Integrating Spatial Data Mining Technique to Identify Potential Landsat Data using K-Means and BPNN Algorithm

N.Naga Saranya
Research Scholar (C.S),
Karpagam University,
Coimbatore, Tamilnadu – 641021.

M.Hemalatha
Head, Software System,
Karpagam University
Coimbatore, Tamilnadu – 641021.

ABSTRACT
Spatial Data mining is one of the challenging field in data mining. The explosive development of spatial data and common use of spatial databases highlight the need for the automated detection of spatial knowledge. Computing data mining algorithms such as clustering on massive spatial data sets is still not feasible nor efficient today. In this research first we elaborate a study on data clustering, particularly on spatial data clustering. Here we introduce a k-means algorithm that is based on the data stream paradigm. Some of the existing classical clustering algorithm and the proposed BPNN were tested with UCI repository datasets for spatial data clustering and classification. Several tests were made on the system and overall significant results were achieved. Proposed method is an influential tool for the classification of multidimensional spatial data sets.

Keywords
Spatial Clustering Techniques, Spatial Land sat Data, Artificial Neural Network (ANN).

1. INTRODUCTION
The Landsat satellite data is one of the many sources of information available for a scene. The interpretation of a scene by integrating spatial data of diverse types and resolutions including multispectral and radar data, maps indicating topography, land use etc. is expected to assume significant importance with the onset of an era characterized by integrative approaches to remote sensing. Clustering of data is a difficult problem that is related to various fields and applications. Challenge is greater, as input space dimensions become larger and feature scales are different from each other. Existing statistical methods are ill-equipped for handling such diverse data types such as spatial data. Note that this is not true for Landsat MSS data considered in isolation (as in this sample database). This data satisfies the important requirements of being numerical and at a single resolution, and standard maximum-likelihood classification performs very well. Consequently, for this data, it should be interesting to compare the performance of other methods against the statistical approach. Major drawbacks have to be tackled, such as curse of dimensionality and initial error propagation, as well as complexity and data set size issues. So in this research, we are going to study the performance of some of the classical clustering algorithm for spatial data clustering and will derive a hybrid model for better spatial data clustering.

2. CLUSTERING APPROACHES IN SPATIAL DATA MINING
Classification algorithms rely on human supervision to train it to classify data into pre-defined categorical classes. For example, given classes of patients that corresponds to medical treatment responses; identify most responsive forms of treatment for the patient.
There are so many methods for data classification. Generally the selection of a particular method may depend on the application. The selection of a particular methodology for data classification may depend on the volume of data and the number of classes present in that data. Further, the classification algorithms are designed in a custom manner for a specific purpose to solve a particular classification scenario.

3. RECENT WORKS IN SPATIAL DATA CLUSTERING
Kharin. Y. S, zhuk. E. E. [3], were investigated the problem of cluster analysis of discrete (multinomial) random observations, assuming the presence of outliers in the sample. They further proposed a robust decision rule based on the truncation principle and demonstrated that the robust algorithm essentially improves the clustering performance approximately twofold. MATHER. L. A. [4], an application of linear algebra to text clustering, a metric for measuring cluster quality was described. The metric was based on the theory that cluster quality is proportional to the number of terms that are disjoint across the clusters.
Vaithyanathan [5], presented an approach to model-based hierarchical clustering by formulating an objective function based on a Bayesian analysis.
Polanco. X [6], suggested using artificial neural networks for mapping of science and technology as a multi-self-organizing maps approach. They proposed the Kohonen self-organizing map (SOM) for clustering and mapping according to a multi maps extension.
Clustering of spatial data using random walks was done by Koren.Y. and Harel. D. They argued that discovering significant patterns that exist implicitly in huge spatial databases is an important computational task, a common approach to this problem is to use cluster analysis.
Cluster-Rasch models for micro array gene expression data were developed by Li, Hongzhe And Hong, Fangxin [7].
Dudoit [8], devised a prediction-based resampling method for estimating the number of clusters in a dataset. Micro array technology is increasingly being applied in biological and

medical research to address a wide range of problems, such as the classification of tumors.

