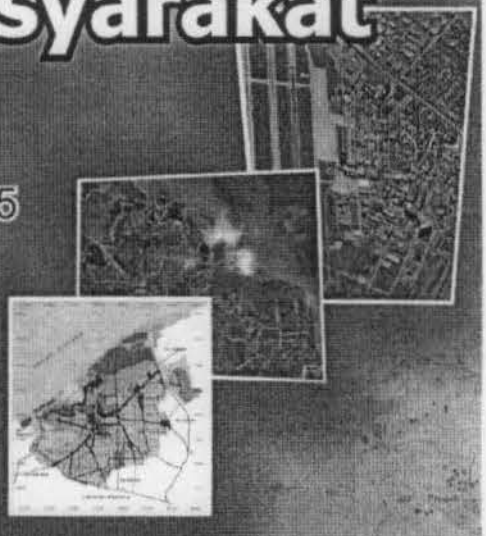
PROSIDING
SIMPOSIUM NASIONAL SAINS GEOINFORMASI I

Meningkatkan Peran dan Kualitas Data Spasial untuk Melayani Masyarakat

17 - 18 November 2009
Gedung Pascasarjana Lt 5



PUSPICS
FAKULTAS GEOGRAFI
UNIVERSITAS GADJAH MADA

ISBN: 978-979-98521-3-7

Prosiding Simposium Sains Geoinformasi – I : Meningkatkan Peran dan Kualitas Data Spasial untuk Melayani Masyarakat
PUSPICS Fakultas Geografi Universitas Gadjah Mada Yogyakarta, 17-18 November 2009. ISBN 978-979-98521-3-7

SPATIAL DECISION TREE: A REVIEW

Imas Sukaesih Sitanggang
Computer Science Department, Bogor Agricultural University
Jl. Meranti, Wing 20 Level V, Kampus IPB Darmaga, Bogor 16680 – Indonesia
E-mail: imas.sitanggang@ipb.ac.id

ABSTRACT

The automated discovery of spatial information has leaded to widespread use of spatial databases. Spatial data have been collected in various computer systems such as Geographical Information Systems (GIS). This fact leads to an increasing interest in mining interesting and useful but implicit spatial patterns. Data mining refers to extracting or "mining" knowledge or patterns from a large number of data. Spatial data mining is a process to discover spatial patterns from huge amounts of spatial data. In the spatial data mining system, the attributes of the neighbors of an object may have a significant influence on the object itself. The research in spatial data mining has gained a high attraction due to the importance of its applications. Classification, which is one of the important tasks in data mining, has been used in learning process in order to develop models (classifiers) from spatial data training. Then, the model can be used to predict the class of new data. Decision tree induction is the widely used method in classification tasks. Spatial decision trees refer to a model expressing classification rules induced from spatial data. In this paper review, we present some works in implementing a spatial data mining algorithm especially spatial decision tree algorithms.

Keywords: *spatial data mining*, *classification*, *spatial decision tree*

1. INTRODUCTION

The automated discovery of spatial information has leaded to widespread use of spatial databases, weather and climatologically data. Spatial data from both direct acquisition such as field surveying and remote sensing and indirect acquisition such as existing paper map and available datasets need huge repositories in order to easy manage and utilize for analysis purposes. In addition, spatial data have been collected in various computer systems such as Geographical Information Systems (GIS). This fact leads to an increasing interest in mining interesting and useful but implicit spatial patterns. Data mining refers to extracting or "mining" knowledge or patterns from a large number of data. Spatial data mining is a process to discover spatial patterns from huge amounts of spatial data. A spatial database contains objects which are characterized by spatial attributes as well as by non-spatial attributes. Spatial attributes are used to define the spatial location and extent of spatial objects (Bolstand (2002) in Shekhar (2004)). Non-spatial attributes such as name and population for a region are the same as the attributes used in the classical data mining. The discovery process for spatial data is more complex that for non-spatial data, because spatial data mining algorithms have to consider the neighbors of objects in order to extract useful knowledge (Ester et. al., 2001). In the spatial data mining system, the attributes of the neighbors of an object may have a significant influence on the object itself. The paper is organized as follows. Section 2 presents a short review about decision trees. Spatial decision trees are discussed in section 3. Finally section 4 summarizes this review paper.

