A PARALLEL RANDOM FOREST ALGORITHM FOR DIGITAL ELEVATION MODEL CLASSIFICATION

Pham Huy Thong[1], **Le Hoang Son**[2] **and Nguyen Dinh Hoa**[3]

VNU University of Science, 334 Nguyen Trai, Thanh Xuan, Hanoi, Vietnam,

[1] thongph@vnu.edu.vn, [2] sonlh@vnu.edu.vn, [3] hoand@vnu.edu.vn

ABSTRACT

Digital Elevation Model classification is the identification of each region on DEM that can be one of terrain types such as lake, river, valley, plain, etc. It plays an important role in GIS, especially in planning Base Transceiver Station (BTS) and traffic operation. Among classification algorithms such as fuzzy k-means and neural network, decision trees, typically ID3 and Random Forest are of higher performance than others. Although ID3 is simpler and widely used than RF, it has some disadvantages such as over-fitted, over-classified data and only one tested attribute at a certain time. RF has overcome this limitation.

In this paper, we propose an innovation of RF algorithm based on parallel computation and the multi-way tree. This algorithm builds multi–way trees instead of traditional binary trees in RF. All trees are divided into some processors to compute simultaneously in order to decrease running time. By experiments, we compare the output of the algorithm with that of ID3. Using benchmark GIS data from Bolzano–Bozen, Italy, we assess the effectiveness of the algorithm. Although the algorithm results in slower running time, its classification performance is better than that of ID3.

Keywords: 3D GIS, DEM classification, ID3, RF.

1. INTRODUCTION

Digital Elevation Model (DEM) is common input of 3D Geographical Information System (3D GIS). A DEM can be represented as a raster (a grid of squares) and it is built based on remote sensing techniques such as photogrammetry, LiDAR, IfSAR and also from land surveying, etc. (Li, Z., Zhu, Q. and Gold, 2005). The classification of DEM plays an important role in GIS and many other fields, especially in planning Base Transceiver Station (BTS) and traffic transportation. However, it is difficult to classify DEM because within the same elevation, DEMs could be different regions such as lakes, rivers, valleys, etc.

Recently, there have been many research related to DEM. The authors (P.A. Burrough et al., 2000), (B.D. Bue et al., 2006), (Ned Horning, 2010) presented some methods to classify images and landforms using DEM as an input. A series of canonical methods were used such as fuzzy k–means, Kohonen network, etc. However, they are non-overlapping algorithm and sensitive to outliers. Another method for classification issues based on the decision tree is ID3 algorithm (Sitanggang, 2011). It is a simple algorithm and widely used for learning from samples. It can be extended to deal with spatial data. The disadvantages of these algorithms are low effective and unable to solve with noisy data. Random forest – RF (Leo Breiman, 2001), which is an overlapping algorithm, can deal with noisy data. The uses of it on GIS are mostly in satellite and aerial images classification (Ned Horning, 2010). Because of its high running time, this motivates us to propose an innovation of RF algorithm with parallel strategy.

The rest of the paper is organized as follow. Section 2 gives an insight about the relevant researches in this field and Section 3 explains the innovative algorithm. Section 4 provides data description and experimental results followed by the conclusion.

2. RELATED WORKS

P. A. Burrough et al. (2000) proposed ways to classify landform. They overcame large datasets by using fuzzy k-means algorithm with distance (the Euclidian, Diagonal or Mahatanobis) metric. In addition, they provided solutions to deal with “overlapping property sets” and artefacts caused by DEM or algorithms used to derive attributes that no one had done before. Using data from Alberta, Canada, and the French pre-Alps, this method was strong and can be easily applied to large datasets without computational difficulties. However, these methods could not deal with multiple divisions of landscape within markedly differing areas.

Bue and Stepinski (2006) proposed an automated method for classifying and characterizing landforms on Mars. The classification was unsupervised and based on the self-organizing map technique. The input data was DEM with the first layer storing elevation values, subsequent layers storing additional topographic information. It was calculated by Tardem software and required some post processing in order to make it ready for clustering. Kohonen neural network technique was then used to classify. The final result of classified procedure was a thematic map of topography. Formally, the output was a matrix map of the same dimension as the planar dimension of the DTM. This design allows for a convenient visualization of results by means of a thematic map of landforms. However, topographic attributes which have been chosen to optimize identification of landforms are common on Mars but rare on Earth. It took too long to process.