The clustering property of corner transformation for spatial database applications was designed by Ju-Won Song [9], Spatial access methods (SAMs) are often used as clustering indexes in spatial database systems.

Vishwanathan [10], worked on Kernel enabled K-means algorithm. They presented a novel method to learn arbitrary cluster boundaries by extending the k-means algorithm to use Mercer kernels.

Chen, c [11], conducted two case studies for visualizing and tracking the growth of competing paradigms. They demonstrated the use of an integrative approach to visualize and track the development of scientific paradigms..

Paulo Gonçalves [13], the goal is to achieve an automatic pixel level classification using a Support Vector Machine (SVM) learning approach.

Nana Liu [15], With more applications of multispectral remote sensing images, how to effectively and correctly make automated classification of multispectral images is still a great challenge..

An image segmentation system is proposed for the segmentation of color image based on neural networks, G. Dong & M. Xie[16].

Xueping Zhang [18], Spatial clustering with obstacles constraints (SCOC) has been a new topic in spatial data mining (SDM).

Spatial image mining for soil classification using diversified domains like Digital Image Processing, Neural Networks, and Soil fundamentals [19]. The three most important algorithms used in implementation are Back Propagation Network (BPN), Adaptive Resonance Theory 1 (ART) and Simplified Fuzzy ARTMAP for soil classification as well as spatial image recognition.

4. PROPOSED SYSTEM'S METHODOLOGY AND DESIGN

In general, clustering methods may be divided into two categories based on the cluster structure, which they produce. The non-hierarchical methods divide a dataset of N objects into M clusters, with or without overlap. These methods are sometimes divided into partitioning methods, in which the classes are mutually exclusive, and the less common clumping method, in which overlap is allowed. Each object is a member of the cluster with which it is most similar, however the threshold of similarity has to be defined. The hierarchical methods produce a set of nested clusters in which each pair of objects or clusters is progressively nested in a larger cluster until only one cluster remains. The hierarchical methods can be further divided into agglomerative or divisive methods.

4.1 The Normal K-Mean Algorithm

K-Means algorithm is very popular for data clustering. The Algorithm goes like this

Step1: Select k Center in the problem space (it can be random).

Step2: Partition the data into k clusters by grouping points that are closest to those k centers.

Step3: Use the mean of these k clusters to find new centers.

Step4: Repeat steps 2 and 3 until centers do not change.

This algorithm normally converges in short iterations.

4.2 Proposed Neural Network Architecture

The following diagram illustrates the proposed multi-layer neural network design which is going to be used in this project [31]. An elementary neuron with *R* inputs is shown in Fig 2. Each input is weighted with an appropriate *w*. The sum of the weighted inputs and the bias forms the input to the transfer function *f*. Neurons can use any differentiable transfer function *f* to generate their output.

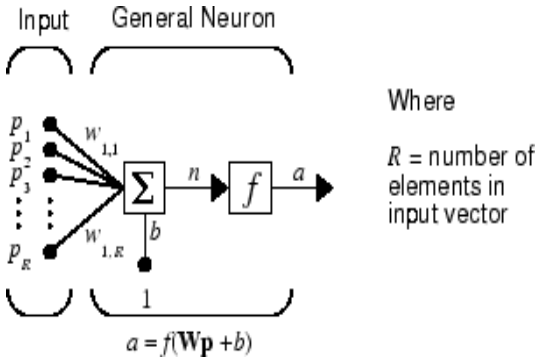

Figure 1: Multi Layer Neural Network Architecture

An actual algorithm for a 3-layer network:

```
#Create the network with 4 inputs, 1 hidden layer with 4 neurons, and 2 outputs
net =Ai4r:: NeuralNetwork::Backpropagation.new([4, 5, 2])
      # Train the network
    2000. times do | i |
          net.train(example[i], result[i])
end
```

