International Conference on Advances in Materials, Computing and Communication Technologies
In Association with International Journal of Scientific Research in Science and Technology
Volume 9 | Issue 1 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)

An Approach to Spatial Data Classification Using Dictionary Based Sequence Classifier

P.D.Sheena Smart[1], Dr.K.K.Thanammal[2]

[1]Department of Computer Science, S.T.Hindu College, Nagercoil, Affiliated to Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli, 627012, India

[2]Associate Professor, Department of Computer Science, S.T.Hindu College, Nagercoil, Affiliated to Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli, 627012, India

ABSTRACT

Data mining and soft computing techniques have been developed to an extend that it becomes possible to automatically mine knowledge from spatial data. Classification problem is prevailing in various disciplines. Developing effective classifier is more challenging for researchers. It is not possible for a single classifier to be highly effective to work with all types of datasets. Therefore classifiers vary based on data distribution. In this paper we propose a classifier called Dictionary based Sequence Classifier (DBSC) technique which classifies the spatial data.this technique classifies the datas from a weather dataset. This method first extracts the features from learned dictionary. Then the attributes are sorted and objects classified using a sequence classifier.

Keywords- Data mining, soft computing, classifier, dictionary

I. INTRODUCTION

Spatial data mining is the process of determining the unknown from spatial datasets. With the development of data mining techniques, it becomes possible to automatically mine information from spatial data. In spatial data, spatial rule extraction with uncertainty is a significant concern in spatial data mining. However, that the information is not regularly distributed. A dataset with the physical location specification such as place, geographic synchronize, landmarks, country are theoretical which used in research work for extracting the relevant information. The multi-class labels are also used to determine the mapping output and determine the uncertainty and uncertainty in positioning map. The several spatial data mining techniques were developed in order to analysis the spatial information. Spatial data mining techniques are mostly applied to extract significance and usual information from large spatial databases. This information is also used to provide the accurate result for spatial and non spatial data and their relationships. This information is essential in Geographic Information Systems (GIS), image processing, remote sensing and so on. Based on, the spatial data classification is a significant in data mining.

Rough-fuzzy set based rule extraction model was introduced in [1] measure the accuracy of the fuzzy decisions on unseen spatial objects. Rough Fuzzy set handled fuzziness and irregularity simultaneously in spatial data tables but spatial correlation information is not considered into account. An Automatic non-parametric cooperative segmentation technique was developed in [2] to create and manage the complexity on spatial object primitives with Very High Spatial Resolution (VHSR) satellite images. However, high spectral satellite images do not improve the classification accuracy on Object-Based Image Analysis (OBIA).

A Spectral un mixing and unsupervised classification method on hyper spectral images was developed in [3] integrate the spectral and spatial information for resolution enhancement. Spatial regularization is coupled with sub-pixels from a spatial viewpoint but discrimination on pure and mixed pixel does not produce the better performance result.

Data classification was performed using local spatial information in [4] to automatically discover wide spatial sparse features of PolSAR data and improve the classification accuracy. In [5] a new algorithm was designed for clustering in spatial data mining process. However, it does not considered for neighborhood data objects. Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) for benthic habitat mapping was introduced in [6] goes to the indiscriminate selection of the group-level and code-level sample spatial objects but classified the maps on relatively poor spatial data with minimal accuracy rate.

A spectral–spatial classification strategy was introduced in [7] based on Deep Belief Network to provides highest classification accuracy. However, an effective spatial information extraction is difficult issue. An effective technique was developed in [8] to automatically build the extended attribute profiles with the standard deviation attribute for spectral spatial classification on remote sensing data. An advanced spectral-spatial classification method was introduced in [9] for classification of hyper spectral images, which combines advances of region-based segmentation and image fusion. In [10], spatial data mining method was designed using cluster analysis. A spatial database stores the spatial data symbolized through the spatial data types and spatial connections.

This paper is organized as below: Section II depicts Related works Section III explains proposed classification technique, Section IV evaluates experimental setup, Section V presents comparisons and Section VI depicts conclusion of paper.

