International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 3, Issue 3, March 2016

# Spatial Data Analysis for Knowledge Discovery Using Segmentation Based Clustering

Ch. Mallikarjuna Rao
Professor, Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and technology,
Hyderabad-500 090.
chmksharma@yahoo.com

*Abstract:* **Segmentation Based Clustering has been accepted widely as a novel method for analysing the Spatial Data. Many types of Modern Global Positioning Systems (GPS) and also other data acquisition mechanisms are widely used for collecting huge amount of geographical data , which is expected to grow exponentially. It is observed that Mining of such huge data can extractunknown and latent information from spatial datasets that are characterized by complexity, dimensionality and large size. However, it is challenging to do so.Geographical knowledge discovery through spatial data mining has emerged as an attractive field that provides methods to leverage useful applications. Remote sensing imagery is the rich source of geographical data. Analysing such data can provide actionable knowledge for making strategic decisions. This paper proposes a Novel methodology that is used to perform clustering on remote sensing data. Thesedata sets are collected and used World Wind application, provided by NASA. The images are with .TIF extension. The methodology includes feature extraction, training, building classifier and cluster analysis. We built a prototype application that demonstrates the proof of concept. The implementation has taken native method support from Fiji and Weka to realize the proposed methodology. The empirical results revealed that the spatial clustering is made with high accuracy.**

*Index Terms*– **Spatial data mining, remote sensing imagery, clustering, classification, segmentation**

## I. INTRODUCTION

Spatial data sources are becoming rich in geographical data. With such voluminous data, the data sources have implicit patterns that are hidden. Extracting such trends has been made using spatial data mining. Ever since spatial data mining has emerged, it attracted considerable research that focused on different techniquesin mining spatial data [1]. In the domain of data mining there are many clustering algorithms for mining patterns from data. These algorithms are used to group similar objects. Based on clustering many algorithms came into existence that strive to obtain clusters that are characterized by high intra-cluster similarity and low inter-cluster similarity. K-Means [4] is one of the clustering algorithms that are widely used. Another example is Particle Swarm Optimization (PSO) as explored in [8]. All the clustering algorithms must use a similarity measure in order to measure similarity between two objects. The clustering is of two types broadly. Hard clustering and soft clustering are the two categories. In case of hard clustering an object can belong to only one cluster while the soft clustering lets an object to belong to different clusters with certain probability. Thus Fuzzy K-Means came into existence to bring about flexibility in clustering and cluster analysis[2]. However, these algorithms are to be adapted to spatial data mining with required changes. Moreover there needs to be visualization techniques to present results as explored in [3].The spatial data mining deals with geographical data that contains images of different formats. From this it is understandable that some kind of image processing is also involved in the spatial data mining when remote sensing imagery is used for experiments. Having understood this, in this paper we proposed a methodology that combines both image processing and data mining to bring about accurate clustering of spatial objects. We built a methodology that involves extraction of features from .TIF files (satellite images), obtaining training data from input labelling, and building a classifier that can help in clustering of objects in given geographical area. We built an application that makes use of native methods supported by Fiji [20] that can be used to combine image processing and data mining capabilities. The application is based on the proposed methodology. We collected geographical information in the form of .TIF files from World Wind application [21] provided by NASA [22]. Our results reveal that there is high accuracy of clustering of objects in the spatial data which can be used in any real world application for spatial data analysis. Our contributions in this paper are described here. We proposed a methodology for spatial data analysis through clustering. The underlying technique also includes

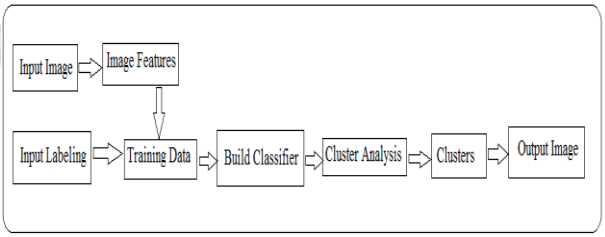
segmentation based classification. We built a prototype application to demonstrate the efficiency of the proposed methodology. The remainder of the paper is structured as follows. Section II provides review of related works. Section III proposes our methodology that is meant for spatial data analysis. Section IV presents experimental results while section V concludes the paper besides providing recommendations for future work.