4.3 Proposed System Design

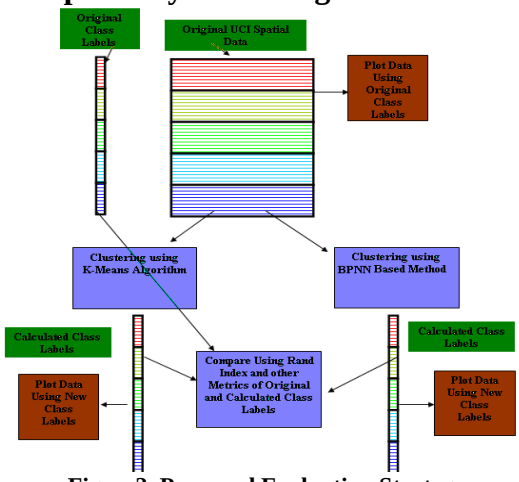

Figure2. Proposed Evaluation Strategy

5. IMPLEMENTATION AND RESULTS

The proposed system has been implemented and the performance of the classification algorithm was tested with the spatial dataset called "UCI Landsat Multi-Spectral dataset".

5.1 About the UCI Datasets

The database consists of the multi-spectral values of pixels in 3x3 neighborhoods in a satellite image, and the classification associated with the central pixel in each neighborhood. The aim

is to predict this classification, given the multi-spectral values. In the sample database, the class of a pixel is coded as a number. This database was generated from Landsat Multi-Spectral Scanner image data. These and other forms of remotely sensed imagery can be purchased at a price from relevant governmental authorities. The data is usually in binary form, and distributed on magnetic tape(s).

5.1.1 Main Interface

The following interface was created for altering various parameters during the evaluation of the algorithm.

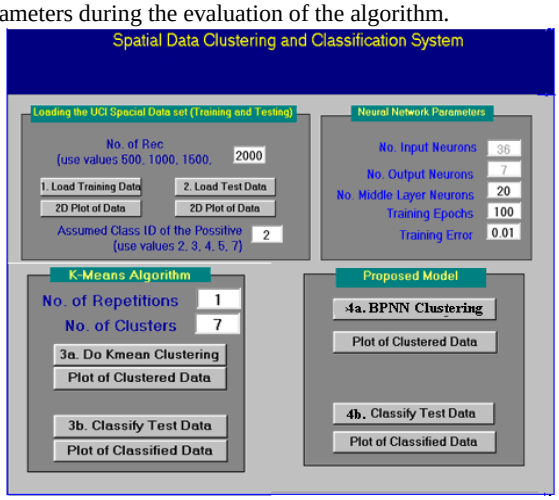

Figure 3: Main Interface Design

In Fig 3, the buttons labeled 1 and 2 is used to load the UCI spatial training and testing data. Buttons labeled 3a, and 4a were used to cluster the data with different methods. Buttons labeled 3b and 4b were used to classify the data using the previously trained clusters with different methods.

5.1.2 Description of the Data

One frame of Landsat MSS imagery consists of four digital images of the same scene in different spectral bands. Two of these are in the visible region (corresponding approximately to green and red regions of the visible spectrum) and two are in the (near) infra-red. Each pixel is a 8-bit binary word, with 0 corresponding to black and 255 to white. The spatial resolution of a pixel is about 80m x 80m. Each image contains 2340 x 3380 such pixels. The database is a (tiny) sub-area of a scene, consisting of 82 x 100 pixels. Each line of data corresponds to a 3x3 square neighborhood of pixels completely contained within the 82x100 sub-area. Each line contains the pixel values in the four spectral bands (converted to ASCII) of each of the 9 pixels in the 3x3 neighborhood and a number indicating the classification label of the central pixel.

The number is a code for the following classes:

Number	Class
1	red soil
2	cotton crop
3	grey soil
4	damp grey soil
5	soil with vegetation stubble
6	mixture class (all types present)
7	very damp grey soil

There are no examples with class 6 in this dataset.

The data is given in random order and certain lines of data have been removed so you cannot reconstruct the original image from this dataset.

