Transactions in GIS, 2006, 10(3): 451–467

Research Article

A Spatial Entropy-Based Decision Tree for Classification of Geographical Information

Xiang Li
Naval Academy Research Institute
Brest, France

Christophe Claramunt
Naval Academy Research Institute
Brest, France

Abstract

A decision tree is a classification algorithm that automatically derives a hierarchy of partition rules with respect to a target attribute of a large dataset. However, spatial autocorrelation makes conventional decision trees underperform for geographical datasets as the spatial distribution is not taken into account. The research presented in this paper introduces the concept of a spatial decision tree based on a spatial diversity coefficient that measures the spatial entropy of a geo-referenced dataset. The principle of this solution is to take into account the spatial autocorrelation phenomena in the classification process, within a notion of spatial entropy that extends the conventional notion of entropy. Such a spatial entropy-based decision tree integrates the spatial autocorrelation component and generates a classification process adapted to geographical data. A case study oriented to the classification of an agriculture dataset in China illustrates the potential of the proposed approach.

1 Introduction

Classification of multi-attribute data is an objective of many information processing domains, particularly when applied to the analysis of financial, economical, health, environmental and demographic phenomena where the data are potentially large, complex, and not easily observable. Amongst many classification algorithms, decision trees have proven to be efficient algorithms for classification of large datasets. Decision trees are widely employed in data analysis and mining. A decision tree is a top-down, divide-and-conquer classification strategy based on automatically selected rules that partition a set of given entities into smaller classes (Breiman et al. 1984, Quinlan 1986). Each class, corresponding to a leaf of the decision tree, consists of a subset of all records

Address for correspondence: Xiang Li, Department of Earth Sciences, University of Memphis, 226A Johnson Hall, Memphis, TN 38152, USA. E-mail: li.xiang.china@gmail.com

belonging to one or several categories according to the values of a specific attribute, named 'target attribute'. Each rule is hierarchically represented by a path from the root node to a leaf via intermediate nodes and branches. The nodes of the path represent the 'supporting attributes' that maximize the distinction among the classes and minimize the diversity within each class. The branches represent the values of the supporting attributes used as the criteria to classify the dataset. The hierarchical process is equivalent to a minimization of the entropy produced by the classification at each successively lower level of the hierarchy. Executing a decision tree implies automatically selecting the appropriate supporting attributes that iteratively split the given dataset into smaller groups according to the different values of these attributes.

Several decision tree learning algorithms have been proposed, such as Classification and Regression Trees (CART) (Breiman et al. 1984), Iterative Dichotomizer version 3 (ID3) (Quinlan 1986), and C4.5 (an industrial version of ID3) (Quinlan 1993). These algorithms differ by the way they quantify the 'distinction' and 'diversity' criteria, but they all share a common hypothesis, that is, entities of the input dataset might be independent. Although these decision trees have been successfully applied to one-dimensional datasets, their application to complex multi-dimensional data, e.g. geographical data, implies that they consider the impact of the spatial dimension on the selection of appropriate classification criteria in order to avoid poor classification performance (Shekhar et al. 2003). The fact that the population of an input dataset is located in space should lead us to consider the spatial autocorrelation between entities that is implicitly equivalent to a dependent factor in the dataset. This should be particularly taken into account in the classification process when the target attribute has a positive spatial autocorrelation.

This paper introduces an experimental solution for applying decision trees to georeferenced datasets. The principle of this solution is to take into account the autocorrelation phenomenon in the classification process, and thus to apply the entropy factor but taking into account the influence of the spatial dimension. We achieve this objective by integrating a 'spatial diversity coefficient' into the decision tree, this coefficient being equivalent to a measure of spatial diversity. Without loss of generality, we illustrate the approach by integrating and applying the coefficient of spatial diversity within the ID3 decision tree, and by an illustrative case study applied to the classification of agriculture data in China.

The reminder of the paper is organized as follows. Section 2 introduces the main principles of the ID3 decision tree. Section 3 presents the principles of spatial diversity and spatial entropy. Section 4 develops the extension of ID3 towards an integration of the spatial entropy, and compares the results of the application of a spatial decision tree with a conventional decision tree. Section 5 introduces the application of the proposed approach to the classification of agriculture data in China. Section 6 briefly describes related work oriented to the application of decision trees to geographical information. Finally, section 7 concludes the paper and outlines further work.

2 Decision Trees

ID3 decision trees were first introduced to minimize the computational cost of classifying a given dataset (Quinlan 1986). They are based on an inductive learning and heuristic concept introduced by Hunt et al. (1966). ID3 decision trees apply the notion of entropy introduced by Shannon (1948) to select the most appropriate attributes as the

Table 1 Landslide sample dataset

Record No.	Landslide	Vegetation	Soil	Gradient
1	Yes	Shrub	Stone	≤40°
2	Yes	Shrub	Stone	>40°
3	No	Grass	Stone	≤40°
4	No	Grass	Sand	≤40°
5	Yes	Grass	Sand	>40°
6	Yes	Shrub	Stone	≤40°
7	No	Shrub	Sand	≤40°
8	No	Grass	Sand	≤40°
9	No	Shrub	Sand	>40°
10	Yes	Grass	Stone	>40°

nodes of a classification tree. An ID3 process generates a classification tree that facilitates the discovery of implicit relationships between a target attribute and some supporting attributes. Supporting attributes constitute the information space from which rules should be inferred to explain and classify the values of the target attribute. The target attribute is used to divide and classify the dataset into several categories. Over the past years, ID3 has proven to be one of the most popular decision trees. We use it as a reference algorithm in our work although the concepts presented can be applied to other decision trees with some minor adaptations.

Let us introduce a simplified example to illustrate the application of the ID3 decision tree. Table 1 shows a synthetic dataset recording the occurrences of landslide values, where 'Vegetation', 'Soil', and 'Gradient' are three supporting attributes and 'Landslide' the target attribute. The dataset is divided into two categories using the values 'Landslide=Yes' and 'Landslide=No'.

