scope of this study. The C4.5 algorithm is presented as a good performer classifier.

The performance of the C4.5 algorithm is presented by S. Oujdi and H. Belbachir in their work [10]. They propose an

extension of the C4.5 algorithm for spatial data, based on two different approaches *join materialization* - favors the processing time in spite of the memory space and *querying on the fly the different tables* - promotes memory space despite the processing time. The modified C4.5 algorithm requires three entry tables: a target table, a neighbor table, and a spatial index join that contains the possible spatial relationship among the objects in the target table and those in the neighbor table. The algorithm is applied to a spatial data pattern in the accidentology domain with a higher performance.

S. Singh and P. Gupta explain in their paper [11] three most commonly used decision tree algorithms, to understand their use and scalability on different types of attributes and feature. The current study presents some advantages and disadvantages of each algorithm. Thus, the C4.5 algorithm allows pruning of the resulting decision trees. This increases the error rates on the training data, but more importantly, decreases the error rates on the unseen testing data. The classifier can also deal with numeric attributes, missing values and noisy data.

III. THE DECISION TREE

A decision tree is a popular classification technique that results in flow-chart-like hierarchical tree structure which is composed of three basic elements: decision nodes corresponding to attributes, edges or branches which correspond to the different possible attribute values. The third component is leaves including objects that typically belong to the same class or that are very similar. Such representation allows us to induce decision rules that will be used to classify new instances. In fact, each path from the root to a leaf corresponds to a conjunction of test attributes and the tree is considered as a disjunction of these conjunctions. Building procedure (induction) and classification procedure (inference) are the two major procedures with which the most decision trees are made [12].

A decision tree classifier appears thus as a recursive, partition-based tree model that predicts the class for each of the points in a d - dimensional space that belongs to the training set.

Decision tree learning algorithms such as ID3 or C4.5 are the most powerful and well known predictive methods for classification. We can also mention the CART algorithm and CHAID algorithm. They are predictive models and they are used in computing for calculating probabilities. The data usually play an important role in determining the quality of the decision trees [13].

Other related topics to building Decision Trees and learning algorithms are: simplifying decision trees and scalable implementation techniques.

Since our intention here is to follow closely a widely tested implementation (Weka), we will not get into details on these topics.

The C4.5 algorithm is an extension of the ID3 algorithm (Iterative Dichotomiser 3) developed by Quinlan (1986). It uses information theory and inductive learning method to construct decision tree. C4.5 improves ID3 which cannot process continuous numeric problem. J48 is an open source

Java implementation of the C4.5 algorithm in the WEKA data mining tool [14].

Decision tree based on the C4.5 algorithm is a commonly used classification technique which extract relevant relationship in the data. It is a program that creates a decision tree based on a set of labeled input data. The decision tree modeling created by this algorithm can be used for classification and for this reason the C4.5 algorithm is often defined as a statistical classifier. The C4.5 algorithm makes decision trees use a set of training data. The training data is a set $S = s_1, s_2, \dots$ of already classified samples. Each sample $S_i = x_1, x_2, \dots, x_p$, consists of a p -dimensional vector $X = (x_1, x_2, \dots, x_p)$, where the x_j represents attribute values or features of the sample. The training data is augmented with a vector $A = a_1, a_2, \dots$ where a_1, a_2, \dots represent the class to which sample belongs [13].

From the C4.5 classification algorithm, the decision tree is constructed, depending on the most effective attributes are given using the Entropy and the Gain information. Hence, to achieve this construction, it is necessary to calculate the entropy of each feature of the training images by using C4.5 algorithm and measure the information gained for each feature. The attribute with the highest normalized information gain is chosen to make the decision [14]. The C4.5 algorithm then recurses on the partitioned sublists. It is computationally efficient and has proven very successful in practice [15].