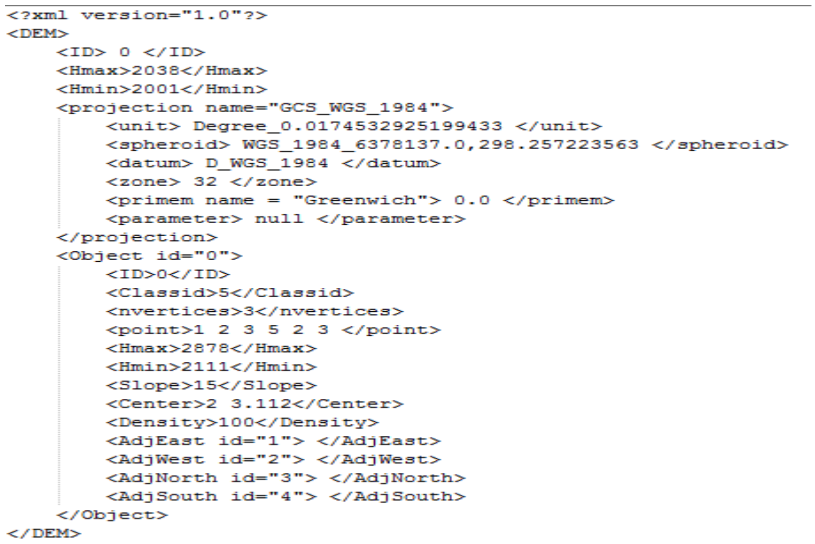
Ned Horning (2010) presented an application of RF algorithm for satellite image classification and generation of continuous data sets. The paper firstly provided an overview of RF algorithm. It was non-parametric, capable of using continuous and categorical data sets, easy to parameterize, not sensitive to over-fitting, good at dealing with outliers in training data. It calculated ancillary information such as classification error and variable importance. Secondly, the author showed the advantages and limitations of the algorithm and some implementations of it for creating land cover, biomass, and percent cover maps using satellite imagery. Though, RF had high running time and it was still not a common approach for image classification and regression largely because many remote sensing practitioners were unaware of the algorithm.

Sitanggang et al (2011) mentioned a new spatial decision tree algorithm based on the ID3 algorithm for discrete features represented in points, lines and polygons. They proposed a new formula for calculating spatial information gain by using spatial measures for point, line and polygon features. Their empirical results demonstrated that their algorithm could be used to join two spatial objects in constructing spatial decision trees with 138 leaves and the accuracy is 74.72%. Despite of their high accuracy, the new algorithm was only applied on small spatial data set and the inputs of it were only two – dimensional spatial data.

3. THE PROPOSED METHOD

In order to apply RF algorithm for DEM classification, we need further study on DEM to select its characteristics to be input attributes. Earth's surface is divided into different forms. However, there are some typical forms such as: mountain, plateau, hill, plain, river, lake. We focus on the classification of six types of terrains. In order to identify an object (a region), we have to determine some characterized parameters (Figure 1). Firstly, determining the border of a region – a set of boundary points wrapping it – is needed. Secondly, inner values of the region, for details, the maximal and minimal heights, the slope, the center and the density are required. Then, neighborhood regions located at the north, south, west and east from it are requested. Finally, parameters of projection are required.

Characteristics of each region in each DEM input will be converted into Digital Markup Language (DML) file. DML, which is based on XML, defines characteristics of its regions and types of its neighborhood regions.



```
<?xml version="1.0"?>
<DEM>
    <ID> 0 </ID>
    <Hmax>2038</Hmax>
    <Hmin>2001</Hmin>
    <projection name="GCS_WGS_1984">
        <unit> Degree_0.0174532925199433 </unit>
        <spheroid> WGS_1984_6378137.0,298.257223563 </spheroid>
        <datum> D_WGS_1984 </datum>
        <zone> 32 </zone>
        <primem name = "Greenwich"> 0.0 </primem>
        <parameter> null </parameter>
    </projection>
    <Object id="0">
        <ID>0</ID>
        <Classid>5</Classid>
        <nvertices>3</nvertices>
        <point>1 2 3 5 2 3 </point>
        <Hmax>2878</Hmax>
        <Hmin>2111</Hmin>
        <Slope>15</Slope>
        <Center>2 3.112</Center>
        <Density>100</Density>
        <AdjEast id="1"> </AdjEast>
        <AdjWest id="2"> </AdjWest>
        <AdjNorth id="3"> </AdjNorth>
        <AdjSouth id="4"> </AdjSouth>
    </Object>
</DEM>
```

Figure 1: An example of DML file.

