12-2-2013 (6)

International Symposium & Exhibition

ISG Geonformation

8th INTERNATIONAL SYMPOSIUM AND EXHIBITION ON GEOINFORMATION

ISG 2009

"Geoinformation For All"

10 - 11 August 2009 | Crowne Plaza Mutiara Hotel, Kuala Lumpur

Organized by:

UPM

UPM UNIVERSITI PUTRA MALAYSIA BERILMU BERBAKTI

NRE

isprs Information from Imagery

Jointly organized by:

USM UNIVERSITI SAINS MALAYSIA

KLIUC

Sponsored by:

Supported by:

JUPEM

PROCEEDINGS

Application of Spatial Decision Tree in Identifying Mangrove Area using C4.5 Algorithm

Imas Sukaesih Sitanggang[1], Napthalena[1], Sony Hartono Wijaya[1]

[1] Computer Science Department, Bogor Agricultural University, Jl. Meranti, Wing 20 Level V, Kampus IPB Darmaga, Bogor 16680 – Indonesia. *E-mail*: imas.sitanggang@gmail.com

Abstract – *East Borneo is one of provinces in Indonesia that has potential coast teritory for mangrove's growth. This work uses a spatial data mining method, especially spatial decision tree using C4.5 algorithm to develop a classifier to predict new data of mangrove area. We use the Spatial Join Index (SJI) and the complete operator to apply the conventional classification technique in the spatial database. The SJI is created using topological relations to find relations between two spatial objects, then the result is simplified using the complete operator. The result shows that classes of mangrove area are described by four attributes: slope, topography, substrate, and landuse. The classifier contains 23 rules with 60,66% accuracy.*

Keywords: *Spatial decision tree; C4.5; Spatial join index; Complete operator*

I. Introduction

Mangrove has a lot of benefits for life, such as beach abration protector, building material and fuel, as well as meal supplier for plankton. Therefore, mangrove forest should be protected and developed. Mangrove forests are located along tropical and subtropical beach that are influenced by tide water. East Borneo is one of provinces in Indonesia that has potential coast teritory for mangrove's growth. This work aims to develop a classifier from Mangrove data in some districts in East Borneo. We applied the data mining technique namely decision tree induction on spatial data. The resulted model can be used to identify potential area for mangrove's growth. The domain of interest for our work is to analyze how large Mangrove's area grow in a district.

Spatial data mining refers to the extraction of knowledge, spatial relationships or other interesting patterns not explicitly stored in spatial databases [4]. One of data mining techniques is classification that can be used to extract models or classifiers describing target classes or to predict unknown class labels from new data. Decision tree induction is one of popular classification techniques that widely used to construct a model from an input dataset. Spatial decision tree differs from the conventional decision tree because this technique considers the specifics of geographical data and their spatial relationships.

This work uses the Spatial Join Index (SJI) [11] and the Complete operator [1] to apply a conventional classification technique in a spatial database. The SJI is created using topological relations to find relations between two spatial objects, then the result is simplified using the Complete operator.

II. Spatial Relations

Spatial relations are usually stored implicitly in spatial databases. We need to compute the relationship using spatial operations. There are three groups of relations between a spatial object and its neighborhood [3]:

- Topological-relations, for examples: meet, overlap, covered-by, contains, inside, equal
- Metric-relation, for example: distance < d
- Direction-relation, for examples: north, south, west, east.

Topological relationships characterize the type of intersection between two spatial objects. Figure 1 illustrates three operators in topological relations: *contains*, *inside*, and *overlap* [7].

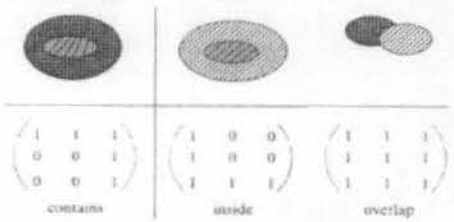

Fig.1: Binary topological relations from [7].