II. RELATED WORKS

A framework for spatial co location pattern mining was described in [11] for ubiquitous GIS. However, it does not address the spatial data classification. A new classification method was introduced in [12] for spatial resolution remote sensing image depends on a strategic method of spatial mapping and reclassification and it also improves the classification accuracy. A novel feature extraction method was developed in [13] for utilization of spectral information and spatial knowledge of hyper spectral images with local covariance matrices.

In [14], a new low-rank decomposition spectral–spatial approach was developed to improve the classification accuracy. However, the neighborhood and nonlocal neighborhood information for low-rank representation were not discovered. An effective sparse discrimination analysis (SDA) was used in [15] to obtain the spectral and spatial information from hyper spectral data for improving the classification and prediction accuracies.

Discriminative dictionary learning was introduced in [16] to recognize different weather conditions. In [17] a sequence classifier approach was proposed to

normalize attributes, sort and rank, and create a sequence of numbers.

III. DICTIONARY BASED SEQUENCE CLASSIFICATION TECHNIQUE

Spatial data mining techniques are mostly applied to extract significance and usual information from large spatial databases. This information is also used to provide the accurate result for spatial and non spatial data and their relationships. This information is essential in Geographic Information Systems (GIS), image processing, remote sensing and so on. Based on, the spatial data classification is a significant in data mining. Spatial data is categorized into different classes or effective usage of data. classification of data is grouping and arranging the data under various classes. It is easy to locate and analyze a data in a dataset using the classification technique.

Our proposed method at first extracts the visible features from the blue region and physical characteristics from the non blue region. To the extracted features we apply a Discriminative Dictionary Learning method instead of conventional methods like SVM, KNN.

The weather conditions can be classified as sunny, cloudy, misty, fog and humidity. From the architecture of the DDL mehod we can see that the image dataset is given as input to the feature selection phase. The training dataset consists of the labeled and unlabelled data. The labeled dataset is used to train the initial dictionary. the sample data are chosen repeatedly from unlabelled dataset based on the factors informative and representative. The labeled dataset is expanded by adding the selected samples to learn dictionary.

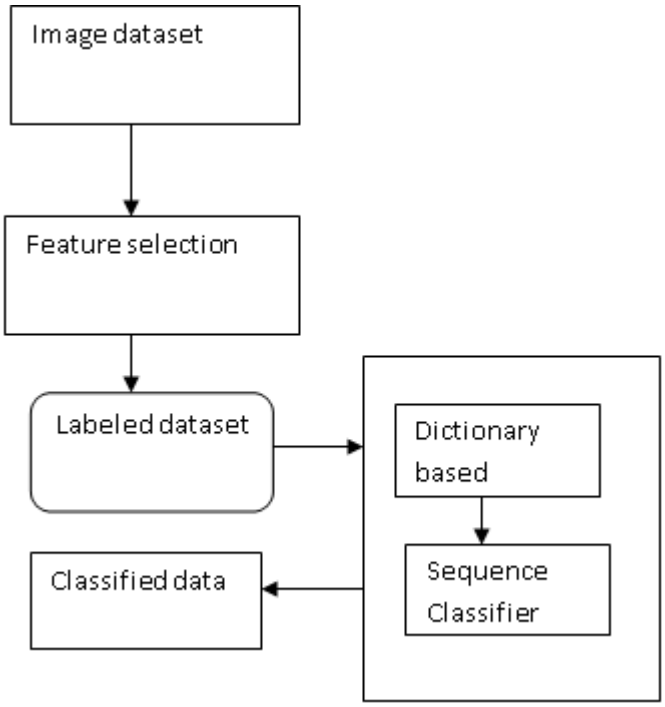

Figure 1. Architecture of DBSC classification technique

A. Feature Extraction

The data of same scene taken under different weather conditions are collected and those which describe the change in weather conditions are analyzed. The images that have the appearances of different visuals depict the properties visually with their physical characteristics. When describing the physical manifestation, we can take the blue sky as a significant feature that plays an important role in finding the weather conditions. If the weather is sunny, the sunlight refracts when passing through he atmosphere and as a result of this, the blue light is scattered and hence, sunlight refracts when passing through it.