Number of Examples In the dataset

Training set 4435
Test set 2000
Number of Attributes in each record
36 (= 4 spectral bands x 9 pixels in neighborhood)

Sample Data:

For Example, the following two records belong to class 4 and 3.
68 94 94 79 76 94 111 79 80 98 106 83 71 83 87 70 76 91 91 74 76 95 104 81 67 75 85 71 67 75 96 79 75 83 96 83 4
80 94 102 83 80 102 111 87 84 106 115 91 84 103 104 85 84 103 108 85 88 107 118 88 79 99 104 83 84 99 113 87 84 99 109 87 3

In each line of data the four spectral values for the top-left pixel are given first followed by the four spectral values for the top-middle pixel and then those for the top-right pixel, and so on with the pixels read out in sequence left-to-right and top-to-bottom. Thus, the four spectral values for the central pixel are given by attributes 17, 18, 19 and 20. If you like you can use only these four attributes, while ignoring the others.

5.1.3 Evaluation Results

Step 1: loading the Training and Testing Data
Loading the Training data.....
The Total Training Records Loaded: 500
The First Ten Sample Records of UCI Spatial Training Data

92	115	120	94	84	102	106
	79	84	102	102	83	101
	85	84	103	104	81	102
	126	134	104	88	121	128
	100	84	107	113	87	Class : 3

Loading the Testing data.....
The Total Testing Records Loaded: 500
The First Ten Sample Records of UCI Spatial Test Data

80	102	102	79	76	102	102
	79	76	102	106	83	76
	88	80	107	118	88	79
	107	109	87	79	107	109
	87	79	107	113	87	Class : 3

…. .. ……..

Step 2: Clustering and Classification using k-means
Clustering the Training Data (Button 3a)
Clustering the UCI Spatial data using k-Means Algorithm
The Time Taken for Clustering: 1.1250 sec
The Accuracy of Clustering
The Sensitivity: 56.67
The Specificity: 100.00
The Accuracy : 94.80
The Rand Index : 0.73

Classifying the Testing Data using previous Cluster Centers (Button 3b)
Classifying the UCI Spatial data using k-Means Algorithm
The Time Taken for Classification: 0.0150 sec
The Accuracy of Classification
The Sensitivity: 90.60
The Specificity: 100.00
The Accuracy : 97.80
The Rand Index : 0.84

Step3: Clustering and Classification using BPNN
Training the ANN with training data **(Button 4a)**
Clustering the UCI Spatial data using BPNN
TRAINGDX, Epoch 0/1000, MSE 0.142439/0.01, Gradient 0.5829/1e-006

TRAINGDX, Epoch 100/1000, MSE 0.0291814/0.01, Gradient 0.01241/1e-006
TRAINGDX, Epoch 200/1000, MSE 0.0231053/0.01, Gradient 0.0209676/1e-006
TRAINGDX, Epoch 300/1000, MSE 0.0220311/0.01, Gradient 0.042881/1e-006
TRAINGDX, Epoch 400/1000, MSE 0.0213853/0.01, Gradient 0.048597/1e-006
TRAINGDX, Epoch 500/1000, MSE 0.0206704/0.01, Gradient 0.0235075/1e-006
TRAINGDX, Epoch 600/1000, MSE 0.0202969/0.01, Gradient 0.0148538/1e-006
TRAINGDX, Maximum epoch reached, performance goal was not met.

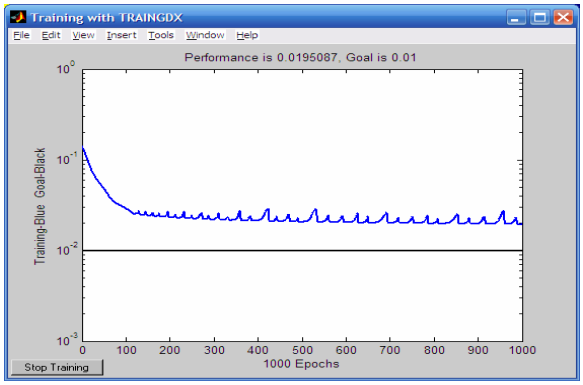

Figure 4: Clustering and Classification using BPNN

The Time Taken for Clustering: 7.2190 sec
The Accuracy of Classification with BPNN
The Sensitivity: 96.67
The Specificity: 99.77
The Accuracy : 99.40
The Rand Index : 0.95