ID3 employs a top-down, greedy search to test all available supporting attributes in order to determine a tree node at each split of the hierarchical classification. The classification criterion used in the search process is the 'entropy' introduced by Shannon in his seminal information theory (Shannon 1948), and the 'information gain' (Quinlan 1979). The measure of entropy is applied hierarchically at each level of the decision tree to sum the diversities exhibited by the target attribute values in each category, and inferred using one to many supporting attributes at the superior levels of the hierarchy. The information gain is used to measure the expected reduction in entropy at the immediate lower level of the hierarchy, where dataset categories are refined using another supporting attribute. At each level of the decision tree hierarchy, the supporting attribute with the most entropy reduction, i.e. the largest information gain, is selected as a tree node. Entropy and information gain are respectively given as follows:

$$Entropy(GA) = -\sum_{i=1}^{n} P_i \log_2 P_i \tag{1}$$

$$Gain(GA, SA) = Entropy(GA) - \sum_{v \in Values(SA)} \frac{|GA_v|}{|GA|} Entropy(GA_v) \tag{2}$$

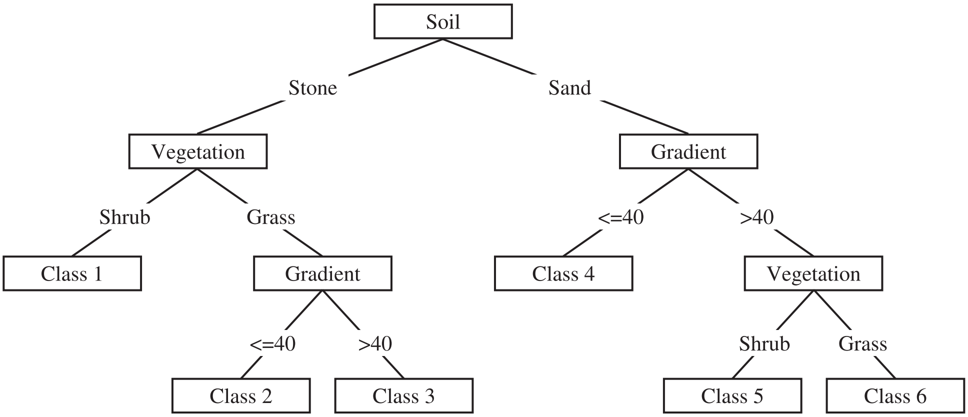


Figure 1 A decision tree based on ID3

where n is the number of categories in the enumerated domain of the target attribute GA; P_i is the proportion of the number of category i elements over the total number of records; *Values(SA)* gives the enumerated domain of the supporting attribute SA; GA_v denotes a subset of GA where the corresponding value of SA is v for each record; and $|GA_v|$ and $|GA|$ denote the cardinality of GA_v and GA, respectively.

The computation of a decision tree is an iterative process. For instance, given the dataset in Table 1, the process that initializes the decision tree is as follows:

- Step 1: *Entropy*(Landslide) = 1.0; *Gain*(Landslide, Vegetation) = 0.029; *Gain*(Landslide, Soil) = 0.28; and *Gain*(Landslide, Gradient) = 0.12. The attribute 'Soil' is selected as the root node of the decision tree since its information gain is the largest of the supporting attributes.
- Step 2: The dataset is split into two subsets according to the values of 'Soil', i.e. 'Stone' and 'Sand'.
- Step 3: Iteratively apply the above steps to branches 'Soil=Stone' and 'Soil=Sand' until there is no other supporting attribute available or the entropy is null.

Figure 1 illustrates the resultant decision tree, from which a set of if-then classification criteria, namely a decision list, is generated as follows:

- Class 1: if Soil=Stone, Vegetation=Shrub, then Landslide=Yes
- Class 2: if Soil=Stone, Vegetation=Grass, Gradient≤40°, then Landslide=No
- Class 3: if Soil=Stone, Vegetation=Grass, Gradient>40°, then Landslide=Yes
- Class 4: if Soil=Sand, Gradient≤40°, then Landslide=No
- Class 5: if Soil=Sand, Gradient>40°, Vegetation=Shrub, then Landslide=No
- Class 6: if Soil=Sand, Gradient>40°, Vegetation=Grass, then Landslide=Yes

This decision tree efficiently classifies the dataset with only one category per class. Note that for a large dataset a resultant class might consist of several categories.

Due to its deterministic and algorithmic nature, ID3 is relatively easy to implement and has become one of the most popular machine learning algorithms. ID3 has been used in a variety of commercial decision tree packages such as C4.5 (Quinlan 1993), and as a support for many classification processes applied to conventional application domains (Utgoff 1989, Mitchell 1997, Pal et al. 2001, Shao et al. 2001, McCoy et al. 2003).

3 Spatial Entropy

The spatial autocorrelation that inherently exists in most geographical phenomena is characterized by the fact that the attribute values of a geographical entity are influenced by the ones of the neighboring entities. This also means that such spatial autocorrelation has an implicit influence on the spatial distribution of a given attribute, and the diversity exhibited. Therefore, such spatial autocorrelation should be considered within a classification process. The problem that needs to be addressed is twofold: how to evaluate the influence of space on a given target attribute's diversity, and how to integrate such an influence into the measures of entropy and information gain?

Existing measures of spatial entropy or spatial diversity might be considered (Batty 1974, O'Neill et al. 1988, Li et al. 1993, Balling et al. 2004). These measures evaluate the dispersion of the entropy measure over some neighbourhoods. In a related work, a measure of spatial diversity has been introduced (Claramunt 2005). This spatial diversity is adapted to either discrete or continuous spaces, and not limited to a bounded boundary as distances rather than neighborhoods are considered (Claramunt 2005). Two supporting rules, derived from the First Law of Geography (Tobler 1970), motivate the expression of the spatial diversity. These rules are defined as follows:

- Rule 1: when different entities are closer, diversity increases.
- Rule 2: when similar entities are closer, diversity decreases.

Expressed in quantitative terms, these rules imply that the spatial diversity coefficient should increase when either the average distance between the entities belonging to a given category decreases, or the average distance between the entities of a given category and the entities of all the other categories increases, and vice versa. These average distances are named *intra-distance* (d_i^{int}) and *extra-distance* (d_i^{ext}), respectively. They are defined as follows:

$$d_i^{int} = \frac{1}{|C_i| \times (|C_i| - 1)} \sum_{j \in C_i} \sum_{k \in C_i; k \neq j} dist(j, k) \quad \text{if} \quad |C_i| > 1; \quad \text{and} \quad d_i^{int} = \lambda, \ \text{otherwise} \tag{3}$$

$$d_i^{ext} = \frac{1}{|C_i| \times |C - C_i|} \sum_{j \in C_i} \sum_{k \in (C - C_i)} dist(j, k) \quad \text{if} \quad C_i \neq C; \quad \text{and} \quad d_i^{ext} = \beta, \ \text{otherwise} \tag{4}$$

where C is the set of spatial entities of a given dataset; C_i denotes the subset of C whose entities belong to the *ith* category of the classification; d_i^{int} is the average distance between the entities of C_i; d_i^{ext} is the average distance between the entities of C_i and the entities of the other categories; *dist(j, k)* gives the distance between the entities j and k; λ is a constant taken relatively small, and β a constant taken relatively high; these constants avoid the "noise" effect of null values in the calculation of the average distances.