Another feature cloudy is a condition that appears in-between sunny and gloomy. When it is cloudy, the clouds float in the sky. The SIFT,HSV ,LBP and the sky parts are extracted. These extracted attributes describe the color, shape and pattern. In [16], the nonsky areas were discarded. The non sky areas of images are not discarded directly. Instead two

attributes based on their features are extracted so that they act as powerful attributes. Bad weather conditions causes the degradation in the contrast of the image. So on a particular scene, contrast must be different when taken in different weather conditions.

Gloomy weather is accompanied by mist. The observation made here is by identifying patches without mist in the non blue areas and it must be with less intensity and atleast anyone of the RGB channel.

The dim channel,

$$S^{dim}(ai) = \min_{c\in\{r,g,b\}}(\min_{i\in\Omega(i)} S^c(i^1))) \quad (1)$$

Where Sc is a color of the images and Ω(i) is the patch.

B. Object Classification

Object classification is done to improve the classification power of the trained set so that it identifies various weather conditions and as a result the training dataset gets increased. Consider a set of I classes denoted by [a1,a2,a3,…ai).the factors informative and representative are used in the feature selection. Informtive is a criteria that chooses only the helpful samples of datas so that it reduces uncertainity. It is measured as

$$I_n(ap) = Er\_S_p + En_p \quad (2)$$

where Er_S_p Represent the reconstructed error with respect to the present dictionary and r represents entropy.

$$Er\_S_p = {}^{min}_{i}||a_p - EiXia_p||^2 \quad (3)$$

Where Ei and Xi denote sub dictionary pairs in a class i. when the error is more. Er_S_p represents that the present dictionary does not describe the sample as well. Similarly, the probability of reconstructed error of image a_p in class i is represented as,

$$RP_i(a_p) = \frac{||a_p - EiXia_p||^2}{\sum_{i=1}^{I}||a_p - EiXia_p||^2} \quad (4)$$

Where, RP is the probability for sample which is calculated as X(=[x1,x2,…xi], which describes the difference between the dictionary and the input . the entropy is estimated as follows:

$$En_p = \sum_{i=1}^{I}(RP_i(a_p)\log RP_i(a_p)) \quad (5)$$

If the value of entropy is high, it indicates that tha sample image is more complex to be classified.

Another factor to be considered is the representative attribute, which acts as an added criterion to extract the useful unlabelled data. As the informative factor takes into account how the sample images are related to the classification model, it discards the structure of the unlabelled dataset.

The unlabelled data that is distributed is helps in training for a good classification technique.

It is calculated by

$$R(a_p) = P(a_p) - P(a_p|I(u_p) = \tfrac{1}{2}\ln\left(\frac{\sigma_{p^2}}{\sigma_{p^2}\,|Up}\right) \quad (6)$$

Algorithm 1. Algorithm for DBSC technique

Input: labeled and unlabelled datasets Li and Lu respectively

Output: classified objects

Begin

Step 1 : initialize

Step 2 : for i=1 to n do

Step 3 : compute In(a_p) and R(a_p) in unlabelled dataset L_u

Step 4 : Extract samples 'n'

Step 5 : Learn updated dictionary D

Step 6 : end for

Step 7 : learned pair of data extracted

Step 8 : Sort the values in ascending order

Step 9 : Provide ranking to normalized values

Step 10 : Apply threshold to perform reduction

End

C. Sequence Classifier

The spectral order of HIS data is not do informative. So it is necessary to re arrange the bands to produce a new band order. Every attribute that is extracted is normalized and arranged in order. The sorted features are then provided ranking with the help of the location without any repletion. This rank provides a strong collaboration between the features.

As the relations are enhanced after normalization. From this we can understand that after normalization the dataset is more informative. Consider an attribute matrix P , which consists of 'x' feature vectors each with 'n' features. Vector 'I' represents 'x' number of classes mentioned as x1,x2,x3,…,xn.

$$P = \begin{matrix} p11 & p12 \cdots & p1n \\ p21 & p22 \cdots & p2n \\ px1 & px2 \cdots & pxn \end{matrix} \quad (7)$$

$$I = i1, i2, \ldots., im$$

Sorting in data is performed each row by row for every feature and ranking is assigned to ech attribute on their sorting result. The dictionary Di is assigned with normalization values, rank value and class labels., which are estimated from P and I. the whole dataset in a dictionary is represented by x instances .Dictionary Di is represented as follows.

$$Di = \begin{matrix} p111 & p121 & p131 \cdots & p1n1 \\ p211 & p221 & p231 \cdots & p2n1 \\ p112 & p122 & p132 \cdots & p1n2 \\ p21i & p22i & p23i \cdots & p2ni \end{matrix} \quad (8)$$

Where denotes ath sample with feature b and class c. After the dictionaries are created they are split into training data and test data. Redundant data must be remove from the dictionary to enhance computational efficiency. Similarity measurement for each entry in the dictionary to detect repeated data. Again, a threshold value is applied to the data to perform reduction.