III. Spatial Join Index and Complete Operator

The important aspect in spatial data mining is that it considers implicit relationships between two spatial objects. Zeitouni et.al [11] proposed the structure namely Spatial Join Index (SJI) to calculate the exact spatial relationship between the locations of two collections of spatial objects. The SJI is an extension of Join Index method proposed by Valduriez [9] to improve the performance of complex operations in Database Management System.

Figure 2 shows the SJI table in which spatial relationships are represented in the scheme: (ID1, *spatial-relationship* (SR), ID2).

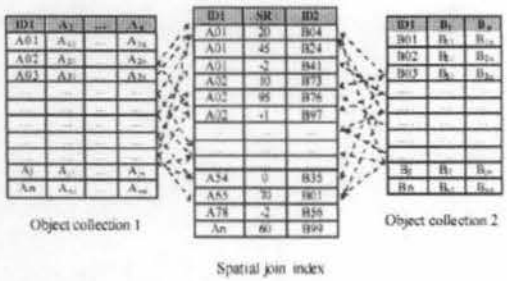

Fig.2: Spatial Join Index from [1].

Each tuple (ID1, *spatial_relationship* (SR), ID2) in the SJI table references a matching object 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.

Chelghoum and Zeitouni [1] proposed the Complete operator to organize the data in a unique table by joining the three tables, including the SJI table, without duplicating objects. The advantage of this operator is to avoid duplication of the objects in a dataset therefore the conventional data mining techniques can be applied to the dataset, without any modification. Figure 3 illustrates the use of Complete Operator.

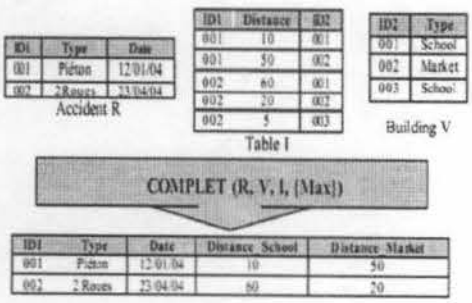

Fig.3: Illustration for the use of Complete Operator from [1].

IV. K-means

K-Means is a partitional-clustering algorithm that assigns data objects into non-overlap clusters in which each object is exactly in one cluster. Square-error, also called within-cluster variation, is a common used criterion in partitional-clustering. By applying this criterion, we will have partitions of objects with minimal total square-error. Below are the steps in K-Means algorithm [8]. Each clusters center is represented by the mean value of objects in the cluster.

Input: K: the number of clusters, D: a data set containing n objects.
Output: A set of K clusters that minimizes the square-error criterion.
Method:
(1) arbitrarily choose K objects from D as the initial cluster centers;
(2) repeat
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, i.e., calculate the mean value of the objects for each cluster;
(5) until no change;

V. Spatial Decision Tree

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). A decision tree as a model expressing classification rules. It 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 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).

C4.5 [6] is a successor of ID3 (Iterative Dihotomiser) that learns decision tree classifiers. This algorithm visits each decision node recursively and selects optimal splitting attributes until the dataset satisfies a stopping criterion. C4.5 uses Information Gain to select optimal splitting attributes [5]. ID3 algorithm is summarized in [4]. Below is the general algorithm for C.5 [6]:

1. Build the decision tree form the training set (conventional ID3).
2. Convert the resulting tree into an equivalent set of rules. The number of rules is equivalent to the

number of possible paths from the root to a leaf node.

3. Prune (generalize) each rule by removing preconditions that increase classification accuracy. Sort pruned rules by their accuracy, and use them in this order when classifying future test examples.

Spatial decision tree induction analyzes spatial objects to derive classification scheme by considering spatial objects relationships. Ester et. al [3] proposed a spatial classification algorithm based on ID3 using the neighborhood graph to represent the spatial relationship. Another approach in spatial decision tree is an extension of the conventional decision tree algorithm, CART, proposed by Chelghoum et. al. [2].