## II. RELATED WORKS

This section reviews the literature on geographical analysis of spatial data. Stated differently it throws light into the prior works on analyzing satellite images for knowledge discovery. Smith *et al*. [5] proposed Tract Based Spatial Statistics (TBSS) which is meant for enhancing spatial data mining in terms of multi-subject diffusion, interoperability, objectivity and sensitivity. They made many diffusion imaging studies that revealed the utility of such spatial data analysis. Wise *et al*. [6] focused on visualization of spatial data analysis results that could provide user-friendly outputs to end users. They could visualize non-visual information as well for high level of sophistication. Ester *et al*. [7] explored spatial databases that can hold huge amount of spatial data besides providing information on knowledge discovery from such databases. Ng and Han [9] proposed a novel method for clustering spatial data objects. The algorithm was named as CLARANS. The algorithm could handle both points and polygon objects well. Their algorithms could find relationships among both spatial and non-spatial objects. Mary and Ber [10] studied knowledge discovery from spatial and temporal data sources. They employed Markov Chain in order to achieve this. They also used unsupervised learning for extracting trends from spatial datasets. The second order hidden markov model was used in order to obtain temporal segmentation as part of spatial data analysis. Thus they could get hidden information that is in the spatial data sets. Keim *et al*. [11] proposed a pixel based approach for mining geo-spatial data. Moreover they could present results in graphical format for fostering better comprehension. Handl *et al*. [12] focused on validation of clusters in spatial data mining. On the other hand Tung *et al*. [13] explored the spatial data analysis in the presence of obstacles.Yang *et al*. [14] employed multi-feature and multi-scale approach in order to mine information from remote sensing imagery. Their model revealed extraction accuracy and efficient way of information retrieval from geographical data. Soh and Tsatsoulis [15] proposed a segmentation technique that could process satellite images of natural scenes using data mining techniques. Their segmentation process has dynamic local thresholding, extraction of spatial features, conceptual clustering for making final clusters of regions. Other such techniques were explored in [16], [17], [18] and [19] for segmentation, spatial constrained clustering using K-means, spatial data mining for clustering, spatial data mining for high resolution image processing. In this paper we combine image processing and spatial data mining methods in our methodology for clustering spatial objects.

## III. PROPOSED METHODOLOGY

Format is supported for spatial data mining. The raster graphics images that are collected from satellites are best represented in this file format. The World Wind is built and maintained by NASA as part of its complete Earth observation system. The proposed methodology consists of machine learning approach that is used to train a classifier. Once classifier is computed, that will take care of the clustering process. The methodology Image segmentation is widely used technique for pre-processing of satellite images. However our proposed method is segmentation based spatial clustering which needs to combine the best features of image processing and data mining. We take an input image and find the features of given image. The extracted image features and labelling of original image are used to train a classifier. The classifier is then used to perform spatial clustering of objects in the given image. The datasets used for experiments are in TIF format. They are collected from World Wind, a virtual globe application, by choosing any geographical area from the selected area. The datasets or saved in .TIF format which is further used for spatial data mining. For high quality graphics, .TIF file format is widely used. The Tagged Image File (TIF) format is suitable for processing high quality images. Moreover, the TIF is presented in Figure 1.




The proposed approach is based on image segmentation features provided by Fiji, one of the best image processing API in Java. The traditional image processing, data mining and spatial data mining capabilities of Fiji where exploited to realize the proposed framework. The image features supported by Fiji native methods are edge detectors, texture filters, noise reduction filters, membrane detectors, and a host of customized features. The underlying processing in the proposed system is pixel based. The resultant classifier leverages the mining API of Fiji in order to perform clustering of objects in given satellite image through classification. This methodology is implemented by building a prototype application using Java

programming language. The reason behind it is that Fiji API supports Java and thus it becomes cross-platform.