After creating a DML file, all the data from it are used as an input of RF algorithm. Random forest is a set of decision trees and each one is built from random subsets of attributes of the training dataset through ID3 algorithm. We improve the original RF algorithm by using multi-way tree substitute for traditional binary tree. Using this kind of tree increases the number of thresholds to split and therefore, also the accuracy of the model is grown up. A multi-way tree is the same to the binary tree of ID3. It is followed the steps to construct in ID'3 algorithm:

At first, we choose some attributes randomly from the training set S to build a decision tree. We split all attributes and classes in table S to sub-tables S_i , where i is an attribute of S . Secondly, $Entropy(S), Entropy(S_i)$ are calculated by the equation:

$$Entropy(S) = \sum_j - p_j \log_2 p_j \qquad (1)$$

If i is discrete value then S_i is corresponding to the probable values of i. Otherwise, if i is continuous value, we choose h in random and split them into h equal domains. Each of them is $S_i^j; j = \overline{1,h}$. h is usually greater than two, then the tree can have more than two branches. After calculating all the $Entropy$, we calculate the $Information\,Gain$ as below:

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \left(|S_v| / |S| \right) \times Entropy(S_v) \qquad (2)$$

We choose an attribute which has maximal value of $Gain$ to a root node. We re-split table S into sub table D_i following the attribute i based on the value of $Gain$. If table D_i has $Entropy(D_i) = 0$, it is the leaf node of the decision tree. Finally, we remove the root node from the data table and repeat the process with other table D_i until $Entropy$ of all tables are zero.

To speed up the tree building process, we divide trees to processors in order to construct in parallel. Assume that we have k processors. We divide all the trees in the forest into k forests (sub-forests) which has the number of trees equally and each processor builds sub-forest separately. After finishing the construction of the forest, each processor will receive the testing data to classify. All the results will be sent to the master processor to gather. The most appearing class among the results will be chosen to the output.

4. EXPERIMENTS

We setup the algorithm using COMGIS system (Le Hoang Son et al., 2011), which was developed to provide some tools for spatial analysis, terrain classification, BTS plans, etc. The experiment is run on the computer with configuration Intel (R) core (TM) i5 CPU M460 (2.53 GHz) and 2GB RAM. Our experiment is carried out by using benchmark GIS data from Bolzano–Bozen, Italy. GIS data included 22 DEM files which were split from a large one. We select 111 regions from 22 DEM files and use the two third of them for the training dataset. The rest of them are used for testing dataset. We compare this algorithm with ID3. Two assessed criterions are classification performance and running time.

Table 1 shows the classification performances and running times of ID3 and sequential RF algorithm. The ID3's classification performance is dissimilar through times because in each time, we choose another region to classify. It fluctuates from 18.92 to 64.86 percent. On the other hand, the fluctuation of RF is smaller, from 54.05 to 70.27 percent. After five times running separately, RF is of much higher classification performance. On average, RF algorithm reaches 62.7% compared with 31.89% of ID3. However, in order to achieve the high performance, the sequential RF algorithm takes more time to solve. It run ten times slower than ID3 does, about 130.8 seconds (s) compared with 10.6s of ID3's running time.

Table 1. Classification performances and running times of ID3 and RF.

Algorithm / Times	ID3		Sequential RF	
	Classification performance (%)	Running time (s)	Classification performance (%)	Running time (s)
1	18.92	10	64.86	186
2	29.73	8	62.16	100
3	27.03	11	62.16	139
4	18.92	14	70.27	99
5	64.86	10	54.05	130
Average	**31.89**	**10.6**	**62.7**	**130.8**

Table 2. Running times comparison between sequential and parallel RF.