VI. Experimental Setup

In this study we apply the conventional decision tree induction to the spatial dataset from coastal area in East Kutai and Tarakan, East Borneo. Attributes in the dataset include district, mangrove, river, topography, landuse, substrate, geology, geomorphology, slope, and soil type. 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 experiment is performed in some steps:

1. Determining topological relationships of two spatial objects using operator *contains*, *overlap* and *inside*. We calculate overlapping area of two objects using intersection operation provided in Arcview.
2. Creating the SJI table to relate a pair spatial objects that contain attributes: Record ID1 (for object 1), Record ID2 (for object 2), spatial relationships (*overlap, contains* and *inside*), intersection area (%), characteristics of intersection area. Figure 4 illustrates the SJI table for two spatial objects: district and landuse. Table 1 represents spatial relationships of the objects as well as number of records in the SJI table.

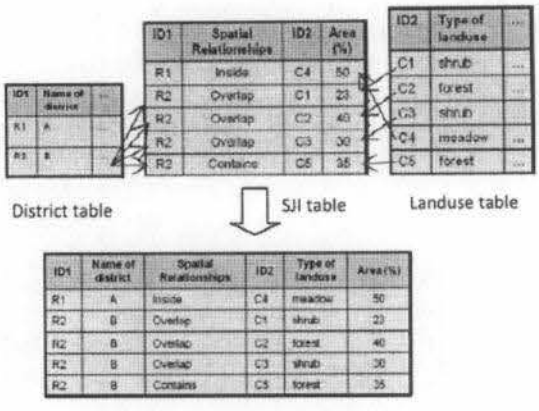

Fig.4: SJI table for two spatial objects: district and landuse.

Table 1: Spatial relationships of the objects

Spatial object	Spatial relationship	Number of record in SJI table
mangrove	Contains, overlap	122
topography	Contains	227
landuse	Contains, overlap	431
substrate	Overlap	92
geology	Contains, inside, overlap	290
geomorphology	Contains, inside, overlap	477
slope	Contains, inside, overlap	296
soil type	Contains, inside, overlap	352
Total		2287

In order to reduce the duplicate records, the Complete operator with aggregation function MAX is applied to the dataset. As the result, the SJI table only contains records that have largest overlapping area. In Figure 5 we simplify the SJI table by applying the Complete operator.

ID1	Name of district	Spatial Relationships	Type of landuse	Area (%)
R1	A	Inside	shrub	0
R2	B	Overlap	shrub	53
R2	B	Contains	forest	40
R2	B	Overlap	forest	35

COMPLETE operator with aggregation function: MAX

ID1	Name of district	Spatial Relationships	Type of landuse	Area (%)
R1	A	Inside	shrub	50
R2	B	Overlap	shrub	53
R2	B	Contains	forest	40

Fig. 5: The use of Complete Operator for SJI table in Figure 4.

3. Clustering numerical attributes using K-means algorithm. This step need to be performed to discretize numerical attributes. There are two tasks in data clustering: 1) for mangrove area in districts and 2) for overlapping area resulted from topological relations between two spatial objects. Below are classes for mangrove area:

Category	Interval for area (%)
No_ Mangrove	0
class1	[0.035 - 8.17]
class2	(8.17 - 26.97]
class3	(26.97 - 66.78]

For attributes landuse, substrate, geomorphology, geology, slope, and soil type, groups of overlapping area are combined to their types. As the result, for example, the attribute landuse has 27 new categories. New labels b1, b2, b3 are assigned to type of landuse Shrub, whereas l1, l2, l3 are for Unirrigated agriculture field, and so on (Table 2).

Table 2: Categories for landuse

Type of landuse	Interval area (%)		
	(0 – 31)	[31 – 70)	>70
Shrub	b1	b2	b3
Unirrigated agriculture field	l1	l2	l3
Settlement area	pm1	pm2	pm3
Plantation	p1	p2	p3
Embankment	e1	e2	e3
Forest	h1	h2	h3
Meadow	r1	r2	r3
Underbrush	s1	s2	s3
Unused land	t1	t2	t3

4. Combining a SJI table with other SJI tables. For conducting the experiment using WEKA, all SJI tables are combined into a single table (for some records in the table, see Table 3). Then the table is converted into the ARFF data file as the standard dataset format in WEKA. To preserve the type (semantic) of the topological relationships, we need to create different names of attributes (relevant features) as given in Table 4.