## IV. EXPERIMENTAL RESULTS:

We built a prototype application that demonstrates the usefulness of the proposed methodology in performing spatial clustering. The segmentation based algorithm is used in the process of building a classifier to perform clustering of objects in the given dataset. The datasets were collected from World Wind resources provided by NASA. Five experiments are made using the proposed methodology. In every experiment different .TIF file is given as input and the performance of the proposed method is noted in terms of number of features extracted from image and the time taken to do the same, the number of underlying random forests (trees used for clustering), the time required to build classifier, and the clustering process.

| Experiment # | # of Features | Feature Extraction Time | # of Random Forest Trees | Training Time | Classification Time |
|---|---|---|---|---|---|
| 1 | 79 | 5625 | 200 | 46 | 735 |
| 2 | 79 | 4823 | 200 | 47 | 890 |
| 3 | 79 | 5016 | 200 | 16 | 578 |
| 4 | 79 | 5156 | 200 | 31 | 703 |
| 5 | 76 | 1687 | 200 | 47 | 516 |

***Table 1 –Experimental results***

As can be seen in Table 1, the experimental results reveal the number of extracted features from input images, the time taken for extracting features, the number of random forest trees constructed, the total time consumed for training and the total time required for classification.

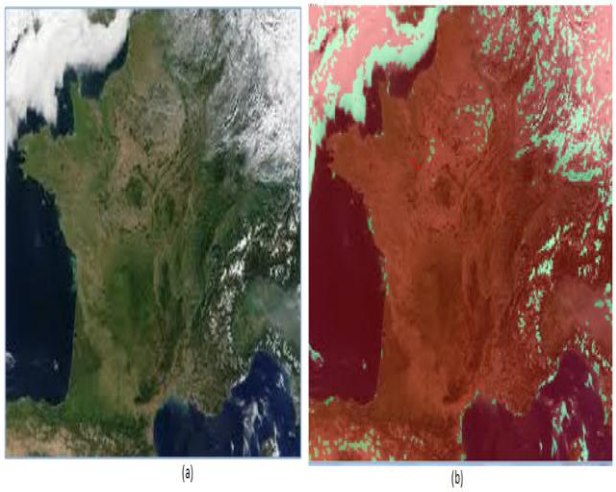



***Figure 2 – Results of first experiment. Original image (a) and resultant image (b)***

As shown in Figure 2, the original image is a .TIF image that is subjected to the proposed methodology. The resultant image shows the clustering of geographical objects based on the outcomes of training phase.

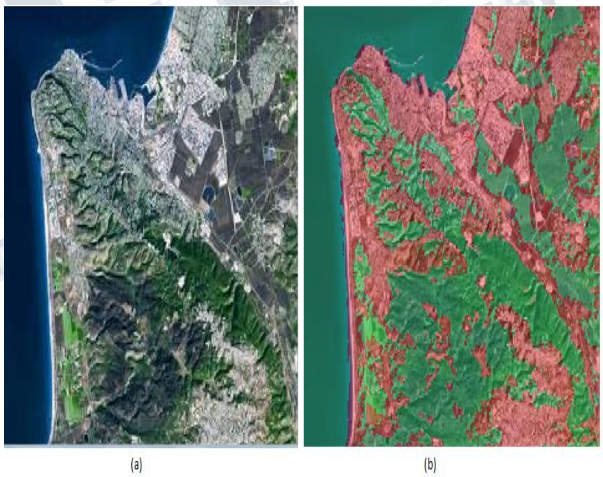



***Figure 3 – Results of second experiment. Original image (a) and resultant image (b).***

As shown in Figure 2, the original image is a satellite image with .TIF extension. It is subjected to the proposed methodology. The resultant image shows the clustering of geographical objects based on the outcomes of training phase.

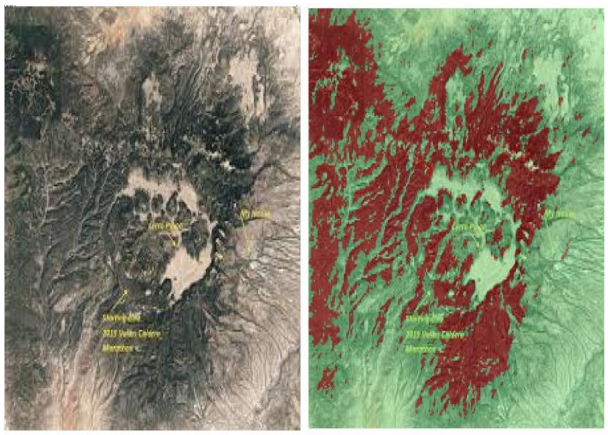


*Figure 4 – Results of second experiment. Original image (a) and resultant image (b).*

As shown in Figure 4, the original image is a satellite image with .TIF extension. It is subjected to the proposed methodology. The resultant image shows the clustering of geographical objects based on the outcomes of training phase.