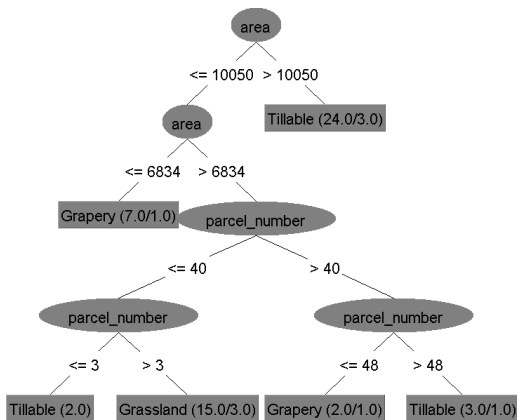


Fig. 1. An example of decision tree.

Fig. 1 shows an example of building a Decision tree using the C4.5 algorithm. This tree model is used to determine a category of use of a parcel in a certain area. The example predicts, depending on the size of the parcel and its number in the parcel plan, the destination of a parcel which means that a parcel can have the category of use Tillable, Grassland or Grapery.

The classification (or class identification) provides a logical description that yields the best partitioning of the entered dataset according to one or few attributes (label class). The classification rules constitute a decision tree where each node contains a criterion on an attribute. The leaves contain objects that belong-in majority-to one label class [16].

IV. SPATIAL DATA CLASSIFICATION

After an approach in [1] we define some details about the

spatial data classification and how to recognize these spatial data. GIS Systems collect data from various data sources from a wide range of communicating devices. These data from the communicating devices are available in different representations and file formats. In GIS these data representation is classified into two main categories: Raster and Vector Data types, presented in Fig. 2.

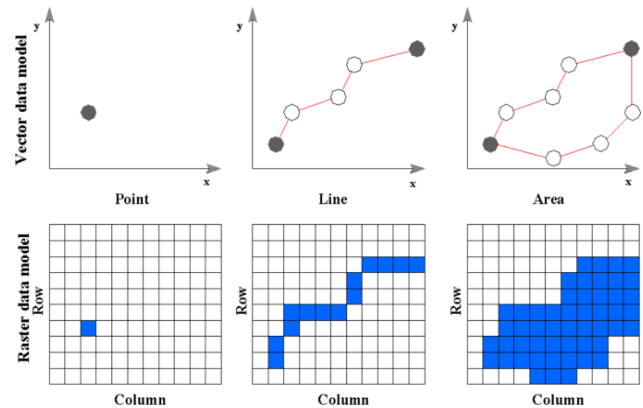


Fig. 2. The data representation in GIS.

Raster represents a two dimensional data type which stores the value of pixel colors of raster images in a cell and the attribute values are continuous in nature. This category is used to represent information from sources such as scanned maps, air photos and elevation layers.

Vector Data types are used to represent discrete features in GIS and have a layered architecture representing point, line and polygon. Vector Data types are used to represent information from sources such as countries, cities, provinces, districts, roads, with a layered hierarchy.

Analysis of spatial data has become important for analysing information with respect to location, shape, size and orientation about a physical object that can be represented by numerical values in a geographic coordinate system.

Spatial objects usually consist of both spatial and non-spatial data. Spatial data represent the data related to spatial description of the objects such as coordinates, area, latitudes, perimeters, spatial relations (distance, topology, direction). Non-spatial data are other data associated to spatial objects. The spatial object has been classified by using its attributes. Each classified object is assigned to a class.

The spatial attributes are classified in three major relations (functions or predicates) [17]:

- 1) Topological: overlaps, contains, touches, disjoint, equals, covers, covered-by, inside;

Topological relations are based on the boundaries, interiors and complements of the two related objects. This relation is always non-spatial data and it requires spatial mapping to convert non-spatial to spatial data.
- 2) Distance: close-to, far-away;

Distance relations compare the distance of two objects with a given constant using one of the arithmetic comparison operators.
- 3) Direction: north, south, east, west, northeast, northwest;

Directions relations consider one representative point of the *object1* as the origin of a virtual coordinate system whose quadrants and half-planes define the directions.

To realize the direction predicate, all point of *object2* have to be located in the respective area of the plane.