Although the distance considered by the spatial entropy might be any form that fulfils the metric properties of a measure of distance, we consider the Euclidean distance in its application to geographical spaces as the First Law of Geography implicitly makes reference to this interpretation of distance. These measures of distance support the extension of the conventional entropy towards a form of spatial entropy when entities are distributed and categorized in space. We integrate these average distances in such a form that exhibits an increase of spatial entropy when the intra-distance d_i^{int} increases and extra-distance d_i^{ext} decreases, and vice versa. The spatial entropy is defined as follows:

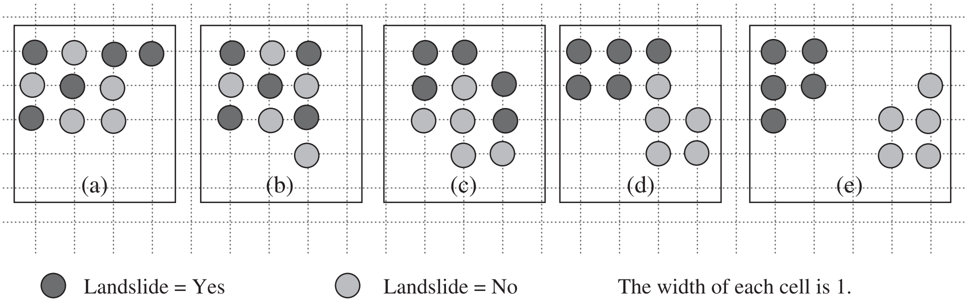


Figure 2 Spatial distribution of 'Landslide' categories

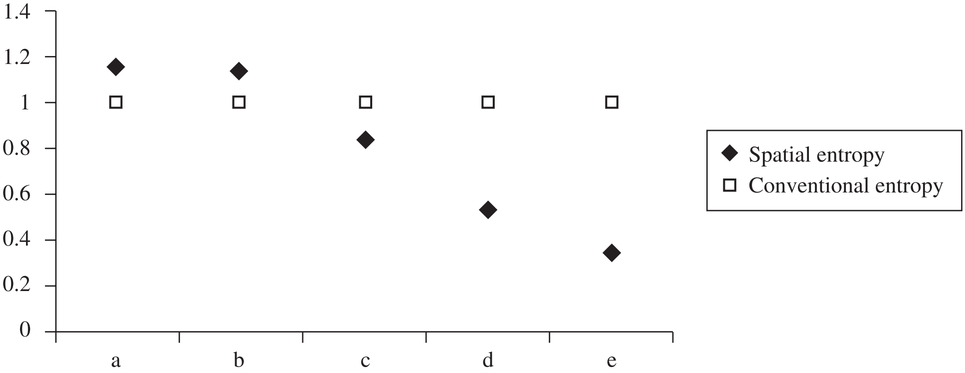


Figure 3 Conventional versus spatial entropy values

$$Entropy_s(A) = -\sum_{i=1}^{n} \frac{d_i^{int}}{d_i^{ext}} P_i \log_2 P_i \qquad (5)$$

As shown in the appendix, the values of the spatial diversity coefficient are given by the interval [0, 2]. The domains of the conventional entropy and the spatial entropy are given by the interval [0, +∞]. In order to illustrate and compare the different and respective properties of the spatial and conventional entropies, let us suppose that each entity of the landslide dataset given in Table 1 is associated with a location represented by a point in 2D space. Without loss of generality, we define five typical spatial distributions of these locations (Figure 2). Conventional and spatial entropy values applied to these spatial distributions are calculated and presented in Figure 3.

As shown in Figures 2 and 3, when the number of categories and their proportions are identical, the conventional entropy remains constant, while the spatial entropy varies with the changes of the spatial distribution of the spatial entities. This is explained by the fact that the conventional entropy is determined by the richness (i.e. number of categories, denoted by n) and the evenness (i.e. proportions of these categories, denoted by p_i) of the distribution. The spatial entropy is determined by not only the richness and the evenness of the distribution, but also by the ways the spatial entities are distributed in space. This means that changes in the spatial distribution of the dataset might increase or decrease the spatial entropy even if the category constitution remains identical. For example, in Figure 2a, the spatial entropy is relatively high as the spatial diversity revealed by the configuration is high. On the contrary, in Figure 2e, as the two categories are respectively closed at two corners of the region of interest, the spatial entropy is much smaller than the conventional entropy. These examples show that the spatial entropy surpasses the conventional entropy in the evaluation of the diversity a given spatial system exhibits.

4 Spatial Decision Tree

The notion of spatial entropy provides a means for an integration of the spatial dimension within the ID3 classification algorithm. The strategy retained is to replace the conventional measure of entropy *Entropy*( ) with the measure of spatial entropy *Entropy*$_s$( ). The information gain at each level of the decision tree is replaced with the following expression *Gain*$_s$( ):

$$Gain_s(GA, SA) = Entropy_s(GA) - \sum_{v \in Values(SA)} \frac{|GA_v|}{|GA|} Entropy_s(GA_v) \tag{6}$$

The main principle of the ID3 decision tree is still valid. At each level of such a spatial form of a decision tree, the supporting attribute that gives the maximum spatial information gain *Gain*$_s$ is selected as a node. This guarantees that the spatial entities of a same category are preferably aggregated. By contrast, application of a conventional decision tree may lead to situations in which entities of different categories might not be spatially distinguished as clearly as those of a spatial decision tree.

Moreover, the integration of the spatial entropy within the classification process can effectively reduce the number of classes generated by a decision tree. While the hierarchy of a conventional decision tree is refined until the entropy is null or there is no more available supporting attributes, the hierarchy of a spatial decision tree is stopped when different categories in a class can be clearly distinguished in space. For example, if Figure 2e presents the spatial distribution of landslide entities introduced in Table 1, where no more supporting attributes are needed to divide the dataset as the spatial distribution clearly favors the identification of two categories, i.e. ‘Landslide=Yes’ and ‘Landslide=No’.

As spatial entities are likely to have a positive spatial autocorrelation (Haining 1990), that is, $d_i^{int} < d_i^{ext}$, the spatial entropy is likely to be smaller that the conventional entropy (Figure 2). Spatial entropy thresholds can be also employed to control and terminate the growth of a spatial decision tree. If the spatial entropy of a given dataset on a given branch is lower than the threshold, this means that the branch has reached a leaf of the tree. Such a threshold can effectively reduce the tree depth and leaf numbers without compromising the classification performance. This also gives a smaller number of classes and facilitates the discovery of the most dominant classification rules. This is especially helpful for geo-referenced datasets with a large number of attributes and entities. Without thresholds, a classification usually leads to scattered results effacing the dominant features of the spatial distribution, and even overfitting. Pruning approaches as suggested in Quinlan et al. (1989) can be also applied to prospectively or retrospectively reduce the size of the resulting trees but these are still to be experimented with in the context of spatial decision trees.