*Figure 5 – Results of second experiment. Original image (a) and resultant image (b).*

As shown in Figure 5, the original image is a satellite image with .TIF extension. It is subjected to the proposed methodology. The resultant image shows the clustering of geographical objects based on the outcomes of training phase.

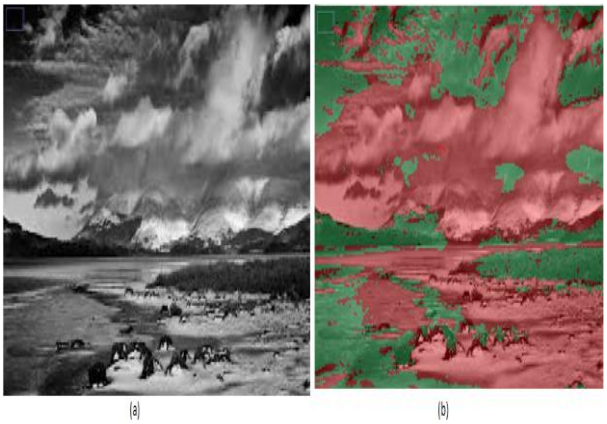


*Figure 6 – Results of second experiment. Original image (a) and resultant image (b).*

As shown in Figure 6, the original image is a satellite image with .TIF extension. It is subjected to the proposed methodology. The resultant image shows the clustering of geographical objects based on the outcomes of training phase.

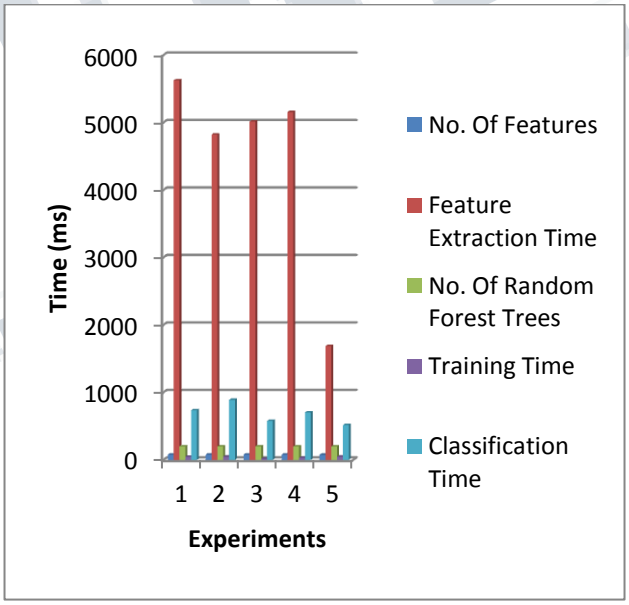



*Figure 7 –Summary of experimental results*

As shown in Figure 7, it is evident that the different experiments have revealed results differently. However, the fifth experiment shows least time for feature extraction and classification time. Another observation is that the classification time is highest for second experiment in which the feature extraction time is relatively less. When number of features is less, there is significant reduction in feature extraction time and classification time but training time is reduced.

## V. CONCLUSION AND FUTURE WORK

In this paper we studied the application of spatial data mining for extracting latent and unknown information from spatial data sets. We used remote sensing images with .TIF extension for experiments. Since they are high dimensional and complex, we combined the image processing and spatial data mining approaches to form a methodology. Our methodology leverages the segmentation based classification in order to perform spatial clustering. We built a prototype application to demonstrate the usefulness of our methodology. We made many experiments with satellite imagery collected from World Wind application provided by NASA. Our experiments revealed that clustering has been made with high accuracy. This can be used in real world applications like verifying forest extent increase or decrease. This research can be extended further to test the methodology for such applications to enhance the quality of geographical analysis.

## AUTHORS

**Dr. Ch.Mallikarjuna Rao** Received his B.Tech degree in computer Science and engineering from Dr. Baba Sahib Ambedkar Marathwada University, Aurangabad, Maharastra, .India in 1998, M.Tech Degree in Computer Science and Engineering from J.N.T.U Anantapuramu Andhrapradesh,India in 2007 and Ph.D degree from JNTU University, Ananthapuramu Andhra Pradesh. India in January 2016. Currently he is working as Professor in the department of Computer Science and Engineering of Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana India.His research interest includes Data bases, data mining and Bigdata.