A. The Model Based on GIS Data

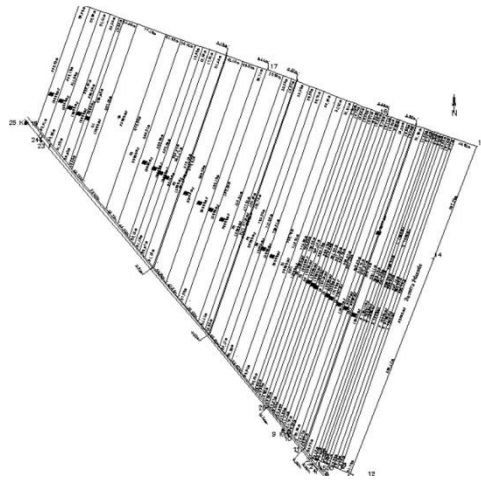


Fig. 3. The parcel plan with 53 instances.

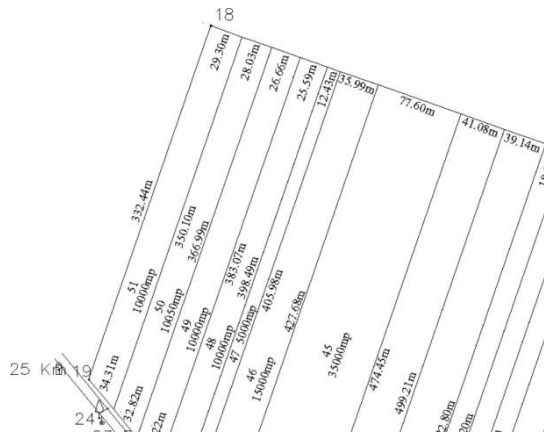


Fig. 4. Parcels with the attributes in the parcel plan.

The dataset that we used in order to run the C4.5 classifier algorithm is based on a model described in the paper [18]. The model is called *Cadastr* and includes 53 instances representing the parcels of a parcel plan from the Dolj district of Romania. The instances are presented in Fig. 3 as a parcel plan. Each instance is characterized by nine attributes, namely:

- 1) *Zone* = area, declared in the model as direct value.
- 2) *Parcel-number* = parcel number, declared numeric attribute.
- 3) *Length1*, *Length2*, *Width1* and *Width2* = represented the two lengths respectively the two widths of a parcel, have been defined by the numerical type and have been expressed on the pattern in meters.
The *Length1* represents the common element of two parcels, the representation of the data for each parcel in the model was done from west to east.
- 4) *Category-of-use* = the category of use of a parcel, attribute that was equal in model to one of the values: *Tillable*, *Grassland*, *Grapery*.
This attribute represents the way we can visualize the type of use of a parcel in the mentioned area.
- 5) *Area* = the area of a parcel, an attribute that we have defined of the numeric type and has been expressed on the pattern in square meters.

- 6) *Cadastral-number* = the cadastral number of a parcel, attribute that we have declared by direct values.
This attribute is the way to view those parcels that are/or not entered into the database as received/or not by a cadastral number.

The data that we used to create the model are the type GIS (Geographic Information System) and have been processed with the AutoCAD software and the Weka file used in the application was called *Date.arff*. In our case one single parcel or an instance is an object. The Fig. 4 shows the attributes of each parcel on the model and also we can see that the main spatial relation is on the topological type with touches.

V. EXPERIMENTAL RESULTS

In this section we present three tables, in which the results vary on the value of *k*. These tables expose a part of the results of the implementation of the C4.5 Decision tree algorithm using Weka tool, according to the attributes *Cadastral-number* and *Category-of-use* which represent two random values on the model.