Table 2 Target attribute and supporting attributes

Target attribute
Gross value of agricultural output (yuan)
Supporting attributes
Rural population (people)
Rural labor force, inclusive (people)
Rural labor force in agriculture, forestry, animal husbandry and fishing (people)
Labor force in rural industry (people)
Total area under cultivation (hectares)
All crops sown area (hectares)
Grain sown area (hectares)
Cotton sown area (hectares)
Oil crops sown area (hectares)
Grain output (tons)
Cotton output (tons)
Oil output (tons)
Meat output (tons)
Agricultural mechanization (watts)
Tractor-plowed area (hectares)
Irrigated area (hectares)
Fertilizer used (tons)
Electricity used (watt-hours)
Value of collective units' product (yuan)
Gross value of rural industrial output (yuan)
Purchased output (yuan)

5 Case Study

The spatial decision tree and the conventional ID3 decision tree have been implemented in a prototype developed in Java. A geo-referenced dataset recording the agriculture statistics for the People's Republic of China in 1990 from the Socioeconomic Data and Application Center (http://sedac.ciesin.org) supports the case study. This dataset can be considered as relatively large as it is composed of 22 attributes and 2,743 entities. Each entity represents a Chinese county. The semantics of these attributes is given in Table 2. The classification purpose is to explore the relationship between an attribute 'gross value of agricultural output' selected as the target attribute and some potential explanatory attributes as supporting attributes. This classification can be considered as part of a search for the spatial imbalance of agricultural development and the most relevant factors restricting or favoring the improvement of agricultural outputs.

Initial numerical values of each attribute are converted into categorical data. We evenly assign them into two categories, i.e. 'high' and 'low'. The dataset is divided into two categories by the values of the target attribute, i.e. 'gross value of agricultural output is high' and 'gross value of agricultural output is low'. The spatial distribution of the categories is illustrated in Figure 4.

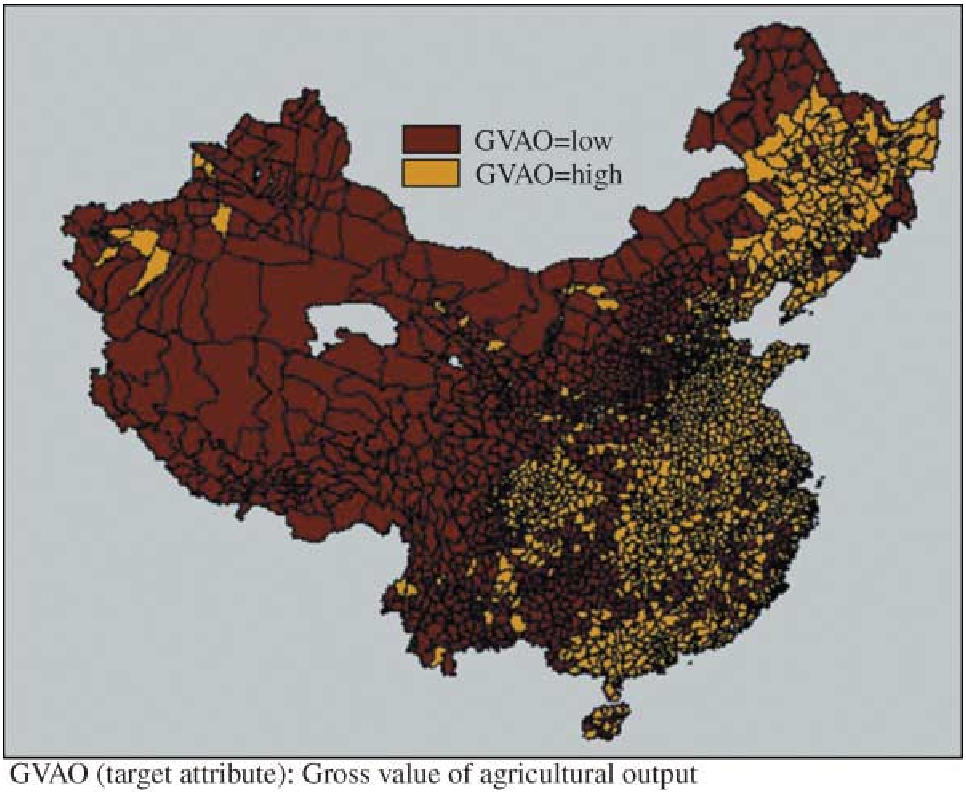
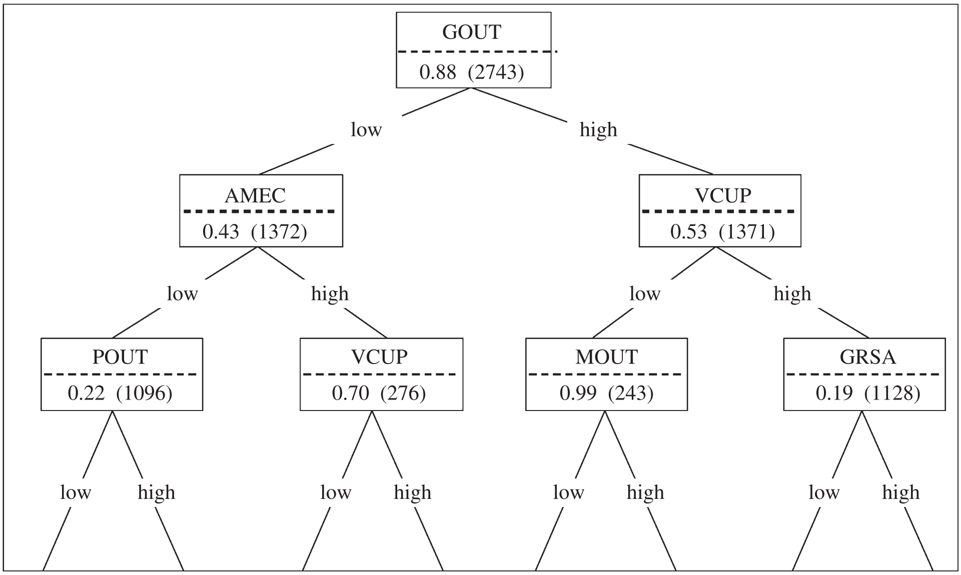


Figure 4 Spatial distribution of the target attribute's categories. This figure appears in colour in the electronic version of this article and in the plate section at the back of the printed journal

As shown in Figure 4, the target attribute has a positive spatial autocorrelation. An overall trend observed is that the eastern part of China has higher gross value of agricultural outputs than the central and western parts. This implies that a classification process should integrate the spatial dimension, and a spatial decision tree should be more adapted than a conventional decision tree for a classification of this dataset. We apply several conventional and spatial decision trees with different entropy thresholds. County centers are considered in the calculation of the measures of distances used by the spatial diversity coefficient.