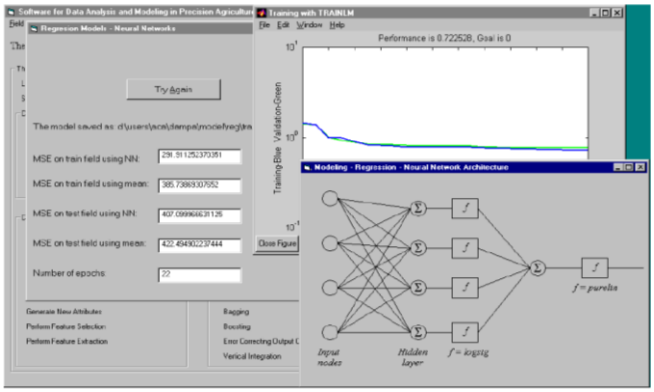

Figure 5: Proposed BPNN Architecture

Testing the BPNN with testing Data (Button 4b)
Classifying the UCI Spatial data using BPNN
The Time Taken for Classification: 1.2500 sec
The Accuracy of Classification with BPNN
The Sensitivity: 96.58
The Specificity: 99.74
The Accuracy : 99.00
The Rand Index : 0.91

5.1.4 The Comparative Results

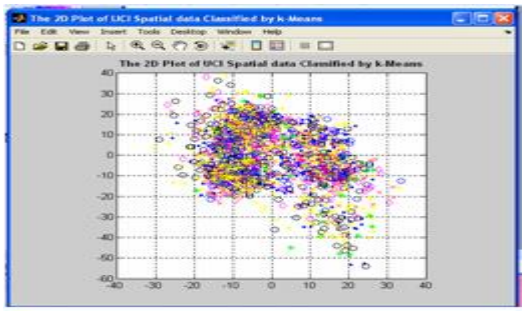

Figure 6: Classification by K-Means

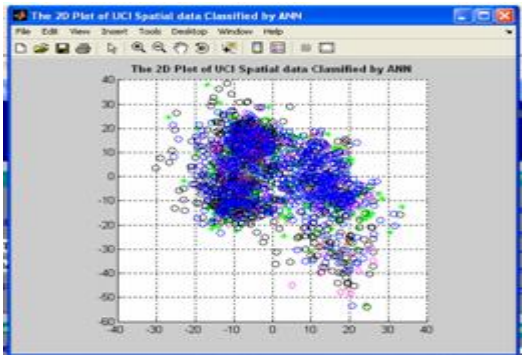

Figure 7: Classification by BPNN

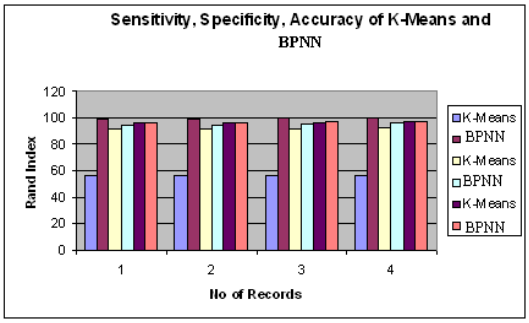

Figure 8: Sensitivity, Specificity, Accuracy of K-Means and BPNN

Table1: The Performance in terms of Rand Index

Sl.No	No. of Records	Rand Index	
		kmeans	BPNN
1	500	0.90	0.90
2	1000	0.82	0.86
3	1500	0.83	0.79
4	2000	0.78	0.79
Avg Performance		0.8325	0.835

