An Application of Spatial Decision Tree for Classification of Air Pollution Index

Minyue Zhao, Xiang Li*
Key Laboratory of Geographic Information Science, Ministry of Education
East China Normal University
Shanghai, China
*Corresponding author, e-mail: xli@geo.ecnu.edu.cn

***Abstract*—A decision tree is an analysis skill and a classification algorithm, whose basic principle is the combination of probability theory and an analysis tool of tree shapes. It derives a hierarchy of partition rules with respect to a target attribute of a large dataset. Nowadays, concrete coordinates exist in lots of datasets, which leads to the spatial distribution of datasets. However, conventional decision tree does not take the spatial distribution of records in the dataset into account, which makes it inadequate to deal with the geographical datasets. A number of new approaches to the analysis of geographical data have been proposed in recent years. In the purpose of evaluating the application of a spatial entropy-based decision tree, a spatial entropy-based decision tree that employed to classify the air pollution index (API) is presented in this paper. A spatial decision tree differs from a conventional tree in the way that it considers the spatial autocorrelation phenomena in the classification process. At each level of a spatial decision tree, the supporting attribute that gives the maximum spatial information gain is selected as a node. A case study oriented to the classification of API, whose study area is main cities in China, deals with the norms of the API, including density of total suspended particulate, density of SO2, density of NO2, and etc. After the process of data processing, and graphical analysis, it demonstrates a tree shape of the classification of the API and a map of the spatial distribution of the target attribute's categories, which illustrate the practicability of spatial decision tree.**

Keywords-spatial decision tree; spatial entropy; API

I. INTRODUCTION

Nowadays, there's a rapid growing demand of spatial information and spatial information system is well identified. In many areas, large quantities of data are generated and collected, therefore, tremendous spatial data in spatial database management system and spatial data warehouse is used for discovering previously unknown knowledge, and geographical pattern in experimental datasets is frequently investigated. To find useful information in spatial data, many methods are introduced like association rules, classification, clustering and etc. Classification is particularly important when applied to the analysis of financial, economical, environmental, and demographic phenomena where the data are potentially large, complex.

There are several ways to classify data like neural network, Bayesian analysis, decision tree and etc. Amongst those algorithms, decision trees have proven to be efficient for processing large datasets. Decision trees are widely applied to data analysis and mining. A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences. It is one way to display an algorithm. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. Another use of decision trees is as a descriptive means for calculating conditional probabilities. 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 [1, 2].

This paper introduces the application of spatial entropy based decision tree, which is proposed by Li et al. in 2006 [3]. This approach applies decision trees to geo-referenced datasets. It modified conventional entropy to spatial entropy on ID3 decision tree. The purpose of this paper is to illustrate the practicability of spatial decision tree by an illustrative case study applied to the classification of API data in China.

The reminder of the paper is organized as follows. Section 2 briefly describes related work oriented to the application of decision tree and spatial decision tree. Section 3 introduces the main principles and steps of spatial decision tree. Section 4 introduces the application of the proposed approach to the classification of API data in China. Finally, section 5 concludes the paper and outlines further work.

II. RELATED WORKS

Many scholars have already proposed several decision tree learning algorithms, such as Classification and Regression Trees, Decision tree classification of land use land cover for Delhi, India using IRS-P6 AWiFS data, and introduction of decision tree [1, 2, 4]. They are related to the goal of finding 'patterns' by selecting and prioritizing variables according to their hierarchy

Supported by National Natural Science Foundation of China (No. 40701142), Science and Technology Commission of Shanghai Municipality (No. 11ZR1410100), and Scientific Research Starting Foundation for Returned Overseas Chinese Scholars, Ministry of Education, China.

978-1-61284-848-8/11/$26.00 ©2011 IEEE

of influence on a dependent outcome or on a specific statistic. Through all these papers, we can conclude that they share a common view, that is, decision tree really matters in the classification of one-dimension datasets.