Algorithm / Times	Sequential RF (s)	Parallel RF (processors)		
		k = 2	k = 3	k = 4
1	186	133	88	64
2	100	82	72	78
3	139	85	118	86
4	99	91	117	63
5	130	90	85	71
Average	**130.8**	**96.2**	**96**	**72.4**

Therefore, we parallelize the algorithm to reduce running time. Assumed that we have *k* processors to solve simultaneously, we divide to each processor some trees to build. After building the forest, each processor uses data from testing data set to test all its decision trees. Outputs are then sent to the master processor to gather and return the final result. Table 2 indicates the running times of sequential and parallel RF algorithm. Although they are not equal in each time, this parallel algorithm obviously has smaller running time than the sequence. For more details, while the sequence has averagely 130.8s to solve, the parallel with two processors takes 96.2s. The use of two processes is not almost different from the three ones. However, when using four processors, the time running is down to 72.4s. Figure 2 shows the speed-up and efficiency of the parallel RF. Although our algorithm results in best running time with four processors, the best efficiency is with the two. Because the efficiency of two processors is 0.68 and it is higher than that of the three.

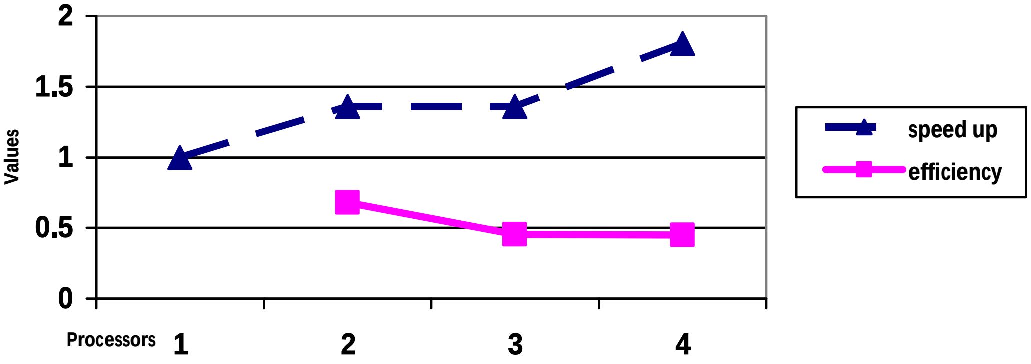
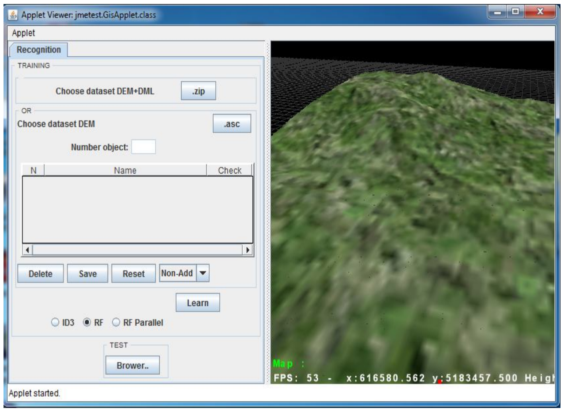


Figure 2: Speed-up & efficiency of parallel RF.

Figure 3 shows our COMGIS system after loading a DEM file. After choosing a region, users need to input training dataset (Figure 4). The system then runs RF algorithm in parallel and displays the classification performance as in Figure 5. Finally, it indicates the type of the input region as in Figure 6.

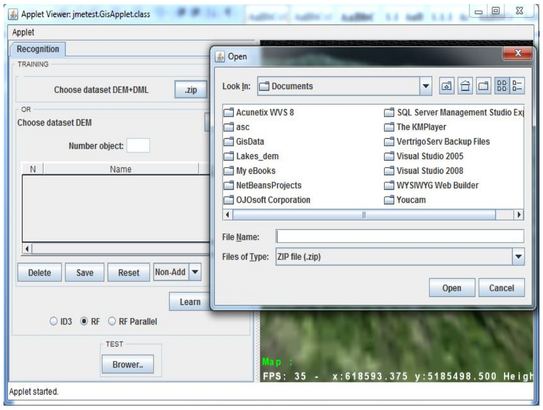


○ ID3 ● RF ○ RF Parallel

FPS: 53 - x:616580.562 y:5183457.500 Heig

Applet started.

Figure 3: COMGIS system.



○ ID3 ● RF ○ RF Parallel

Look In: Documents

Files of Type: ZIP file (.zip)

FPS: 35 - x:618593.375 y:5185498.500 Heigh

Applet started.