A sample dataset SD is arranged and undergone ranking to get SDr. Sometimes, more than one class may be generated and therefore Tie conditions occur. During such situation, voting is carried on for class data. If there is no tie available, then the sample data that is currently under test is classified as class c. But if there is a tie, then the dictionary D with the samples and reference to locations are selected. Therefore, the DBSC procedure improves the performance of spatial data classification result with the help of dictionary learning to improve the classification accuracy rate.

IV. EXPERIMENTAL EVALUATION

In this paper, an DBSC Procedure is implemented using java language with El Nino dataset[18] extracted from UCI repository. El Nino dataset consists of oceanographic and surface meteorological analysis from a succession of buoys in equatorial Pacific. There are 178080 instances and attributes are 12. The results are compared with rough fuzzy sets model [1] and Automatic non-parametric cooperative segmentation [2].

Table 1. Spatial Attributes Information using El Nino dataset

Attribute	Description
date	Date of the readings taken
Latitude	Values stayed within a degree from the approximate location
Longitude	Values were sometimes as far as five degrees off of the approximate location.
Zonal winds (west<0, east>0)	zonal winds are fluctuated between -10 m/s and 10 m/s
Meridional winds (south<0, north>0)	Meridional winds fluctuated between -10 m/s and 10 m/s
Relative	Relative humidity values in the

humidity	tropical Pacific were typically between 70% and 90%
Air temperature	Air temperature fluctuated between 20 and 30 degrees Celsius
Sea surface temperature	Sea surface temperature fluctuated between 20 and 30 degrees Celsius
Subsurface temperatures	Subsurface temperatures down to a depth of 500 meters

V. COMPARISON AND DISCUSSIONS

For experimental analysis, the data classification using different techniques is compared with number of spatial data. Parameters like classification accuracy and classification time are used for comparison.

A. Classification Accuracy

It is defined as the ratio of number of data that are accurately classified to total number of data and its measurement is in percentage (%).

$$\text{Classification Accuracy} = \frac{\text{No. of spatial data correctly classified}}{\text{Total number of data}} \quad (9)$$

From (9), the classification accuracy is estimated. If the classification accuracy is high, then the technique is said to be efficient. Table 1 depicts the performance of classification accuracy with respect to number of spatial data ranging from 10 to 70.

Table 2. TABULATION FOR CLASSIFICATION ACCURACY

Number of spatial data	Classification Accuracy (%)		
	Automatic non-parametric cooperative segmentation technique	Rough fuzzy sets	DBSC
10	65.85	73.25	76.61
20	67.62	74.67	79.45
30	70.2	75.13	83.22
40	71.35	78.35	86.67
50	72.38	80.36	90.55
60	74.35	81.2	93.97
70	76.98	82.24	92.73

From the above table, it is obvious that, the classification accuracy do not increase in a linear way as the number of spatial data increase.

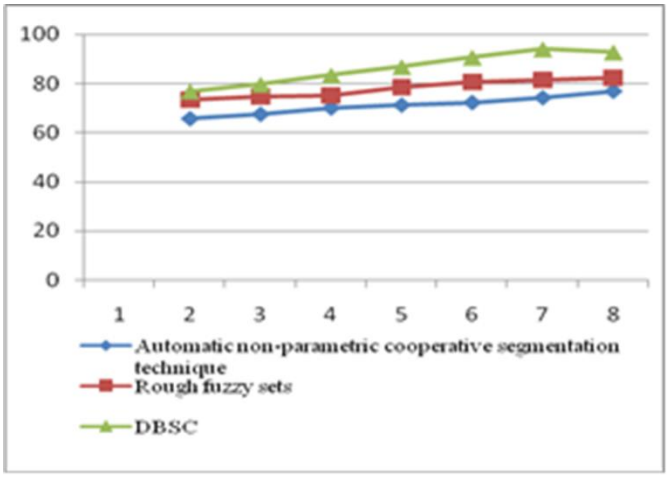

Figure. 2. Measure of Classification Accuracy

Figure 2 describes the classification accuracy performance of three different techniques with respect to the number of spatial data.