However, when it comes to complex spatial dimension datasets, e.g. geographical data, decision trees always cannot play a good role. It is difficult to make truthful classifications because most spatial objects are impartially gathered rather than equally distributed. Any conventional decision tree can be applied to spatial data mining without reflecting spatial correlation in classification, which is very important and should not be ignored. To extract spatial information in classification, many new decision trees have been proposed. The algorithm proposed by Easter et al. is based on ID3 and uses the concept of neighborhood graph. The limit of this method is that it takes into account only one spatial relationship of the objects [5]. Another algorithm proposed by Balling et al. evaluates the dispersion of the entropy measure over some neighborhoods [6]. Ding et al. also proposed a different algorithm. It is a decision tree based model to perform classification on spatial data streams, which use the Peano Count Tree structure to build the classifier [7]. Another algorithm is proposed by Li et al. in 2006, which integrates and applies the coefficient of spatial diversity within the ID3 decision tree [3]. It is a spatial entropy based decision tree, which takes into account the spatial distribution, and spatial gain is an important criteria when selecting the nodes.

The approach proposed by Li et al. provides an efficient algorithm for classifying and exploring large geo-referenced datasets and it is without loss of generality, so it is selected to be applied in the paper.

III. SPATIAL DECISION TREE

A. Principle

The notion of spatial decision tree provides an approach for an integration of the spatial dimension within the ID3 decision tree. It differs from the conventional tree in that it replaces the measure of entropy with the measure of spatial entropy, denoted by $Entropy_s$. For including a spatial factor on entropy, intra-distance and extra-distance are proposed. Intra-distance means the average distance between the entities of a same category. Extra-distance means the average distance between the entities of one category and the entities of the other categories. These support the extension of the conventional entropy towards a form of spatial entropy when entities are distributed and categorized in space.

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, denoted by *Gains,* is selected as a node. Nodes are where trees branch or split the data set; terminal nodes are called leaves. Each class, corresponding to a leaf of the decision tree, consists of a subset of all records 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.

B. Procedure

A spatial decision tree is based on ID3 and spatial entropy. The spatial entropy is defined as follows:

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

where n is the number of categories in the enumerated domain of the target attribute A; P_i is the proportion of the number of category i elements over the total number of records.

Besides, d_i^{int} and d_i^{ext} are respectively defined as follows:

$$d_i^{\text{int}} = \frac{1}{|C_i| \times (|C_i| - 1)} \sum_{j \in C_i} \sum_{k \in C_i, k \neq j} dist(j,k)$$

$$\text{if } |C_i| > 1; \quad \text{and } d_i^{\text{int}} = \lambda, \text{ otherwise} \tag{2}$$

$$d_i^{ext} = \frac{1}{|C_i| \times (C - C_i)} \sum_{j \in C_i} \sum_{k \in (C - C_i)} dist(j,k)$$

$$\text{if } C_i \neq C; \quad \text{and } d_i^{ext} = \beta, \text{ otherwise} \tag{3}$$

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.

The information gain at each level of the spatial decision tree is defined as follows:

$$Gain_s(GA, SA) = Entropy_s(GA) - \sum_{v \in Values(SA)} \frac{|GA_v|}{|GA|} Entropy_s(GA_v) \tag{4}$$

where $Values(SA)$ 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.

IV. CASE STUDY

In our case study, spatial decision tree is applied to the classification of API, whose study area is main cities in China. Spatial decision tree classifies the API according to some norms, in order to analysis the influence of API. The spatial decision

tree has been implemented in a prototype developed in C#. This application will serve as an example to highlight the specificity and the importance of spatial decision tree.

A. Study Area and Data

Figure 1. Spatial distribution of study area

Study area of this case covers main cities in People's Republic of China. In this case, we concern 111 cities in China, covering China's 23 provinces, 4 municipalities and 5 autonomous regions, but not including two special administrative regions and Taiwan Province. Fig. 1 shows the locations of these 111 cities.

The investigation samples of API involved many factors, such as density of total suspended particulate, density of SO2, density of NO2, and etc. To have further analysis on the effect of various factors, the best solution of our study area is to divide them into various groups. The initial materials for analysis include air quality data, socioeconomic data, and city distribution map.