Figures 5 and 6 respectively show the top levels of the spatial decision tree and the conventional decision tree when no thresholds are given. Each tree node shows the selected supporting attribute's name, its entropy or spatial entropy, and the number of entities under this node. Figure 7 illustrates the resulting number of classes using different thresholds.

Figures 5 and 6 show that the conventional and spatial decision trees generate different nodes, and that entropy values are often lower in the spatial decision tree. Figure 7 shows that, although the two trees have similar numbers of classes when no entropy thresholds are given, the spatial decision tree always generates smaller numbers of classes than the conventional decision tree. For instance, given an entropy threshold



GOUT: Grain output
VCUP: Value of collective units' product
MOUT: Meat output
AMEC: Agricultural mechanization
POUT: Purchased output
GRSA: Grain sown area
A number without parentheses refers to the value of spatial entropy.
A number within parentheses refer to the number of entities under the node.

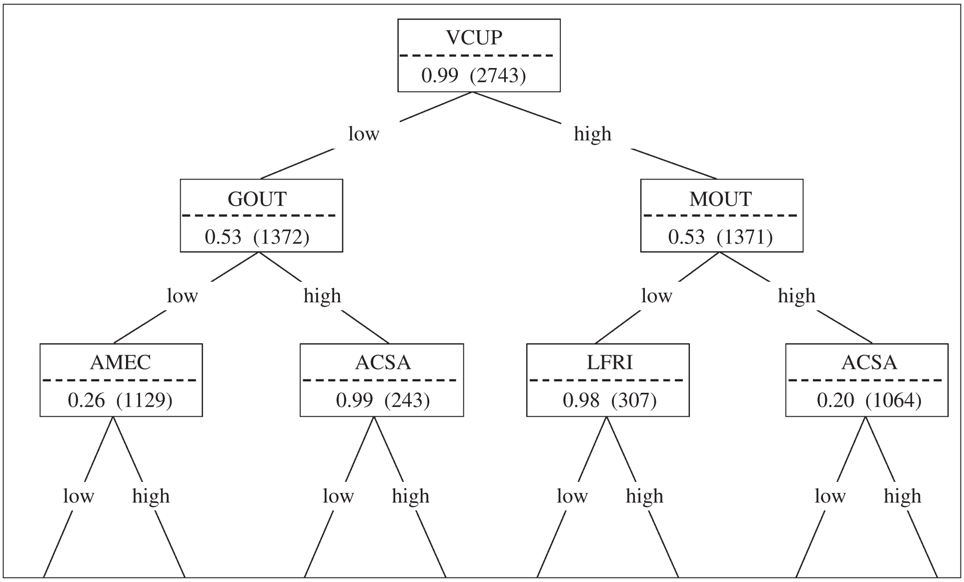
Figure 5 Top levels of the spatial decision tree

Table 3 Class description

			Number and percent of entities	
	Supporting attributes	Entropy	GVAO=low	GVAO=high
Spatial decision tree:				
S1	GOUT=low	0.431	1192 (43%)	180 (7%)
S38	GOUT=high; VCUP=high	0.199	54 (2%)	1074 (39%)
Conventional decision tree:				
C1	VCUP=low; GOUT=low	0.265	1078 (39%)	51 (2%)
C101	VCUP=high; MOUT=high	0.204	34 (1%)	1030 (38%)

GVAO: Gross value of agricultural output; GOUT: Grain output.
VCUP: Value of collective units' product; MOUT: Meat output.

of 0.5, let us compare the conventional and spatial decision trees at the class level. Some well populated classes of the spatial (e.g. classes S1 and S38) and conventional (e.g. classes C1 and C101) decision trees are selected. For classes S1 and C1, Table 3 shows that most of the spatial entities belong to the category 'low' of the target attribute, while for classes S38 and C101 most spatial entities belong to the category 'high' of the target attribute. It also appears that in classes S1 and C1 the spatial entities have similar spatial



VCUP: Value of collective units' product GOUT: Grain output
MOUT: Meat output AMEC: Agricultural mechanization
ACSA: All crops sown area LFRI: Labor force in rural industry
A number without parentheses refers to the value of entropy.
A number within parentheses refer to the number of entities under the node.

Figure 6 Top levels of the conventional decision tree

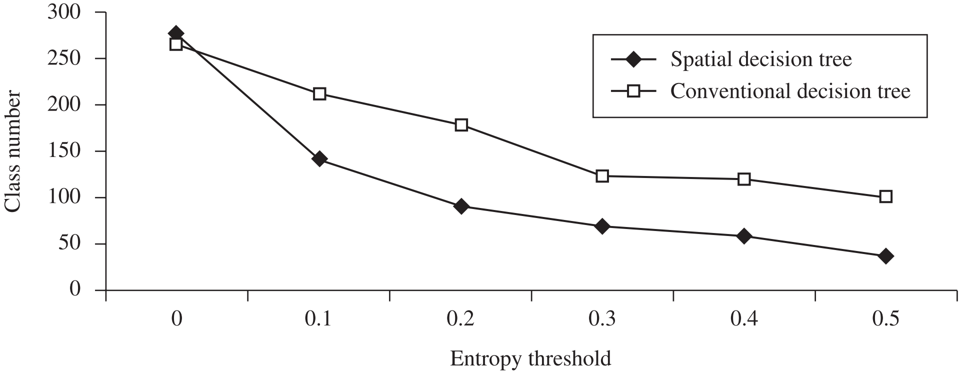


Figure 7 Class cardinalities of spatial and conventional decision trees

distributions, while in classes S38 and C101 the spatial entities have remarkably different distributions in two areas as illustrated in Figure 8.