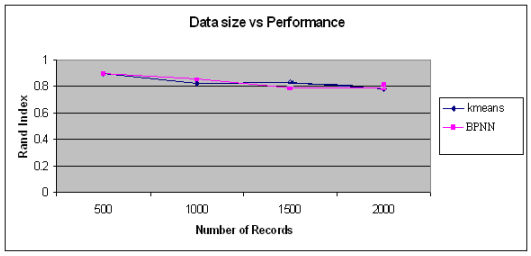

Figure 9: Data Size vs. Performance of k-means, ANN

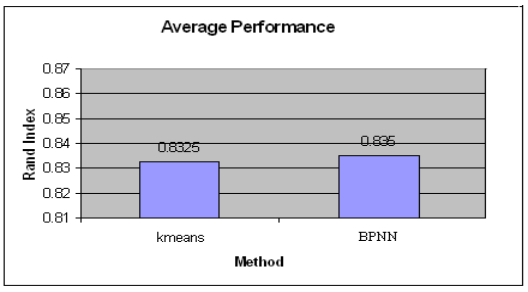

Figure 10: Average Performance of k-means, ANN

6. CONCLUSION

The average accuracy of classification has been measured using the metrics Rand index as well as Sensitivity, Specificity and Accuracy. As shown the performance graphs in the previous section, the transformation of data using BPNN leads to better accuracy of clustering and classification. The results proves that the traditional methods like k-means clustering will not give better accuracy in the case of spatial data due to its nature of complexity and dimensionality. The arrived results were significant and comparable. This makes the proposed method as a simple and powerful tool for the classification of multidimensional spatial data sets and to reduce the dimensionality.

7. ACKNOWLEDGMENT

We are very thankful to our Karpagam University for the encouragement to do this research as a successful one.

8. REFERENCES

[1] Kharin. Y. S, zhuk. E. E. "Robust classification of multinomial observations with possible outliers", J CLASSIF 17(1): 51-65, 2000.

[2] Laura A. Mather, "A linear algebra measure of cluster quality", Journal of the American Society for Information Science and Technology, Article first published online: 18 MAY 2000, DOI: 10.1002/(SICI)1097-4571(2000) 51:7<602::AID-ASI3>3.0.CO;2-1, 2000.

[3] Shivakumar Vaithyanathan, Byron Dom, "Model-Based Hierarchical Clustering (2000), In Proc. 16th Conf, Uncertainty in Artificial Intelligence.

[4] Polanco X; Francois C; Lamirel JC, "Using artificial neural networks for mapping of science and technology: A multi-self-organizing-maps approach", Scientometrics 51 (1): 267-292, May 2001 .

[5] Koren. Y. And harel. D., "Clustering of spatial data using random walks", Arnetminer, In Proc KDD, pages 6, 2001.

[6] Sandrine Dudoit and Jane Fridlyand, "prediction-based resampling method for estimating the number of clusters in a dataset", Genome Biol. 2002; 3(7): research0036.1–research0036.21, Published onlineJune 25, 2002.

[7] Ju-Won Song, Kyu-Young Whang, Young-Koo Lee, Min-Jae Lee, Wook-Shin Han, and Byung-Kwon Park, "The clustering property of corner transformation for spatial database applications", Information and Software Technology, 44(7), , Pages 419-429, 15 May 2002.

[8] Vishwanathan, SVN and Murty, Narasimha M (2002), "Kernel Enabled K- Means Algorithm".

[9] Chaomei Chen, Timothy Cribbin, Robert Macredie, Sonali Morar, "Visualizing and tracking the growth of competing paradigms: Two case studies", Article first published online: 9 APR 2002, DOI: 10.1002/asi.10075.

[10] Jin-Tsong Hwang, Hun-chin Chiang, "The study of high resolution satellite image classification based on Support Vector Machine", Geoinformatics, 2010 18th International Conference, Print ISBN: 978-1-4244-7301-4, DOI: 10.1109/GEOINFORMATICS. 2010.5567755, 09 September 2010.

[11] Paulo Gonçalves, Hugo Carrão, Andre Pinheiro & Mário Caetano, "Land cover classification with Support Vector Machine applied to MODIS imagery", Pages 517-525. doi=10.1.1.134.4769, May 2002.

[12] Madhubala M, S.K.Mohan Rao, G. Ravindra Babu, "Classification of IRS LISS- III Images by using Artificial Neural Networks", IJCA Journal, Number 3 - Article 2, 2010.

[13] Rui Xu and Donald Wunsch, "Survey of Clustering Algorithms", IEEE, Vol. 16, No.3, pp.645-678, May 2005.

[14] G. Dong & M. Xie, "Color Clustering & Learning for Image Segmentation Based on Neural Networks", IEEE, 16(4), Pp.925- 936.