Classification accuracy of rough fuzzy sets model is comparatively higher than that of Automatic non-parametric cooperative segmentation. DBSC extracts the spectral information from HIS dataset. The essential spatial data in dissimilar sizes of identical areas is taken out through multi-scale feature extraction. Research in DBSC framework produces 15% higher classification accuracy than Automatic non-parametric cooperative segmentation Architecture and 9% higher classification accuracy than rough fuzzy sets model.

B. Classification Time (CT)

Classification time is the measure between the starting time and ending time of spatial data processing and measured in milliseconds (ms).,

CT=End time-Star time of classification (10)

From (10), the classification time is calculated. When classification time is minimal, the method is more effective.

Table 3. TABULATION FOR CLASSIFICATION TIME

Number of spatial data	Classification Time(ms)		
	Automatic non-parametric cooperative segmentation technique	Rough fuzzy sets	DBSC
10	16.7	15.4	12.1
20	21.6	17.6	13.5
30	22.5	18.2	16.8
40	24.1	20.7	18.4
50	25.9	21.6	19.4
60	30.2	23.8	20.5
70	31.5	24.7	22.6

Table 3 explains the results of classification time with respect to number of spatial data ranging from 10 to 70. Classification time comparison takes place on existing rough fuzzy sets model and Automatic non-parametric cooperative segmentation. From the table, it is observed that, the classification time does not increase linearly when the spatial data increase.

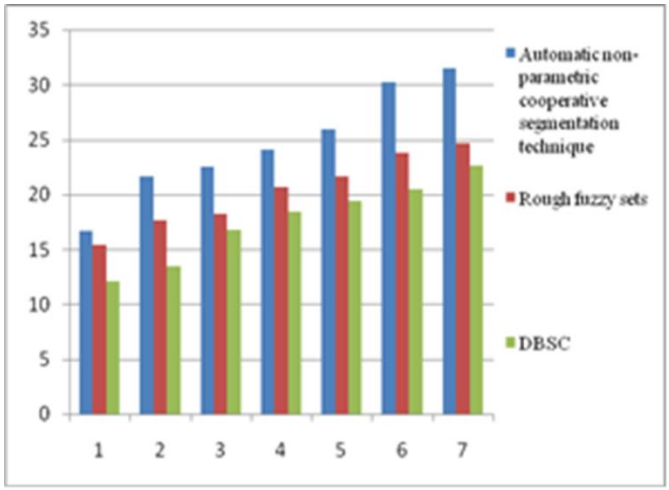

Figure 3. Measure of Classification Time

Figure 3 explains the classification time performance of three different techniques.

Classification time of DBSC is comparatively lesser than that of rough fuzzy sets model and Automatic non-parametric cooperative segmentation framework. Privileged order statistics methods are used in hyperspectral categorization to classify phase correlation of spectral curves.

VI. CONCLUSION

A novel Dictionary based Sequence Classifier (DBSC) technique is developed to enhance the spatial data mining efficiency through classification. The goal of the DBSC Technique is acquired with the application of the Distributed Discriminative Learning method and Sequence Classifier. Also , DBSC Technique takes minimum time to classify the spatial data as compared to existing works. Thus, DBSC Technique enhances the performance of spatial data mining as compared to existing works.

The efficiency of the DBSC Technique is estimated in terms of classification accuracy, time complexity and compared with two existing methods. The results demonstrate that DBSC Technique performed better with an enhancement of classification accuracy and reduction of time.