In particular, Figure 8 shows that spatial entities with values 'high' of the target attribute in class C1 are scattered in the east and northeast parts of China, while spatial entities with values 'high' in class S1 have a more clustered spatial distribution. The clustered distributions are caused by the fact that, in the above two areas, fishing output usually contributes much to a high gross value of agricultural output of the counties

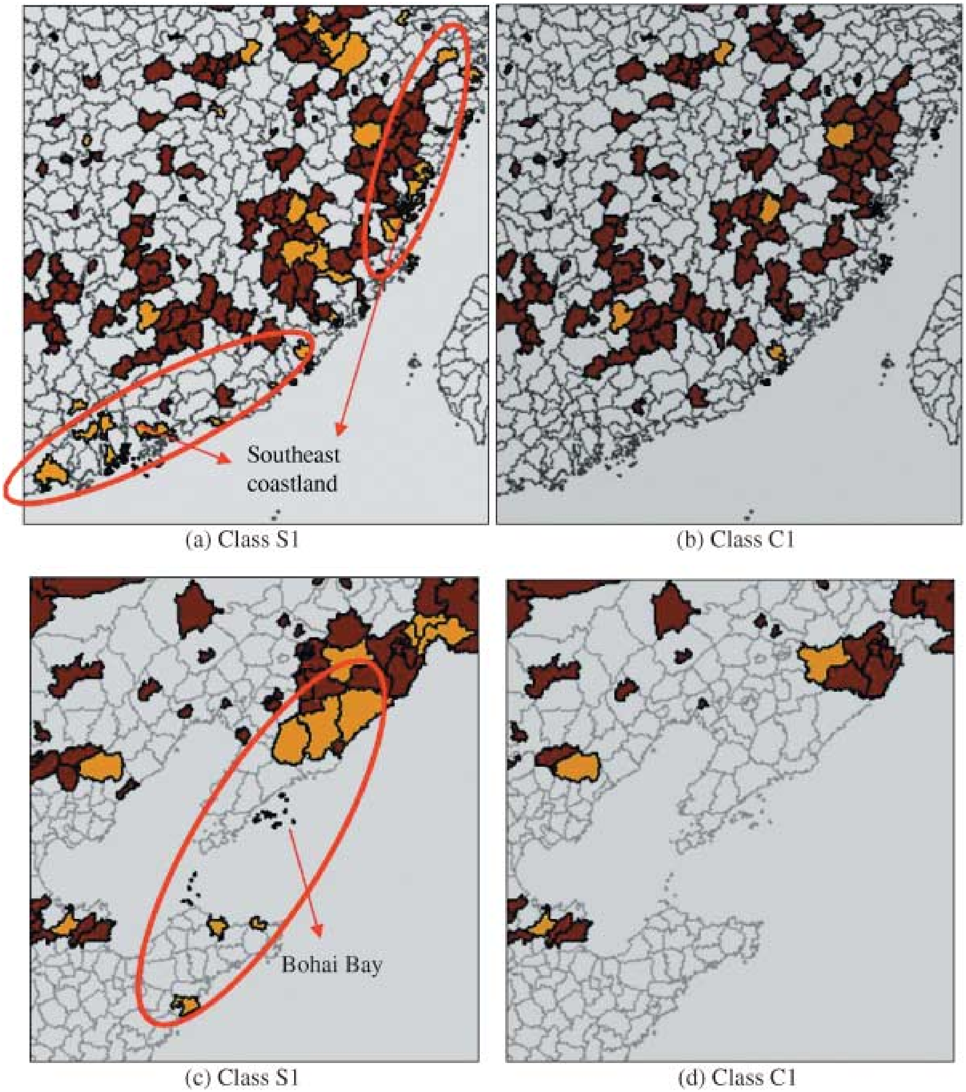
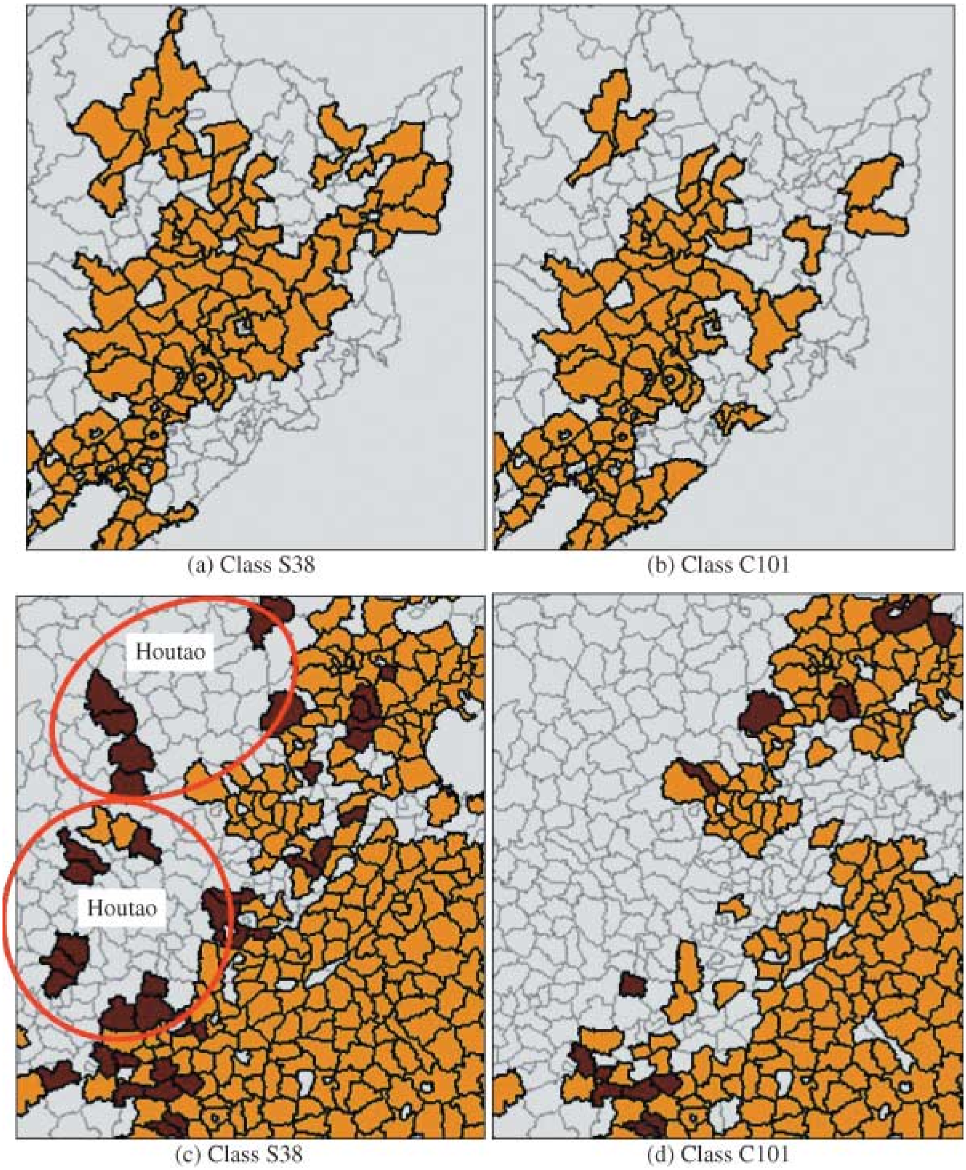


Figure 8 Spatial distribution comparisons (S1 and C1). This figure appears in colour in the electronic version of this article and in the plate section at the back of the printed journal

which have limited infield and low grain output. This fact is reflected by class S1 from which the following rule can be exhibited 'if the grain output of a county is low, then the gross value of the agricultural output of such a given county is likely to be low, or it might be high when the county is located in the southeast coastland, i.e. Figure 8a, or around Bohai Bay, i.e. Figure 8c'. One can remark that the above statement cannot be completely generated from class C1.