TABLE I: TEST MODE=7-CROSS VALIDATION FOR THE DECISION TREE

k=7	Decision tree	
	Category-of-use	Cadastral-number
K statistic	0.4253	- 0.2019
Mean absolute error	0.2733	0.5309
Root mean squared error	0.4360	0.5890
Relative absolute error	67.6627 %	106.9505 %
Root relative squared error	97.2184 %	118.1346 %
Precision	0.683	0.400
Recall	0.660	0.415
F-Measure	0.661	0.404

The Table I, for *k=7* and for the attributes *Category-of-use* and *Cadastral-number*, contains a summary of the results of the implementation of the Decision tree about the accuracy defined by the classifier accuracy measures with *Precision* values, *Recall* values and *F-Measure* values, 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.

TABLE II: TEST MODE=12-CROSS VALIDATION FOR THE DECISION TREE

k=12	Decision tree	
	Category-of-use	Cadastral-number
K statistic	0.4985	- 0.1848
Mean absolute error	0.2451	0.5076
Root mean squared error	0.4110	0.5734
Relative absolute error	60.5813 %	102.2957 %
Root relative squared error	91.4956 %	115.0750 %
Precision	0.726	0.396
Recall	0.698	0.434
F-Measure	0.704	0.401

TABLE III: TEST MODE=25-CROSS VALIDATION FOR THE DECISION TREE

k=25	Decision tree	
	Category-of-use	Cadastral-number
K statistic	0.5006	- 0.4009
Mean absolute error	0.2445	0.5328
Root mean squared error	0.3957	0.5946
Relative absolute error	60.2073 %	107.2607 %
Root relative squared error	87.7599 %	119.2047 %
Precision	0.705	0.294
Recall	0.698	0.321
F-Measure	0.697	0.304

The Table II, for *k=12* and for the attributes

Category-of-use and *Cadastral-number*, contains a summary of the results of the implementation of the Decision tree about the accuracy defined by the classifier accuracy measures with *Precision* values, *Recall* values and *F-Measure* values, 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 III, for $k=25$ and for the attributes *Category-of-use* and *Cadastral-number*, contains a summary of the results of the implementation of the Decision tree about the accuracy defined by the classifier accuracy measures with *Precision* values, *Recall* values and *F-Measure* values, 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.

Analyzing the percent of errors in Tables I, II and III, we can observe major differences between the *Category-of-use* attribute and the *Cadastral-number* attribute.

In case *Category-of-use* we obtained values for parameter *Relative absolute error* below the 100% threshold, values that decrease with the increased value of the factor k . In the second case, for the *Cadastral-number* attribute, the values for parameter *Relative absolute error* are higher than 100% and generally increase with the value of factor k . Therefore, we can affirm in the case of the Decision tree for *Cadastral-number* attribute, the model does not score very well and the value over 100% is an indication of it.

In the experimental section we also present the Fig. 5 which illustrates the ROC curve. The name ROC stands for Receiver Operator Characteristic and comes from the signal detection theory that was developed during the Word War II for the analysis of radar images. The ROC chart shows the tradeoff between the false positive rate or specificity (horizontal axis) and the true positive rate or sensitivity (vertical axis) for a given model.

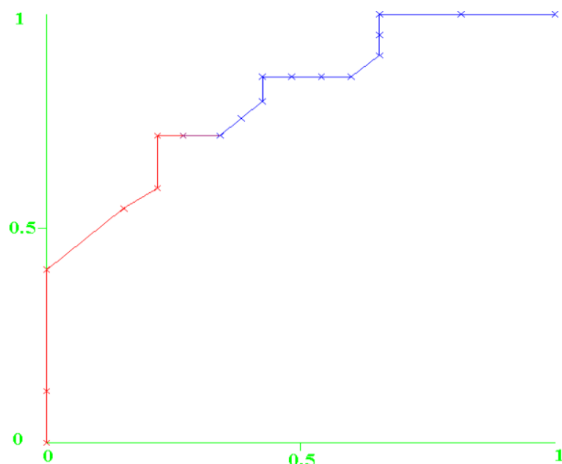


Fig. 5. Decision tree = ROC curve.

Another way to evaluate the accuracy of the model is represented by the measuring of the area under ROC curve (AUC). The closer the area is to 0.5, the less accurate the corresponding model is. A model with a perfect accuracy will have an area of 1.0.