TABLE I. TARGET ATTRIBUTE AND SUPPORTING ATTRIBUTES

	Attributes	Abbreviation
Target attribute	Air pollution index	API
Supporting attributes	Density of SO_2 (Mg / m^3)	SO_2
	Density of NO_2 (Mg / m^3)	NO_2
	Density of total suspended particulate (Mg / m^3)	TSP
	Green coverage rate(%)	GCR
	Per capita green area (m^2 / person)	PCGA
	Proportion of secondary industry to GDP (%)	PSI
	Average temperature (0.1 ° C)	AT
	Average precipitation (0.1 mm)	AP
	Average wind speed (0.1 m/s)	AWS

A geo-referenced dataset recording the API statistics for 111 main cities of China in February 25, 2011 from China National Environmental Monitoring Center (http://www.cnemc.cn) supports the case study. This dataset is relatively large as it is composed of 10 attributes and 111 entities. Each entity represents a Chinese city, and each attribute represents a factor. The specific definition of these attributes is illustrated in Table 1. The purpose of this classification is to explore the relationship between an attribute 'API' selected as the target attribute and some potential explanatory attributes as supporting attributes.

B. Data Pre-processing

Generally, data pre-processing includes data selection (select the relevant data), purification (elimination of redundant data), transformation and etc. The adequacy of the data pre-processing plays an important role in the efficiency and accuracy of data mining.

In this case, for the purpose of applying spatial decision tree to a geo-references dataset, initial numerical values should be extracted from the spatial data firstly. In this case, two spatial datasets are available. One records 111 cities' coordinates of the locations and the other records 287 cities' socioeconomic data respectively. The former is utilized to obtain the *x*, *y* coordinates of the city and calculate intra-distance and extra-distance which is employed to calculate the spatial entropy. The latter is used to obtain the specific value of green coverage, per capita green area, and proportion of secondary industry to GDP through the table file. Specific approach to get the value is using the Add Join operation by ArcMap, and adding 3 fields from the attribute table of one spatial dataset to the attribute table of the other spatial dataset. Finally, export the attribute table to Microsoft Excel form for storage and transform it to text file for program input.

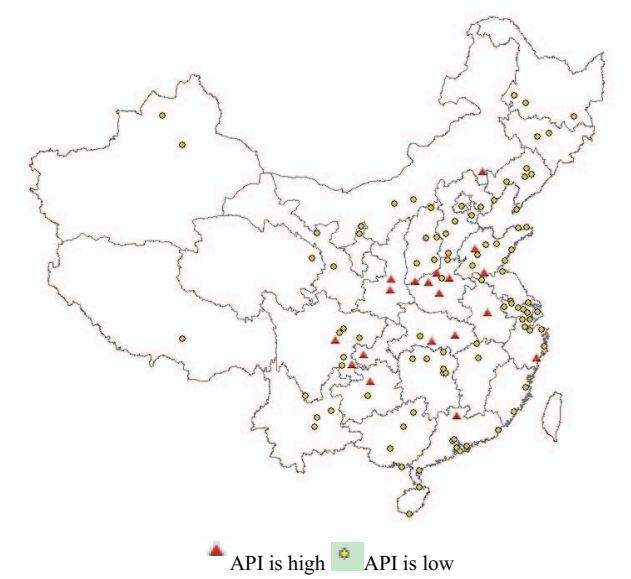

Figure 2. Spatial distribution of the target attribute's category

Next step is converting initial numerical values of each attributes into categorical data (Boolean values or Enumeration values). 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. ‘API is high’ and ‘API is low’. The spatial distribution of categories is illustrated in Fig. 2.

C. Data Calculation

The core of spatial decision tree algorithm is the selection of attributes through calculated spatial gain at all levels of decision tree nodes, in order to obtain the largest category information of the tested records at each test of non-leaf nodes.