2. DECISION TREE

Classification is a process in which a classifier (a model) is built from data training. Then, the model can be used to predict the class of new data. The task of classification aims to discover classification rules that determine the label class of any object (Y) from the values of its attributes (X). Decision trees are logical models formulated as a tree structure that show how the target variable/attribute can be determined or predicted by the set of predictive attributes. The target attribute is the variable whose the value modeled by others variables, whereas the predictor attributes are the variable whose values used to predict the value of the target variable. A decision tree contains three types of nodes: 1) a root node, 2) internal nodes, either a root node or an internal node contains attribute test conditions to separate records that have different characteristics, 3) leaf or terminal nodes, each leaf node is assigned a class label. A decision tree is a model expressing classification rules. A rule obtained from a decision tree consists of test attributes and their value in tree paths starting from the root node to the leaves node (terminals). Rules extracted from decision trees can help us to understand the data. There are many algorithms for decision tree induction. Some of them are ID3 Iterative Dihotomiser) developed by J. Ross Quinlan during the late 1970s and early 1980s, C4.5 as a successor of ID3, and CART (Classification and Regression Tree) proposed by L. Breiman et.al in 1984. CART generates binary decision trees.

3. SPATIAL DECISION TREE

Spatial decision trees refer to a model expressing classification rules induced from spatial data. The training tuples for this task consist of not only object attributes but also the neighbors of objects. Besides, the attributes of the neighboring object may have a significant influence on the object itself. Spatial decision trees differ from conventional decision trees by taking account implicit spatial relationships in addition to other object attributes (Zeitouni and Nadjim, 2001).

Many works have developed the decision tree algorithms for spatial data. This section will outline some work in applying the decision tree algorithm for spatial data. The discussion is divided into two groups. The first is the applications of the conventional decision tree algorithms (non-spatial algorithms) in spatial data. The later is the applications of spatial decision tree algorithms in spatial data.

3.1. Application of Conventional Decision Tree Algorithms on Spatial Data

In general, conventional decision tree algorithms require a single table as the training dataset to develop the classifier. The training set consists of some predicting attributes that can be numerical or categorical and a target attribute (a target class). When the algorithms will be executed on spatial data in order to extract useful knowledge, some preprocessing data steps need to be performed. In this case, a dataset not only contains the object attributes but also other data including the relationship between the objects and their neighbors as well as the attribute of the neighbors. For example, if we analysis the spread of hotspot locations in a particular region then we may consider other spatial features as neighbors of the hotspot locations such as roads, rivers, locations of plantation, land use types in the region, etc. Some topological and metric operations relate a spatial feature to their neighbors. When we apply the conventional decision tree algorithm in the such a dataset, we should integrate at least two different tables: 1) an analyzed objects table consists of predicting attributes and a target class, and 2) a relationship table which

store the relation between target objects and their neighbors. In addition to the two tables, in some case we may have additional tables contain other predictive attributes from the neighbor objects. Below are some works in applying conventional decision tree algorithms to spatial data summarized from some references. The review outlines the input dataset, the method used, and the results (decision trees).