To make the graphic we used the following dates: $k=25$ and the class *Tillable* belonging to the attribute

Category-of-use.

Applying Decision tree with a value of $k=25$ onto the class value *Tillable* we obtained results as the graphical representations in Fig. 5 for the model evaluation. In this case the area under ROC is equal to 0.8421 value, a value close to 1.0, and in the graphical representation the Plot Area under ROC appears above the diagonal line, from which it results that the Decision tree is a very good classifier for the model presented in this study.

In the paper [18] we compare the K-nearest neighbor algorithm and Naïve Bayes algorithm on the same model presented in this article, with the same values of k and the same attributes. From the point of view of accuracy, the area under ROC, for the K-nearest neighbor algorithm, is equal to 0.7100 and the area under ROC, for the Naïve Bayes algorithm, is equal to 0.6471. Thus, based on these results, we can define some comparative conclusions.

VI. CONCLUSIONS AND FUTURE WORK

This paper highlights the concepts of spatial data mining, spatial classification, statistical spatial data mining and also we focused on the presentation of the Decision tree with C4.5 algorithm.

The main purpose of this work is to demonstrate the efficiency and performance of the Decision tree in context of spatial data mining. Based on the statistical results obtained in the experimental part, after a rigorous analysis, new research can be developed in the future in order to apply in practice the data mining algorithms in the field of the topographic engineering.

Starting from the results of the three tables and of the chart we can define some advantages and disadvantages of the Decision tree classifier algorithm in the context of the model based on the dataset built using GIS data.

Thus, we can say that the Decision tree algorithm, in generally, presents a higher accuracy in comparison with K-nearest neighbor algorithm and Naïve Bayes algorithm on the same spatial dataset. For example, we can view the results of the area under ROC where the Decision tree classifier has the best result. Also it is an easier algorithm because it is represented in the form IF-THEN rules which means a better understanding for the readers and it is fast with a less error rate.

The Decision tree can have a more complex representation from some concepts due to the replication problem. In our case, comparing with the data results presented in the study on the K-nearest neighbor algorithm and Naïve Bayes algorithm, we can affirm that the Decision tree algorithm in case of the *Cadastral-number* attribute, shows lower classifier accuracy measures than the other two classification algorithms. Therefore, in the case of the spatial data mining it does not always have a good accuracy, depending also on the long training time.

The analysis of the spatial data mining presented in this work shows that the domain of spatial data mining is a promising field, with advantageous research results and many challenging issues. As further directions, we will apply the artificial neural networks on models based on spatial data mining or we will discuss the spatial clustering in the context

of hierarchical algorithms.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Dana Mihai proposed the research and the data model used in the paper; Dana Mihai analyzed the data, wrote the paper and concuted the research; Mihai Mocanu verified and coordinated the research; all authors had approved the final version.