Table 4: Names and data type of attributes

Names for attributes	Data Type
topography	numerical
river	numerical
contains_landuse	categorical
overlaps_landuse	categorical
overlaps_substrate	categorical
contains_geology	categorical
overlaps_geology	categorical
contains_geomorphology	categorical
overlaps_geomorphology	categorical
contains_slope	categorical
contains_overlaps	categorical
contains_soil_type	categorical
overlaps_soil_type	categorical
class label (target attribute)	categorical

5. Building classifier using the data mining toolkit Weka 3.6.0 [10]. The dataset was analyzed with J4.8 module as Java implementation of C4.5 in WEKA. The algorithm for induction of decision trees uses the greedy search technique to induce decision trees for classification. We divide the dataset into two groups: data training and data testing. Data training is used to develop a classification model, whereas data testing is for calculating accuracy of the model.

Widely used method to predict the error rate of the classifier is 10-*folds cross validation*. In this method, dataset is divided randomly into 10 partitions in which the class is represented in approximately the same proportion as in the full dataset [10]. Each partition will be the testing set and a classification algorithm runs on the nine remaining partitions as a training set. Therefore we have 10 classification models from 10 different learning procedures. 10 error rates from 10 learning processes are averaged to calculate an overall error rate.

Table 3: Integration of the SJI tables

Topo-graphy	river	contains_ landuse	overlaps_ landuse	overlaps_ substrate	contains_ geology	overlaps_ geology	contains_ geomor-phology	overlaps_ geomor-phology	contains_ slope	overlaps_ slope	contains_ soil_type	overlaps_ soil_type	Target class
525.00	0.31	p1	h2	ncs	gg1	gg1	gmb1	gmb1	la1	lc2	tb1	tb2	No
74.00	0.02	ncl	ncl	ncs	ncg	gj2	ncgm	gma2	nclr	lc3	ncl	tc3	No
28.83	3.54	r1	b2	ncs	gf1	gg2	gmi1	gmb2	lb1	lb3	ta1	ta2	Class2
62.00	0.00	ncl	ncl	ncs	ncg	ncg	ncgm	gma3	nclr	lc3	nct	tc3	No
62.00	0.01	ncl	h3	ncs	gg1	gc2	gmb1	gmh1	nclr	la3	nct	tb3	No
107.25	5.11	p1	h3	ncs	gg1	gc2	gmf1	gmh1	nclr	la2	nct	tb3	Class1
130.29	0.58	ncl	b3	ps1	ncg	gj2	ncgm	gmf2	nclr	lc2	nct	tb2	Class1
58.33	0.11	h1	b2	in1	gg1	gb1	gmb1	gmf2	nclr	la2	ncl	tb3	Class1
86.67	0.76	b1	p1	ps1	gf1	gl3	gmi1	gmf2	lb1	lb2	ta1	tc3	Class2
4.00	2.31	ncl	ncl	in1	ncg	gf2	gmi2	ncgm	lb2	nclr	nct	nct	Class3

VII. Results

The result we obtained from the experiment is a decision tree (Figure 6) containing four test attributes: contains_slope, overlaps_substrate, topography, and contains_landuse. From the tree we can generate 23 rules, some of them are as follows:

***Rules* 1**: IF less than 31% area has somewhat steep slope THEN the area has No mangrove.

***Rules* 2**: IF less than 31% area has flat slope THEN the area has mangrove with category Class2.

***Rules* 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.

***Rules* 4** : IF less than 31% area contains type of landuse *Settlement* THEN the area has mangrove with category Class1.

***Rules* 5**: IF less than 31% area overlaps with type of substrate *Silt* THEN the area has mangrove with category Class1.

***Rules* 6** : IF greater than 71% area contains type of landuse *Shrub* THEN the area has No mangrove.

The performance of the classifier in term of accuracy is 60,66%. In the future work, we would like to apply other spatial classification techniques and some feature selection methods to improve the accuracy of the model.