Jianting et.al (1999) proposed the procedure for building the decision tree. They calculated the information entropy to determine the attribute partition. Input dataset is the Yellow River Delta (YRD) soil dataset consisting of three explanatory attributes: soil structure, soil essence and soil salinity, and the decision attribute (the target attribute): soil type. All attributes are represented in form of polygons in thematic maps. The output include the association relationship between soil structure, soil essence, soil salinity and soil type in form of a decision tree and a set of classification rules. The decision tree rules derived from decision tree algorithm (Jianting, 1999). Some of classification rules are as follows:

1. If soil salinity is greater than 0.8%, then soil type solely belongs to Salic Fluvisols.
2. If soil salinity is between 0.1% and 0.8% then soil type solely belongs to Gleyic Solonchaks.
3. If soil salinity is less than 0.1% and soil essence belongs to one of ....then soil type belongs to Calcaric Fluvisols.
4. If soil salinity is less than 0.1% and soil essence belongs to one of ... and soil structure belongs to Clay, then soil type belongs to Gleyic Combisols.

Fengqi and A-Xingzhu (2003) extracted knowledge of soil-landscape models from a soil map using See5 decision tree algorithm (http://www.rulequest.com/). See5 recursively grows a tree top-down through batch processing of the training data, using a greedy heuristic to search for a simple tree based on Information gain (Fengqi and A-Xingzhu, 2003). The attribute with the highest information gain is chosen as the splitting attribute at a particular node. Input dataset is the training dataset consists of relevant environmental variables and spatial attributes including elevation, slope gradient, planform curvature, profile curvature, and geology, and the classified category (or 'label') is the soil type. In addition to the aforementioned attributes, (Fengqi and A-Xingzhu, 2003) added the two variables representing the topological relations between soil types: the upslope neighbor and downslope neighbor of a given soil type. The part of a decision tree with spatial neighbor information is as follows:

```
Bedrock = Oneota:
:...Elevation <= 1304.62:
:   :...downNeighbor = Dorerton: Elbaville (5)
:   :   downNeighbor = Elbaville: Dorerton (45/1)
:   :   downNeighbor = Churchtown: Elbaville (25)
:   Elevation > 1304.62:
:   :...upNeighbor = Valton: Lamoille (38)
:       downNeighbor = Elbaville: Lamoille (2)
:       upNeighbor = None: Valton (41/1)
Bedrock = Alluvium:
:...OffGlauconite <= 15.57835:
    :...Slope <= 0.327846: Churchtown (50/6)
    :   Slope > 0.327846: Elbaville (13)
    OffGlauconite > 15.57835:
    :...Elevation <= 860.18: Orion (47/5)
        Elevation > 860.18:
        :...Wetness <= 6.705073: Council (47/6)
            Wetness > 6.705073: Kickapoo (31/1)
```

Figure 1. Decision tree (Fengqi and A-Xingzhu, 2003)

An example of rule obtained from the decision tree (Figure 1) is if Bedrock is *Oneota*, elevation is less that or equal to 1304.62, the downNeighbor is *Dorerton*, the soil type would be *Elbaville*, and there are 5 examples of this soil type in the training set, which are all correctly classified using the tree structure.

3.2. Application of Spatial Decision Tree Algorithms

Ester et. al. (1997) proposed an algorithm based on well-known ID3 algorithm (Quinlan 1986) that was designed for spatial databases. The algorithm considers not only attributes of the object to be classified but to consider also attributes of neighboring objects. Suppose o is the attributes of the object to be classified. Ester et. al. (1997) defined a generalized attribute for some neighborhood path $p = [o_1, \ldots, o_k]$ as a tuple (attribute-name, index) where index is a valid position in p representing the attribute with attribute-name of object o_{index}. The proposed classification algorithm Ester et. al. (1997) allows the input of a predicate focusing the search for classification rules on the objects of the database fulfilling this predicate. For the detail algorithm in pseudo code notation, refer to Ester et. al. (1997). Below are two rules derived from the decision tree:

> IF population of city = low AND amount of taxes of city = very high THEN economic power of city = high (87%).
>
> IF population of city = high AND type of neighbor of city = road AND type of neighbor of neighbor of city = airport THEN economic power of city = high (95%)

There are some notes for the spatial classification algorithm (Ester et. al., 1997):

1. The method does not analyze aggregate values of non-spatial attributes for the neighboring objects. For example, if a city is close to three regions with medium population, such a city may have similar properties as a city close to a single region with large population (Koperski et. al., 1998).
2. The algorithm does not perform relevance analysis and thus, it may produce overspecialized, poor quality trees (Koperski et. al., 1998).
3. The algorithm does not take into account concept hierarchies that may exist for the non spatial and spatial attribute values (Koperski et. al., 1998).
4. The algorithm does not make distinction between thematic layers. It takes into account only one spatial relationship (Zeitouni and Nadjim, 2001).