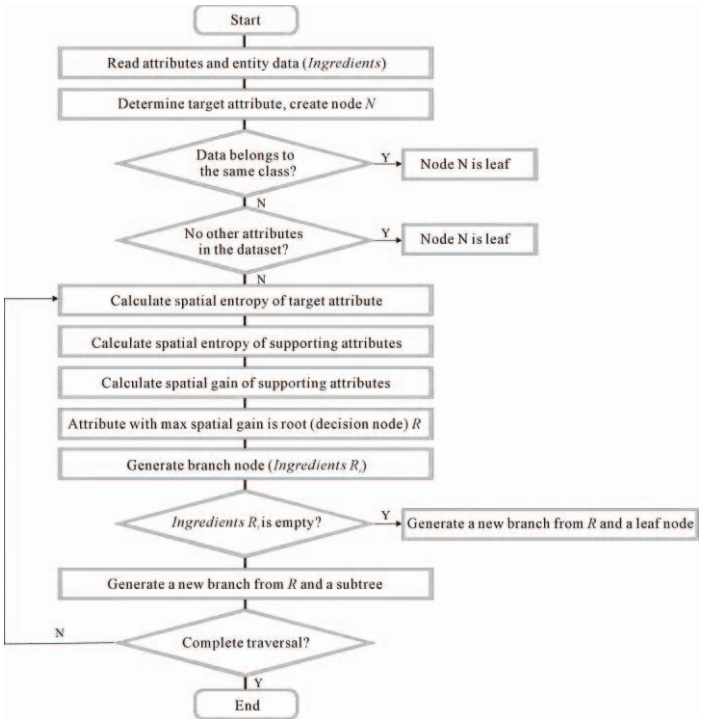

Figure 3. Overview of workflow

The calculation of spatial decision tree algorithm is the selection of the largest spatial gain of attribute as a spatial decision tree node, and then generate branches according to different values. After that, recursively invoke the method to establish the branch of the spatial decision tree node aiming at the subset of each branch, until all subsets contain the same type of data.

The main idea of programming can be concluded to 7 steps:

(1) Read the text file as input and save the attributes and entity data in array *Ingredients;*

(2) Determine the target attributes among all the attributes, and creat node *N*;

(3) Calculate spatial entropy of attributes and spatial gain of supporting attributes if code *N* is not a leaf;

(4) Choose the attribute with max spatial gain as a root *R* (or decision node *R*);

(5) Generate the branch node and save all the branch nodes in array *IngredientsRi;*

(6) Generate a new branch from *R* and a subtree if *IngredientsRi* is not empty;

(7) Traversal until *IngredientsRi* is empty;

An overview of work flow is shown in Fig. 3.

Calculation environment is the same as TABLE II:

TABLE II. CALCULATION ENVIRONMENT

Machine	***Desktop PC converted***
CPU	Core processor 3600+, 1.9GHz
Main Storage	1 GB
OS	Microsoft Windows XP
Program Language	C#

D. Results and Discussion

Under the calculation environment of Table 2, and based on the geo-referenced dataset recording the API statistics, spatial entropy and spatial information gain can be calculated, according to which we can draw a spatial decision tree. Spatial decision tree is showed in Fig. 4.

The number under the attribute means the spatial entropy of the attribute, and the number inside the parentheses means the number of entities under this branch.

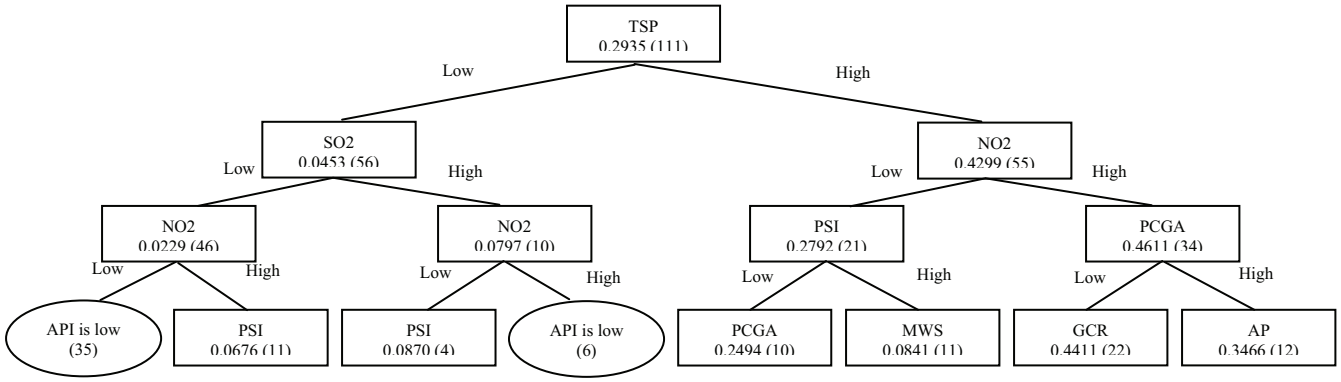

Figure 4. Top levels of the spatial decision tree

Fig. 4 shows that different factors affect API to a different degree. TSP, SO2, and NO2 are the three main factors of API. The density of TSP, SO2, and NO2 directly influence the air quality.