VII. REFERENCES

[1]. Hexiang Bai., Yong Ge., Jinfeng Wangb., Deyu Li., Yilan Liao., Xiaoying Zheng., "A method for extracting rules from spatial data based on rough fuzzy sets," Knowledge-Based Systems, Elsevier Journal., 2014

[2]. Imane Sebari and Dong-Chen He., "Automatic fuzzy object-based analysis of VHSR images for urban objects extraction," ISPRS Journal of Photo grammetry and Remote Sensing., Elsevier Journal., 2013

[3]. Villa a,b,n,1, J. Chanussot c, J.A. Benediktsson d, C. Jutten c, R. Dambreville., "Unsupervised methods for the classification of hyperspectral

images with low spatial resolution," Pattern Recognition., Elsevier Journal., 2013

[4]. Lu Zhang,Wenping Ma, and Dan Zhang, "Stacked Sparse Autoencoder in PolSAR Data Classification Using Local Spatial Information", IEEE Geosciences and Remote Sensing Letters, Volume 13, Issue 9, September 2016

[5]. Arvind Sharma, R. K. Gupta, and Akhilesh Tiwari, "Improved Density Based Spatial Clustering of Applications of Noise Clustering Algorithm for Knowledge Discovery in Spatial Data", Mathematical Problems in Engineering, Hindawi Publishing Corporation, Volume 2016, June 2016, Pages 1-9

[6]. Caiyun Zhang., Donna Selch., Zhixiao Xie., Charles Roberts., Hannah Cooper., Ge Chen., "Object-based benthic habitat mapping in the Florida Keys from hyperspectral imagery," Estuarine, Coastal and Shelf Science., Estuarine, Coastal and Shelf Science, Volume 134, 2013 Pages 88-97

[7]. Yushi Chen, Xing Zhao, and Xiuping Jia, "Spectral–Spatial Classification of Hyperspectral Data Based on Deep Belief Network", IEEE Journal of selected Topics in Applied Earth Observations and Remote Sensing, Volume 8, Issue 6, June 2015

[8]. Prashanth R. Marpu, Mattia Pedergnana, Mauro Dalla Mura, Jon Atli Benediktsson,n and Lorenzo Bruzzone, "Automatic Generation of Standard Deviation Attribute Profiles for Spectral–Spatial Classification of Remote Sensing Data", IEEE Geosciences and Remote Sensing Letters, Volume 10, Issue 2, March 2013, Pages 293-297

[9]. Zelang Miao andWenzhong Shi, "A New Methodology for Spectral-Spatial Classification of Hyperspectral Images", Hindawi Publishing Corporation, Journal of Sensors, Volume 2016, February 2015, Pages 1-12

[10]. Ch.N.Santhosh Kumar, V. Sitha Ramulu,K.Sudheer Reddy, Suresh Kotha and Ch.Mohan Kumar, "Spatial Data Mining using Cluster Analysis", International Journal of Computer Science & Information Technology (IJCSIT), Volume 4, Issue 4, Pages 71-77, August 2012

[11]. Seung Kwan Kim, Jee Hyung Lee, Keun Ho Ryu, Ungmo Kim, "A framework of spatial co-location pattern mining for ubiquitous GIS", Multimedia Tools and Applications, Springer, Volume 71, Issue 1, Pages199-218, July 2014,

[12]. Guizhou Wang, Jianbo Liu, and Guojin He, "Method of Spatial Mapping and Reclassification for High-Spatial-Resolution Remote Sensing Image Classification", The ScientificWorld Journal, Hindawi Publishing Corporation, Volume 2013, November 2013, Pages 1-7

[13]. Ugur Ergul, Gokhan Bilgin, "Integration of spectral and spatial information via local covariance matrices for segmentation and classification of hyperspectral images" Turkish Journal of Electrical Engineering & Computer Sciences, 2015, Pages 1-15

[14]. Yang Xu, ZebinWu, and Zhihui Wei, "Spectral–Spatial Classification of Hyperspectral Image Based on Low-Rank Decomposition" IEEE Journal of Selected Topics in Applied earth Observations and Remote Sensing, Volume 8, Issue 6, June 2015,

[15]. Learning for Weather Recognition", Hindawi Publishing Corporation Mathematical Problems in Engineering ,Volume 2016, 12 pages

[16]. Ram Narayan Patroa , Subhashree Subudhia et al., "Dictionary-based classifiers for exploiting feature sequence information and their application to hyperspectral remotely sensed data", International Journal of Remote Sensing, March 8, 2019

[17]. [dataset] El Nino dataset: UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/El+Nino,