REFERENCES

- [1] M. Perumal, B. Velumani, A. Sadhasivam, and K. Ramaswamy, "Spatial data mining approaches for GIS - A brief review," *Springer International Publishing Switzerland*, vol. 2, pp. 579-592, 2015.
- [2] J. Han and M. Kamber, *Data Mining - Concepts and Techniques*, 2nd ed., Elsevier, 2006.
- [3] M. Vignesh, "Spatial data mining - progress and challenges," *International Journal of Computer Science and Information Technology Research*, vol. 2, pp. 1-16, 2014.
- [4] G. Zhou, L. Wang, D. Wang, and S. Reichle, "Integration of GIS and data mining technology to enhance the pavement management decision making," *Journal of Transportation Engineering*, vol. 136, pp. 324-341, April 2010.
- [5] S. D. Jadhav and H. P. Channe, "Comparative study of K-NN, Naive Bayes and decision tree classification techniques," *International Journal of Science and Research (IJSR)*, vol. 5, pp. 1842-1845, 2016.
- [6] M. Karim and R. M. Rahman, "Decision tree and Na ve Bayes algorithm for classification and generation of actionable knowledge for direct marketing," *Journal of Software Engineering and Applications*, vol. 6, pp. 196-206, 2013.
- [7] K. Ganesan, "Classification framework based on C4.5 algorithm for medicinal data," *International Journal of Computer Science and Information Security*, vol. 13, no. 4, pp. 63-67, April 2015.
- [8] N. Settouti, M. E. A. Bechar, and M. A. Chikh, "Statistical comparisons of the top 10 algorithms in data mining for classification task," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 1, pp. 46-51, 2016.
- [9] N. Dogan and Z. Tanrikulu, "A comparative analysis of classification algorithms in data mining for accuracy, speed and robustness," *Inf Technol Manag*, vol. 14, pp. 105-124, 2013.
- [10] S. Oujdi and H. Belbachir, "Spatial data mining by decision trees," *International Journal of Computer and Information Engineering*, vol. 8, no. 12, pp. 2181-2184, 2014.
- [11] S. Singh and P. Gupta, "Comparative study ID3, CART and C4.5 decision tree algorithm: A survey," *International Journal of Advanced Information Science and Technology*, vol. 24, no. 27, pp. 97-103, July 2014.
- [12] I. Jenhani, N. B. Amor, and Z. Elouedi, "Decision trees as possibilistic classifiers," *International Journal of Approximate Reasoning*, vol. 48, pp. 784-807, 2008.
- [13] P. P. Patil, P. Revankar, and P. Joshi, "The application of data mining for direct marketing," in *Proc. Second International Conference on Emerging Trends in Engineering and Technology*, 2009, pp. 312-316.

- [14] H. Hashim, A. A. Talab, A. Satty, and S. A. Talab, "Data mining methodologies to study student's academic performance using the C4.5 algorithm," *International Journal of Computer Science and Information Security*, vol. 13, no. 4, pp. 104-113, April 2015.
- [15] S. Olafsson, X. Li, and S. Wu, "Operation research and data mining," *European Journal of Operational Research*, vol. 187, pp. 1429-1448, 2008.
- [16] K. Zeitouni and N. Chelghoum, "Spatial decision tree - application to traffic risk analysis," presented at IEEE International Conference on Computer Systems and Applications, Beirut, Lebanon, 25-29 June 2001.
- [17] M. Ester, H.-P. Kriegel, and J. Sander, "Algorithms and applications for spatial data mining," in *Proc. Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS, Taylor and Francis*, 2001, pp. 1-32.
- [18] D. Mihai and M. Mocanu, "A practical comparison on GIS data of two data mining algorithms," presented at the 2nd European Conference on Electrical Engineering and Computer Science, Bern, Switzerland, December 20-22, 2018.

Copyright © 2021 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).



Dana Mihai received the B.Sc. degree from the Computer Science Department, University of Craiova, Romania, in 2010. She obtained her M.Sc. degree from the same institute, in 2012. She received B.Sc. degree from the Land Measurements and Cadastre Department of the same university, in 2018. Currently, she is a Ph.D. student at the Department of Computers and Information Technology, University of Craiova, Romania. Presently, she works as a teacher at the Department of Computer Science of the same university. Her current research focuses on data mining, spatial data mining, etc.



Mihai L. Mocanu received his Ph.D. degree from the University of Craiova, Romania, in 1999. He was a visiting research scholar and consultant for software development for image-guided surgery projects with the Imaging Science and Information Systems (ISIS) Center, Department of Radiology, Georgetown University, Washington D.C., U.S.A., from 2001 to 2002. He is currently a full professor and PhD supervisor within the Department of Computers and Information Technology, and vice-dean of the School of Control and Computers, since 2016. He has authored/coauthored 10 books and more than 150 papers published in journals or conference proceedings, and was a lead researcher/director for more than 20 research contracts. His research interests include parallel computing, data analysis and modeling, (discrete event) systems simulation, computational (medical) imaging, advanced software design methods and software engineering (UML, agile methodologies).