Figure 5. Spatial distribution of the target attribute's category if TSP is low and SO_2 is low

In Fig. 5, there are 46 cities in all if TSP is low and SO_2 is low, of which 45 cities' API are low while only 1 city's API is high.

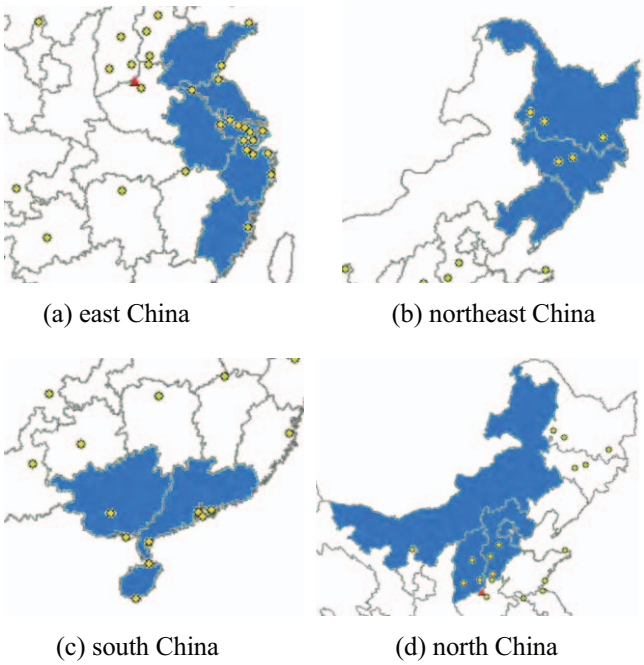

Figure 6. Clustered spatial distribution of the target attribute's category

In Fig. 6, Spatial entities with values 'low' of the target attribute API in this class have a relatively clustered distribution in east China, northeast China, south China and north China. For cities in east China, northeast China and south China are located in the coastland, and climate of these regions are characterized by monsoon climate, which reduce the density of TSP, SO_2 and NO_2. This case reveals that a spatial decision tree is well adapted to a geo-referenced dataset as it reflects the feature of spatial distribution.

978-1-61284-848-8/11/$26.00 ©2011 IEEE

V. CONCLUSION

The paper applied the spatial decision tree to a geo-referenced dataset, and spatial decision is proven to be practicability. In addition, the case shows that spatial entropy-based decision tree is well applied to the classification process of a geo-referenced dataset.

This method reflects the relation between non-spatial data and spatial data because it uses distribution of spatial objects for spatial dimension. Moreover, the proposed method enables a spatial decision tree to reflect the true relationship between non-spatial and spatial data because its spatial entropy informs not only distances but also distribution of spatial data.

Future works concern the general applicability of spatial entropy-based decision tree, and how the spatial entropy-based decision tree can be widely used in geographical research.

ACKNOWLEDGMENT

The authors would like to express appreciations to colleagues in our laboratory for their valuable comments and other helps.

REFERENCES

[1] L. Breiman, J. Freidman, R. Olshen, and C. Stone, Classification and Regression Trees, Monterey: Wadsworth and Brooks, 1984.

[2] J. R. Quinlan, "Introduction of decision tree," Machine Learning Journal, vol. 1, 1986, pp. 81-106.

[3] X. Li and C. Claramunt, "A spatial entropy-based decision tree for classification of geographical information," Transaction in GIS, vol. 10(3), 2006, pp. 451-467.

[4] M. Punia, P. K. Joshi and M. C. Porwal, "Decision tree classification of land use land cover for Delhi, India using IRS-P6 AWiFS data," Expert Systems with Applications, 2010, pp. 5577-5583.

[5] M. Easter, H. Kriegel and J. Sander, "Spatial data mining: A database approach," Springer Lecture Notes in Computer Science, vol. 1262, 1997, pp. 48-66.

[6] R. C. Balling and S. S. Roy, "A spatial entropy analysis of temperature trends in the United States," Geophysical Research Letters, vol. 31, 2004, pp. 11-2.

[7] Q. Ding, Q. Ding, and W. Perrizo, "Decision Tree Classification of Spatial Data Streams Using Peano Count Trees," Proceedings of the ACM Symposium on Applied Computing, pp. 413-417, 2004.