## REFERENCES


1) Wei Wang, Jiong Yang, and Richard Muntz **(1997).** STING : A StatisticalInformationGrid Approach to Spatial Data**,** p.23-37.

2) JOHANNES GRABMEIER, ANDREAS RUDOLPH(2002). Techniques of Cluster Algorithms in Data Mining. and Knowledge Discovery, p.12-20.

3) Daniel A. Keim, Florian Mansmann, J¨orn Schneidewind, and Hartmut Ziegler(2006).Challenges in Visual Data Analysis.IEEE, p.45-55.

4) ZHEXUE HUANG(1998).Extensions to the Means Algorithm for ClusteringLarge Data Sets with Categorical Values.HUANG p.32-42.

5) Stephen M. Smith, Mark Jenkinson,a Heidi Johansen-Berg, Daniel Rueckert,Thomas E. Nichols, Clare E. Mackay, Kate E. Watkins,Olga Ciccarelli, M. Zaheer Cader,Paul M. Matthews, and Timothy E.J. Behrens(2006). Tract-based spatial statistics:Voxelwise analysis of multi-subject diffusion data. Elsevier, p.25-35.

6) James A. Wise, James J. Thomas, Kelly Pennock, David Lantrip,Marc Pottier, Anne Schur, Vern Crow(1995). Visualizing the Non-Visual: Spatial analysis and interactionwith information from text documents.IEEE, p.45-55.

7) Martin Ester, Hans-Peter Kriegel, Jtirg Sander(1997).Spatial Data Mining: A Database Approach, p.12-19.

8) DW van der Merwe, AP Engelhrecht(2003). Data Clustering using Particle Swarm Optimization, p.34-45.

9) Raymond T. Ng and Jiawei Han(2002). CLARANS: A Method for Clustering Objectsfor Spatial Data Mining.IEEE,p.45-56.

10) **J**.-F. Mari · F. Le Ber(2006). Temporal and spatial data mining with second-orderhidden markov models. Springer-Verlag 2005, p.23-34.

11) Daniel A. Keim, Christian Panse, Mike Sips, Stephen C. North(2004). Pixel based visual data miningof geo-spatial data. Elsevier p.1-17.

12) [12] Julia Handl., Joshua Knowles and Douglas B. Kell(20054). Computational cluster validation in post-genomic data analysis, p.23-33.

13) Anthony K. H. Tung Jean Hou Jiawei Han(2001).Spatial Clustering in the Presence of Obstacles.IEEE, p.34-44.

14) X.M • DYang a, W. Cui b, J.M. Gong a, T. Zhang (2005). INFORMATION MINING FROM REMOTE SENSING IMAGERYBASED ON MULTI-SCALE AND MULTI-FEATURE PROCESSING TECHNIQUES, p.120-135.

15) Leen-Kiat Soh, Costas Tsatsoulis (1999). Segmentation of Satellite Imagery of Natural

16) Scenes Using Data Mining.CSE Journal Articles, p.23-34.

17) Leen-Kiat Soh , Costas Tsatsoulis (1999).Segmentation of Satellite Imagery of NaturalScenes Using Data Mining, p.12-18.

18) Ming Luo, Yu-Fei Ma2, Hong-Jiang Zhang(2001). A Spatial Constrained K-Means Approach to Image Segmentation, p.25-35.

19) Diansheng Guo , Jeremy Mennis(2009). Spatial data mining and geographic knowledge discovery—An introduction. Elsevier, p.12-19.

20) Md Ateeq Ur Rahman, Shaik Rusthum(2009). High Resolution Data Processing for Spatial Image Data Mining.INTERNATIONAL JOURNAL OF GEOMATICS AND GEOSCIENCES, p.25-35.

21) AneliaHackathon.(2015). *Collaboration.* Available: htt p://fiji.sc. Last accessed 12th aug 2015.[21] Patrick. (2011). *National aeronautics and space administration.* Available: http://worldwind.arc.nasa.gov/java/. Last accessed 9th July 2011.

22) Markavenue. (2015). *Soyuz Spacecraft with Three crew Members Aboard Lands Safely at.* Available:https://www.nasa.gov/. Last accessed 01 janu 2015.