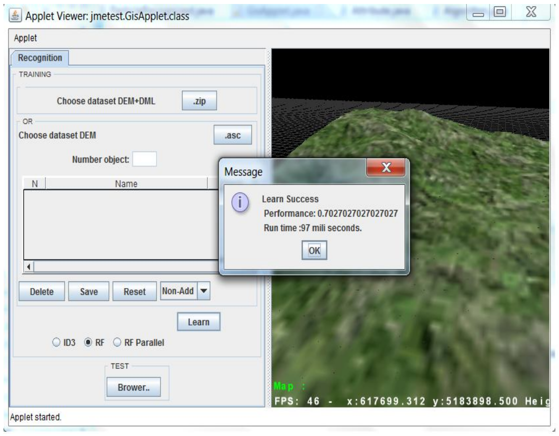
Figure 4: Choosing the training data.

5. CONCLUSION

This paper proposed an innovation of Random Forest using multi-way tree and enabling parallelization for DEM classification. Although our parallel strategy was simple, it contributed to reduce the running time in building the forest. Moreover, parallel implementations will open up possibilities for other experiments with large-scale random forests. A direction for future research is to improve the algorithm in order to distribute equal tasks for each process and enhance the effectiveness in processing.

6. ACKNOWLEDGEMENT

This work is sponsored by a research grant of VNU University of Science [TN-12-48] and VNU Asia Research Centre.

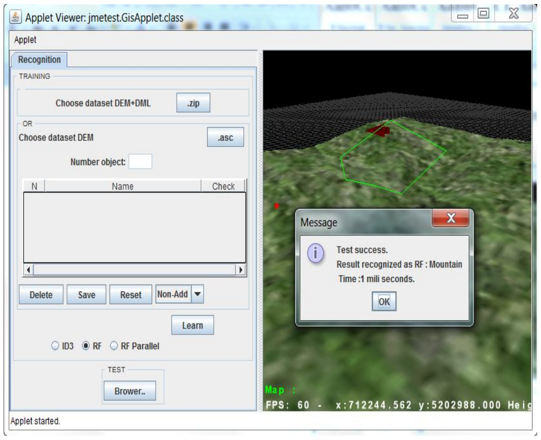


Learn Success
Performance: 0.7027027027027027
Run time :97 mili seconds.

Map :
FPS: 46 - x:617699.312 y:5183898.500 Heig

Applet started.

Figure 5: The learning process.



Test success.
Result recognized as RF : Mountain
Time :1 mili seconds.

Map :
FPS: 60 - x:712244.562 y:5202988.000 Heig

Applet started.

Figure 6: The recognition process.

7. REFERENCES

B.D. Bue, T.F. Stepinski, 2006. Automated classification of landforms on Mars. *Computers & Geosciences* 32 (5), 604-614.

J.R. QuinLand, 1986. Induction of Decision Trees. *Machine Learning* 1(1), 81–106.

Leo Breiman, 2001. Random forests. *Machine Learning* 45(1), 5–32.

Leo Breiman and Adele Cutler, 2002. Random Forests. *http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm*, [accessed on: 3/10/2012].

Li, Z., Zhu, Q. and Gold, 2005. *Digital terrain modeling: principles and methodology*. CRC Press, Boca Raton.

Le Hoang Son, Pham Huy Thong, Nguyen Duy Linh, Truong Chi Cuong and Nguyen Dinh Hoa, 2011. Developing JSG Framework and Applications in COMGIS Project, *International Journal of Computer Information Systems and Industrial Management Applications* 3, 108 – 118.

Ned Horning, 2010, Random Forests : An algorithm for image classification and generation of continuous fields data sets. *International Symposium on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences*, Hanoi.

P.A. Burrough, P.F.M. van Gaans, R.A. MacMillan, 2000. High-resolution landform classification using fuzzy k-means. *Fuzzy Sets and Systems - Special issue on Uncertainty in geographic information systems and spatial data* 113(1), 37–52.

Sitanggang I.S, Yaakob R, Mustapha N, Nuruddin A.A.B, 2011. An extended ID3 decision tree algorithm for spatial data. *International Conference on Spatial Data Mining and Geographical Knowledge Services* (ICSDM), Fuzhou, China, 48 – 53.