To overcome some problems in the spatial classification method proposed by Ester et.al. (1997), Koperski et.al (1998) introduced a new algorithm to develop a decision tree from spatial data. Their approach to spatial classification is based on both (1) non-spatial properties of the classified objects and (2) attributes, predicates and functions describing spatial relation between classified objects and other features located in the spatial proximity of the classified objects. Object spatial may be characterized by different types of information:

- Non-spatial attributes of object (including both classified object, for example O, and other objects used for description), for example the number of salespersons in a store.
- Spatially related attributes with non-spatial values, for example population living within 1 km from a store.
- Spatial predicates, for example *distance_less_than_10km*(X, sea).
- Spatial functions, for example *Driving_distance*(X, beach)

The two-step approach was implemented in finding the spatial predicates and functions. The first step is some rough computation and the other is fine computations for the promising patterns. In addition to simple decision tree, the use of concept hierarchies result in faster computations. Koperski et.al. (1998) developed the RELIEF algorithm using nearest neighbor

approach to find coarse predicates, functions, and attributes which are relevant to the classification task. To handle aggregate values of non-spatial attributes in the thematic maps, the buffer operations is used to compute aggregates for all relevant attributes of thematic maps. The aggregate data can be also generalized and merged with the predicate data so finally each object can be classified using a set of predicates describing properties of both thematic map and other object intersecting area for each object. An example is presented in Table 1 and the resulted decision tree is shown in Figure 2. The tree is generated using ID3 algorithm modified to allow processing of description in the form of sets of predicates (Koperski et.al, 1998).

Table 1. Generalized decsriptions of classified obejcts presented as sets of predicates

OID	high_profit	Predicates
1	Y	sum_population(x, MEDIUM), avg_income(x, SMALL), close_to(x,park), close_to(x,water)
2	Y	sum_population(x, LARGE), avg_income(x, MEDIUM), close_to(x,water)
3	N	sum_population(x, MEDIUM), avg_income(x, LARGE), close_to(x,park), close_to(x,water)
4	N	sum_population(x, SMALL), avg_income(x, MEDIUM)
5	N	sum_population(x, LARGE), avg_income(x, LARGE), close_to(x,park)

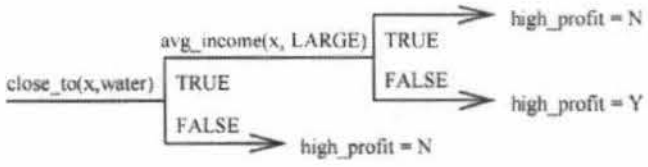


Figure 2. Decision tree for data from Table 1

Chelghoum et. al. (2002) proposed the SCART (Spatial CART) as the extension of the CART method. CART (Classification and Regression Trees) is proposed by Brieman et. al. in 1984. The basic methodology of divide and conquer described in C4.5 is also used in CART. The differences are in the tree structure, the splitting criteria, the pruning method, and the way missing values are handled (Kohavi and Quinlan, 1999). CART analysis is a form of binary recursive partitioning. The term "binary" implies that each group of objects, represented by a "node" in a decision tree, can only be split into two groups. Thus, each node can be split into two child nodes, in which case the original node is called a parent node. The term "recursive" refers to the fact that the binary partitioning process can be applied over and over again. Thus, each parent node can give rise to two child nodes and, in turn, each of these child nodes may themselves be split, forming additional children. The term "partitioning" refers to the fact that the dataset is split into sections or partitioned.