[15] Suresh Subramanian, Nahum Gat, Michael Sheffield, Jacob Barhen, Nikzad Toomarian, "Methodology for hyperspectral image classification using novel neural network ", Algorithms for Multispectral and Hyperspectral Imagery III, SPIE Vol. 3071-- Orlando, FL, April 1997.

[16] Nana Liu, Jingwen Li, Ning Li, "A graph-segment-based unsupervised classification for multispectral remote sensing images", WSEAS Transactions on Information Science and Applications , Volume 5 Issue 6, June 2008.

[17] S. Nagaprasad , "Spatial Data Mining Using Novel Neural Networks for Soil Image", International Journal of Engineering Science and Technology Vol. 2(10), 2010, 5621-5625.

[18] Rekik, A. Zribi, M. Benjelloun, M. ben Hamida, A, "A k-Means Clustering Algorithm Initialization for Unsupervised Statistical Satellite Image Segmentation", E-Learning in Industrial Electronics, 2006 1ST IEEE International Conference, Print ISBN: 1-4244-0324-3 , DOI: 10.1109/ICELIE.2006.347204 , 16 April 2007.

[19] Ahmed Rekik, Mourad Zribi, Ahmed Ben Hamida and Mohamed Benjelloun, "An Optimal Unsupervised Satellite image Segmentation Approach Based on Pearson System and k-Means Clustering Algorithm Initialization", World Academy of Science, Engineering and Technology 59 2009, pages 640-647.

[20] W. Meyer, D. W. Paglieroni, C. Astaneh , "K-Means Re-Clustering Algorithmic Options with Quantifiable Performance Comparisons ",The International Society for Optical Engineering, Jan 2003.

[21] Rekik, Ahmed; Zribi, Mourad; Hamida, Ahmed Ben; Benjelloun, Mohamed, "An optimal unsupervised satellite mage segmentation approach based on Pearson system and k-means clustering algorithm initialization", International Journal of Signal Processing, January 1, 2009.

[22] G. Dong & M. Xie, "Color Clustering & Learning for Image Segmentation Based on Neural Networks", IEEE, 16(4), Pp.925- 936.

[23] Hierarchical clustering for multivariate spatial patterns. In Proceedings of the 10th ACM International Symposium on Advances in Geographic Information Systems, pages 131 – 136, 2002.

[24] Mayank Tyagi, Francesca Bovolo, Member, IEEE, Ankit K. Mehra, Subhasis Chaudhuri, and Lorenzo Bruzzone, Senior Member, IEEE, "A Context-Sensitive Clustering Technique Based on Graph-Cut Initialization and Expectation-Maximization Algorithm", IEEE Geoscience And Remote Sensing Letters, VOL. 5, NO. 1, pages 21-25, Jan 2008.

[25] Neukirchen, J. Rottland, D. Willett, and G. Rigoll. A continuous density interpretation of discrete HMM systems and MMI-neural networks. IEEE Transactions on Speech and Audio Processing, 9(4):367–377, 2001.

[26] Li D.R., Wang S.L., Li D.Y., 2006. The theories and application of spatial data mining, science press.

[27] Li D.Y., Du Y., 2007. Artificial Intelligence with Uncertainty, CRC Press, USA.

[28] Samani A, Bishop J, Plewes D. A constrained modulus reconstruction technique for breast cancer assessment. IEEE Trans Med Im 2001;20(9):877–85.

9. AUTHOR PROFILE

N.Naga Saranya, currently pursuing her Ph.D. degree Under the guidance of Dr. M.Hemalatha, Head, Dept of Software Systems, Karpagam University, Tamilnadu, India. One year of research experience and she published four international journals and also presented four papers in various national and international conferences. Area of interest is Data Mining.

Dr..M.Hemalatha completed MCA M.Phil., PhD in Computer Science and currently working as a Asst Professor and Head, Dept of Software Systems in Karpagam University. Eleven years of Experience in teaching and published sixty papers in International Journals and also presented seventy papers in various National conferences and international conferences. Area of research is Data Mining, Software Engineering, Bioinformatics, and Neural Network. Also reviewer in several National and International journals.