VIII. Conclusion

This work built the spatial decision tree for identifying categories of Mangrove area. The experiment was performed using the Spatial Join Index (SJI) and the Complete Operator proposed by Zeitouni et.al [11] and Chelghoum et.al [1] respectively in order to apply the conventional decision tree algorithm: C4.5. The dataset for learning and testing process consists of a target table to store analyzed objects derived from the mangrove area thematic layer, and geographical environment tables of the target attribute include thematic layers: district, landuse, substrate, geology, geomorphology, slope, and soil type. Other predictive attributes are also used i.e. river and topography. The SJI table represents topological relationships between mangrove area and other spatial objects in geographical environment tables. Operators used in topological relationships are contains, overlap, and inside. To discretize overlapping area of two spatial objects we performed clustering using K-means. In order to reduce the duplicate records in the SJI table, the Complete operator with aggregation function MAX is applied to the dataset. The classifier contains 23 rules with 60,66% accuracy.

References

[1] Chelghoum N, Zeitouni K. 2004. *Spatial Data Mining Implementation: Alternatives and performance*. Versailles. Prism Laboratory University of Versailles.

[2] Chelghoum N, Zeitouni K. and Azedine B. 2002. *A Decision Tree for Multi-layered Spatial Data*. Versailles. Prism Laboratory University of Versailles.

[3] Ester M., Hans-Peter K., and Jorg S. 1997. *Spatial Data Mining: A Database Approach*. Proc. of the Fifth Int. Symposium on Large Spatial Databases.

[4] Han J, and Kamber M. 2006. *Data Mining Concepts and Techniques*. San Diego, USA: Morgan-Kaufmann.

[5] Larose, T Daniel. 2005. *Discovering Knowledge In Data : An Introduction To Data Mining*. New Jersey. Wiley-Interscience.

[6] Quinlan, J.R. 1993. *C4.5: Programs For Machine Learning*. San Mateo. CA: Morgan Kaufmann.

[7] Sekhar S, and Chawla S. 2003. *Spatial Databases a Tour*. New Jersey. Prentice Hall.

[8] Tan P, Michael S. and Vipin K. 2006. *Introduction to Data Mining*. Addison Wesley.

[9] Valduriez P. 1987. *"Join indices"*, ACM Trans. on Database Systems, 12(2); 218-246, June

[10] Witten I. H and Eibe F. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series.

[11] 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, Lecture Notes in Artificial Intelligence, 102-114, 2007.Versailles. Prism Laboratory University of Versailles.

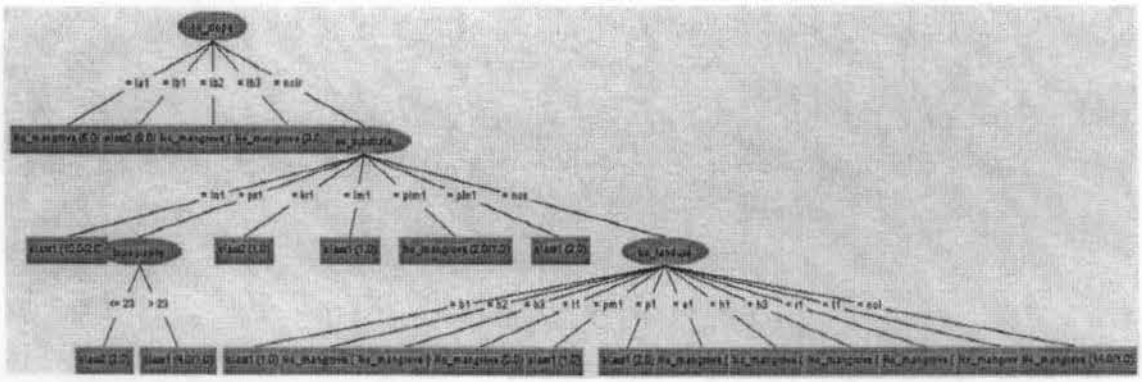

Fig. 6: Decision tree for Mangrove Area Identification.