SCART determines which combination of the attribute values and spatial relationship of neighboring objects provide the best criteria for growing the tree. This algorithm considers the geographical data organized in thematic layers, and their spatial relationships. In SCART, the measure used to split a node is Information Gain. The attribute with the highest information gain, is chosen as the splitting attribute at a node. A decision tree generated by SCART has some properties (Chelghoum et. al., 2002)):

1. a node may be partitioned according to a criterion resulting from neighboring objects, which may have a particular spatial relationship with the target objects.

2. The tree is binary. In order to avoid duplications, the right son of a node is defined as the complement of the left son (right_son = node – left_son). An encoding technique has been adopted to identify each node. The root has a value code of 1. A node code is then defined recursively by:

   *left_son_code = 2 * father_code*
   *right_son_code = 2 * father_code + 1*
3. Computation of the information gain combines the neighbors' attributes and their distance or their topological relationships with target objects.

To calculate the exact spatial relationship between the locations of two collections of spatial objects SCART has the Spatial Join Index (SJI) table (Zeitouni et.al (2000)) as one of input parameters. Figure 3 shows the SJI table in which spatial relationships are represented in the scheme: (ID1, *spatial-relationship* (SR), ID2).

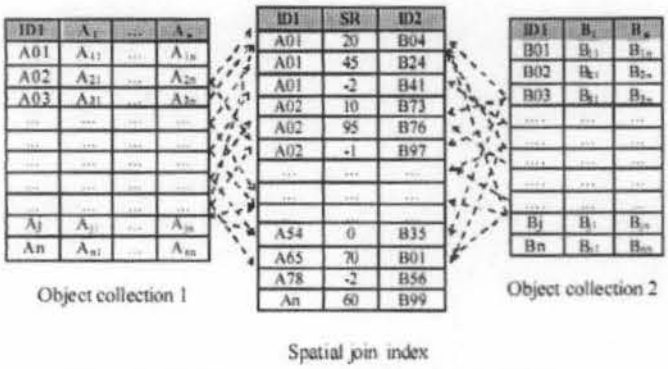


Figure 3. Spatial Join Index (Zeitouni et.al, 2000)

Each tuple (ID1, *spatial_relationship* (SR), ID2) in the SJI table references matching objects from thematic layers of object collection 1 and object collection 2. Relations between two spatial objects can be topological or metric. In case of metric relationship, the SJI table will store the exact distance value. The input parameters in SCART include (Chelghoum et. al., 2002):

1. A target table containing the analyzed objects (i.e. the analysed thematic layer)
2. A neighbor table stores thematic layer objects (neigbors of analyzed objects)
3. The spatial join index
4. A target attribute (i.e. class labels)
5. Predictive attributes from a target table or neighbor table that could be used to predict the target attribute
6. Saturation conditions in which the split is considered invalid. The node split is stopped when all objects in the node are in the same target attribute class. The other possible criteria may be a minimal occupation of the node, a maximal depth of tree or a thresholds value for the information gain.

The SCART has been implemented to develop relevant risk models by combining accident information with thematic information about the road networks, the population census, the buildings, and other geographic neighborhood detail. The model classifies accidents according to the involved categories (pedestrians, two-wheels – bicycles and motorcycles, or others vehicles).

Chelghoum and Zeitouni (2004) proposed three alternatives of multi-tables data mining in the context of the spatial data mining. One of them is reorganizing the data. This alterative reorganizes the data in a unique table by joining the three tables without duplicating the analyzed

objects. Chelghoum and Zeitouni (2004) proposed a new operator COMPLETE. The principle of this operator is to generate for each attribute value of the linked table an attribute in the result table. The advantage of this alternative is to avoid the duplication of the analyzed objects and to allow the use of any data mining method, without modification. Figure 4 illustrates the use of Complete Operator.