As shown in Figure 9, a second comparative analysis is given by the study of the populations of the spatial entities with values 'high' of the target attribute in class S38 which have a more clustered distribution than those in class C101. Spatial entities with



(a) Class S38

(b) Class C101

(c) Class S38

(d) Class C101

Figure 9 Spatial distributions of classes (S38 and C101). This figure appears in colour in the electronic version of this article and in the plate section at the back of the printed journal

values 'low' of the target attribute in class C101 are mostly scattered, while spatial entities with values 'low' of the target attribute in class S38 have a clustered distribution in the Houtao region, i.e. Figure 9c. Considering class S38, an if-then statement with a spatial semantics can be derived as follows: 'if the grain output of a county is high, and the value of the collective units' product is high, then the gross value of the agricultural output of such a county is likely to be high, or it might be low when the county is located in the Houtao region'. The statement is consistent with the fact that, in the

Houtao region, onefold agricultural structure usually leads to a low gross value of agricultural output of a county, though its grain output might be high.

These comparisons reveal that a spatial decision tree is more adapted to a georeferenced dataset with positive spatial autocorrelation than a conventional decision tree. This is exemplified by the fact that the attribute 'Grain output' of the root node of the spatial decision trees generates a better classification than other supporting attributes. This is explained by the fact that in most parts of China, the grain output determines the gross value of agricultural output of a county, but there are exceptions in some areas which are discovered from the resultant classes of the spatial decision tree.

Further refinements of the resultant classes of spatial decision trees with smaller entropy thresholds can be applied. To a great extent, results are consistent with that of the spatial decision tree when the threshold is equal to 0.5. For example, when the threshold is fixed to 0.4, one of the resulting classes (C18) mostly consists of spatial entities with values 'high' of the target attribute. These spatial entities match almost exactly those exhibited by Figures 8a and c. Similarly, an if-then statement with a spatial explanation can be derived as follows: 'if the grain output of the target attribute of a given county is low, and the agricultural mechanization is high, and the value of collective units' product is high, then such a county is most likely to have a high gross value of agricultural output in the southeast coastland, i.e. Figure 8a, or around Bohai Bay, i.e. Figure 8c'.

This case study also shows that the application of a spatial decision tree should be a flexible process where different thresholds can be gradually given in order to observe the distribution of the resulting classes. The case study shows that compactness is effectively an emerging property inherent to the spatial decision tree. Considering the spatial dimension as a supporting attribute has not been systematically considered in our classification algorithm as this implies to organize space in homogeneous regions and thus categories. This is a different objective from the one of our classification algorithm where no prior knowledge of the spatial distribution is assumed. However, modeling the spatial dimension as an attribute in which space was assigned to homogeneous regions does not imply that we did not integrate space and the distance factor in our spatial entropy measure. In this case, the spatial diversity coefficient can in fact still influence the internal spatial distribution of the attributes of the homogeneous regions.

6 Related Work

Decision trees have been successfully applied to many application domains, particularly in the field of image classification, where several extensions to conventional decision trees have been made (Ding et al. 2002; Fayyad et al. 2002). Several research proposals have also suggested extending and applying decision trees to the classification and mining of entity-based geographical data whose geometrical properties differ from image-based representations. Ester et al. (1997) introduced an algorithm applied to spatial databases and extended from the ID3 decision tree initially designed for relational databases (Quinlan 1986). The concept of the neighborhood graph is applied to represent a spatial relationship of interest, that is, topological and/or metric properties. The algorithm proposed considers the properties of neighboring objects in addition to those of a given object. Since the influences from neighboring objects and their attributes decrease when the distance to other objects increases, the length of the relevant neighborhood paths is controlled by a distance threshold given by the user. However, delimitation of the length of the neighborhood graph, and the number of neighborhood objects to consider is rather empirical, and no systemized approach for specifying these parameters has been given so far.

Under similar principles, Koperski and Han (1998) introduced a decision tree derived from the ID3 decision tree, and applied to spatial data where the influence of the neighborhood objects also considers similar and different classes. Despite its computational performance and implementation, the proposal still suffers from the fact that the influence of the neighborhood objects which is taken into account is rather empirical and not specified in a general way (i.e. weights of a given predicate are reinforced or decreased when objects of similar or different values are present in the neighborhood, respectively). Zeitouni (2001) also extended the ID3 decision tree to spatial data using a layer-based method and relational joins, but the approach is computationally costly, and limited to a very restricted number of neighborhood objects of external classes.

Although most extended, implemented, and took advantage of the ID3 decision tree, these methods suffer from an integration of a limited number of neighborhood objects. This limit makes the consideration of the spatial autocorrelation relatively imprecise. Most of these decision trees are also applied to relational data associated to spatial data using external relationships and pointers. This is a logical representation limitation, particularly as the spatial semantics of the data to be classified are not directly integrated as a proper component of the data. Last but not least, no integration of the spatial dimension is made when computing the entropy and the information gain, which are fundamental decision parameters to apply the divide-and-conquer classification strategy.

7 Conclusions

The research introduced in this paper proposes a novel method to extend the application of conventional decision trees towards geo-referenced datasets. The conventional entropy used in the processing of a decision tree is replaced with a spatial entropy measure that takes into account the influence of space and spatial autocorrelation. This leads to a spatial entropy-based decision tree, which can effectively use its hierarchical structure to reflect the spatial distribution of geographical data, and generate a classification that takes into account the spatial dimension. A case study compares spatial decision trees with conventional decision trees and shows that spatial decision trees improve the classification process of a geo-referenced dataset. Although the spatial entropy-based decision tree is designed as an extension of a conventional ID3 decision tree, the same principles can be also applied to other decision trees.

The spatial decision tree provides an efficient algorithm for classifying and exploring large geo-referenced datasets. Future work concerns the application of the proposed approach to different sorts of urban and environmental geo-referenced datasets, and the combination of its capabilities with other data mining techniques such as neural networks and association rules.

Acknowledgements

The authors are grateful to the anonymous referees for their helpful comments and suggestions.

References

Balling R C and Roy S S 2004 A spatial entropy analysis of temperature trends in the United States. *Geophysical Research Letters* 31: 11–2

Batty M 1974 Spatial entropy. *Geographical Analysis* 6: 1–31

Breiman L, Freidman J, Olshen R, and Stone C 1984 *Classification and Regression Trees*. Monterey, CA, Wadsworth and Brooks

Claramunt C 2005 A spatial form of diversity. In Mark D M and Cohn A (eds) *Spatial Information Theory: Proceedings of COSIT 2005*. Berlin, Springer Lecture Notes in Computer Science No 3693: 218–31

Ding Q, Ding Q, and Perrizo W 2002 Decision tree classification of spatial data streams using Peano count trees. In *Proceedings of the ACM Symposium on Applied Computing*, Madrid, Spain: 413–7

Ester M, Kriegel H, and Sander J 1997 Spatial data mining: A database approach. In Scholl M and Voisard A (eds) *Proceedings of the Fifth International Symposium on Large Spatial Databases (SSD'97)*. Berlin, Springer Lecture Notes in Computer Science No 1262: 48–66

Fayyad U, Grinstein G G, and Wierse A 2002 *Information Visualization in Data Mining and Knowledge Discovery*. San Francisco, CA, Morgan Kaufmann

Hunt E, Martin J, and Stone P 1966 *Experiments in Induction*. New York, Academic Press

Koperski K, Han J, and Stefanovic N 1998 An efficient two-step method for classification of spatial data. In *Proceedings of the International Symposium on Spatial Data Handling (SDH98)*, Vancouver, British Columbia: 45–54

Li H and Reynolds J F 1993 A new contagion index to quantify spatial patterns of landscapes. *Landscape Ecology* 8: 155–62

McCoy S A, Martin T P, and Baldwin J F 2003 Learning rules for odor recognition in an electronic nose. *International Journal of Uncertainty Fuzziness and Knowledge-based Systems* 11: 517–43

Mitchell T M 1997 *Machine Learning*. New York, McGraw-Hill

O'Neill R V, Krummel J R, Gardner R H, Sugihara G, Jackson B, DeAngelis D L, Milne B T, Turner M G, Zygmunt B, Christensen S W, Dale V H, and Graham R L 1988 Indices of landscape pattern. *Landscape Ecology* 1: 153–62

Pal N R and Chakraborty S 2001 Fuzzy rule extraction from ID3-type decision trees for real data. *IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics* 31: 745–54

Quinlan J R 1979 Discovering rules by induction from large collections of examples. In Michie D (eds) *Expert Systems in the Micro Electronic Age*. Edinburgh, Edinburgh University Press: 168–201

Quinlan J R 1986 Introduction of decision tree. *Machine Learning* 1: 81–106

Quinlan J R 1993 *C4.5: Programs for Machine Learning*. San Mateo, CA, Morgan Kauffman

Quinlan J R and Rivest R L 1989 Inferring decision trees using the minimum description length principle. *Information and Computation* 80: 227–48

Shannon C E 1948 A mathematical theory of communication. *Bell System Technical Journal* 27: 379–423, 623–56

Shao X Y, Zhang G J, Li P G, and Chen Y B 2001 Application of ID3 algorithm in knowledge acquisition for tolerance design. *Journal of Materials Processing Technology* 117: 66–74

Shekhar S, Zhang P, Huang Y, and Vatsavai R 2003 Trends in spatial data mining. In Kargupta H, Joshi A, Sivakumar K and Yesha Y (eds) *Data Mining: Next Generation Challenges and Future Directions*. London, AAAI Press: 357–80

Tobler W R 1970 A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46: 234–40

Utgoff P E 1989 Incremental induction of decision trees. *Machine Learning* 4: 161–86

Zeitouni K and Chelghoum N 2001 Spatial decision tree: Application to traffic risk analysis. In *Proceedings of the ACS/IEEE International Conference on Computer Systems and Applications (AICCSA'01)*, Beirut, Lebanon: 203–7

Appendix

Theorem: When the Euclidean distance is considered in the expression of the spatial diversity coefficient, then $d_i^{int} \leq 2 \cdot d_i^{ext}$ (if and only if $C_i = C$ or $C_i = \Phi$ then the spatial diversity coefficient is null).

Proof: We have to show that $d_i^{int} \leq 2 \cdot d_i^{ext}$ whatever the category (i) for a given attribute. Suppose: $P_j, P_k \in C_i$; $Q_m \in C - C_i$

then

$$dist(P_1, P_2) \leq dist(P_1, Q_1) + dist(Q_1, P_2) \Rightarrow$$

$$\sum_{P_j \in C_i; P_j \neq P_2} dist(P_j, P_2) \leq \sum_{P_j \in C_i; P_j \neq P_2} dist(P_j, Q_1) + (|C_i| - 1) dist(Q_1, P_2) \Rightarrow$$

$$\sum_{P_k \in C_i} \sum_{P_j \in C_i; P_j \neq P_k} dist(P_j, P_k) \leq \sum_{P_k \in C_i} \sum_{P_j \in C_i; P_j \neq P_k} dist(P_j, Q_1) + (|C_i| - 1) \sum_{P_k \in C_i} dist(Q_1, P_k) \Rightarrow$$

$$\sum_{P_k \in C_i} \sum_{P_j \in C_i; P_j \neq P_k} dist(P_j, P_k) \leq |C_i| \sum_{P_j \in C_i} dist(P_j, Q_1) - \sum_{P_k \in C_i} dist(P_j, Q_1) + (|C_i| - 1) \sum_{P_k \in C_i} dist(Q_1, P_k) \Rightarrow$$

$$\sum_{P_k \in C_i} \sum_{P_j \in C_i; P_j \neq P_k} dist(P_j, P_k) \leq (|C_i| - 1) \sum_{P_j \in C_i} dist(P_j, Q_1) + (|C_i| - 1) \sum_{P_k \in C_i} dist(Q_1, P_k) \Rightarrow$$

$$\sum_{P_k \in C_i} \sum_{P_j \in C_i; P_j \neq P_k} dist(P_j, P_k) \leq 2(|C_i| - 1) \sum_{P_j \in C_i} dist(P_j, Q_1) \Rightarrow$$

$$|C - C_i| \sum_{P_k \in C_i} \sum_{P_j \in C_i; P_j \neq P_k} dist(P_j, P_k) \leq 2(|C_i| - 1) \sum_{Q_m \in C - C_i} \sum_{P_j \in C_i} dist(P_j, Q_m) \Rightarrow$$

$$\frac{\sum_{P_k \in C_i} \sum_{P_j \in C_i; P_j \neq P_k} dist(P_j, P_k)}{|C_i| \times (|C_i| - 1)} \leq 2 \frac{\sum_{Q_m \in C - C_i} \sum_{P_j \in C_i} dist(P_j, Q_m)}{|C_i| \times |C - C_i|} \Rightarrow d_i^{int} \leq 2 d_i^{ext}$$