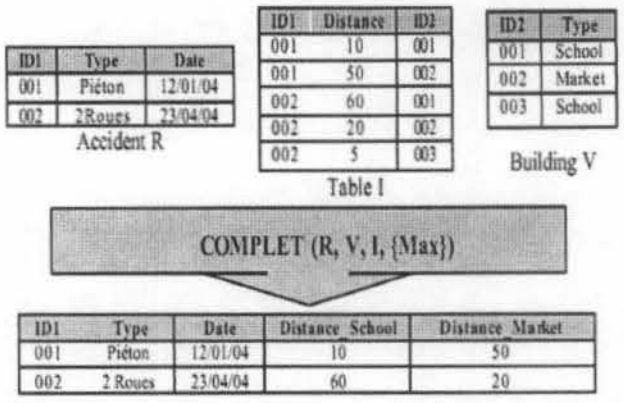


Figure 4. Illustration for the use of Complete Operator (Chelghoum and Zeitouni 2004).

In Figure 4, table I is a weighted correspondence table that joins the target table R with the table V which contains other considered dimension. This operator is only recommended when the attributes B_i of V do not contain many distinct values. The result of COMPLETE operation all objects of R without duplication and that this one is completed in right part by the weights of the "dimension" coming from V and I.

Chelghoum and Zeitouni (2004) have applied and tested the their method in the data on the road accidents and others on the geographical environment (for example road, building, etc) to construct a predictive model for traffic risk analysis. An example of decision trees obtained is shown in Figure 5.

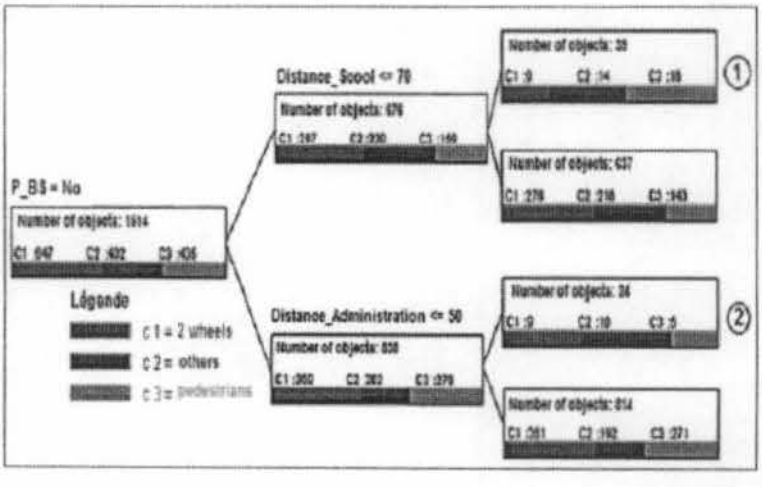


Figure 5. Spatial decision tree (Chelghoum and Zeitouni, 2004)

The decision tree classifies the accidents according to the categories (Pedestrians, 2 Wheels- bicycles and motorcycles- or other-vehicles-. The explanatory attributes are whether

linked to the road sections where the accidents are localized (P_SB: meaning the presence of stop bus), or are linked to the urban environment environment (School, market, administration, etc). As shown in Figure 5, the root correspond to the test attribute the presence of bus stop. The left son of the root corresponds the accidents localized in road sections not having a bus stop. The left node contains the accidents that are near the schools. Class target distribution in the node is 2 Wheels : 9 objects, others: 14 objects, and Pedestrians: 16.

Sitanggang et. al. (2009) applied the conventional decision tree induction namely C4.5 on Mangrove dataset using Spatial Join Index (SJI) (Zeitouni et. al., 2000) and the Complete operator (Chelghoum and Zeitouni, 2004). The dataset consists of three groups:

1. A target table contains analyzed objects i.e. the Mangrove area thematic layer.
2. Geographical environment tables of the target attribute include some thematic layers: district, landuse, substrate, geology, geomorphology, slope, and soil type.
3. A target attribute: mangrove area and its categories.
4. Predictive attributes: river, topography and other attributes obtained from geographical environment tables.

The SJI table related the spatial objects with topological relationships of two spatial objects using operator contains, overlap and inside. Below are some rules generated from the decision tree:

1. IF less than 31% area has somewhat steep slope THEN the area has No mangrove.
2. IF less than 31% area has flat slope THEN the area has mangrove with category Class2 (8.17% - 26.97%].
3. IF less than 31% area overlaps with type of substrate Sand AND the area has topography greater than 23 THEN the area has mangrove with category Class1 [0.035% - 8.17%].

4. SUMMARY

Spatial decision trees refer to a model expressing classification rules induced from spatial data. The training tuples for this task consist of not only object attributes but also the neighbors of objects. Many works have developed the decision tree algorithms for spatial data. Some of them applied the conventional decision tree algorithms (non-spatial algorithms) in spatial data. Other works developed spatial decision tree algorithms to generate classification rules from spatial dataset. Spatial decision trees have been applied in some areas such as soil-landscape, traffic risk analysis, and identifying categories of Mangrove area.

REFERENCES

Chelghoum N., Karine Z. and Azedine B., (2002), A Decision Tree for Multi-Layered Spatial Data. Symposium on Geospatial Theory, Processing and Applications, Ottawa.

Chelghoum N, and Zeitouni K., (2004), Spatial Data Mining Implementation: Alternatives and performance. Versailles. Prism Laboratory University of Versailles. Available from http://www.geoinfo.info/geoinfo2004/papers/6354.pdf

Ester M., Hans-Peter Kr., and Jörg S., (1997), Spatial Data Mining: A Database Approach. In: Proc. of the Fifth Int. Symposium on Large Spatial Databases, Berlin, Germany. Available fromhttp://eprints.kfupm.edu.sa/66096/1/66096.pdf

Ester M., Hans-Peter K., and Jörg S, (2001), Algorithms and Applications for Spatial Data Mining. Published in Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS,

Taylor and Francis. Available from http://www.dbs.informatik.uni-muenchen.de/Publikationen/Papers/Chapter7.revised.pdf

Fengqi and A-Xingzhu, (2003), Knowledge Discovery from Soil Maps using Inductive Learning. International Journal of Geographical Information Science. ISSN 1365-8816 print/ISSN 1362-3087 online © Taylor & Francis Ltd.

Jianting Zhang, Diansheng Guo, Qing Wan, (1999), Geospatial Data Mining and Knowledge Discovery using Decision Tree Algorithm—A Case Study of Soil Data Set of The Yellow River Delta. The Proceedings of Geoinformatics Conference, Ann Arbor 19-21 June 1999, pp 1-8.

Kohavi R. and Quinlan. R, (1999), Decision Tree Discovery. Available in: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.4.5353

Koperski K., Han J., and Stefanovic N, (1998), An Efficient Two-Step Method for Classification of Spatial Data. Available from http://eprints.kfupm.edu.sa/24433/1/24433.pdf

Shekhar S., Pusheng Z., Yan H., and Ranga R. V., (2004), Trends in Spatial Data Mining. In: Data Mining: Next Generation Challenges and Future Directions. AAAI Press/The MIT Press.

Sitanggang I. S., Napthalena, Sony H. W., (2009), Application of Spatial Decision Tree in Identifying Mangrove Area using C4.5 Algorithm. International Symposium and Exhibition on Geoinformation 2009. Kuala Lumpur, August 10-11, 2009.

Zeitouni K. and Nadjim C., (2001), Spatial Decision Tree – Application to Traffic Risk Analysis. IEEE Computer Systems and Applications.

Zeitouni K, Yeh L, Aufaure MA., (2000), Join Indices as a Tool for Spatial Data Mining. International Workshop on Temporal, Spatial and Spatio-Temporal Data Mining.