International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 06, June-2022

Machine Learning Pre-processing using GUI

Md Arshad Anwar, Yuvraj Bansal, Nagesh Jadhav
CSE-MIT SOE

***Abstract*— Machine Learning is a subset of the larger field of artificial intelligence (AI) that focuses on teaching computers how to learn without the need to be programmed forspecific tasks. In fact, the key idea behind ML is that it is possible to create algorithms that learn from and make predictions on the data. Examples of Machine Learning are present everywhere including the spam filter that flags messages in your email, the recommendation engine Netflix uses to suggest content you might like, and the self-driving cars being developed by Google and other companies. But before applying Machine Learning on any dataset, you need to convert it in such a way that the algorithms could understand the dataset.Imagine your wolf pack decides to watch a movie you haven't heard of. There is absolutely no debate about that,it will lead to a state where you find yourself puzzled with lot of questions which needs to be answered in order to make a decision. Being a good chieftain the first question you would ask, what is the cast and crew of the movie? As a regular practice, you would also watch the trailer of the movie YouTube. Furthermore, you'd find out ratings and reviews the movie has received from the audience. Whatever investigating measures you would take before finally buying popcorn for your clan in theater, is nothing but what data scientists in their lingo call 'Exploratory Data Analysis'.**

Keywords*— *python, preprocessing, machine learning

INTRODUCTION

Data pre-processing is an integral step in Machine Learning as the quality of dataand the useful information that can be derived from it directly affects the ability of your model to learn. Data is said to be unclean if it is missing attribute, attribute values, contains noise, or outliers, and duplicate or wrong data.Therefore, it is imperative that you pre-process your data before feeding it into your model. But data pre- processing is often considered time consuming and tedious by many Machine Learning developers. This project covers: Data Description - Document is to record all information about the data filesand their contents so that someone can use the data in a future research project and understand the data's content and structure. Handling NULL Values - Missing values can be handled by deleting the rows or columns having null values. If columns have more than half of therows as null then the entire column can be dropped. The rows which are having one or more columns values as null can also be dropped. Encoding Categorical Data - Machine learning models require all input and output variables to be numeric. This means that if your data contains categorical data, you must encode it to numbers before you can fit and evaluate a model. The two most popular techniques are an Ordinal Encodingand a One-HotEncoding. Feature Scaling - Feature Scaling is a technique to standardize the independent features present in the data in a fixed range.

...If feature scaling is not done, then a machine learning algorithm tends to weigh greater values,higher and consider smaller values as the lower values, regardless of the unit of the values. Univariate analysis - Univariate analysis is the technique of comparing and analyzing the dependency of a single predictor and a response variable. The prefix "uni" means one, emphasizing the fact that the analysis only accounts for one variable's effect on a dependent variable. Bivariate analysis - is performed to find the relationship between each variable in the dataset and the target variable of interest (or) using 2 variables and finding the relationship between them. Download the pre-processed Dataset – this option is also available. Image processing is the general issue in today's era, when we work with computer vision. It is in itself, a broad view to be considered. In order to process the image, we need to segment it so that it would become easier for the computer to understand. Image segmentation is the process of segmenting the image into various segments, that could be used for the further applications such as: Image understanding model, Robotics, Image analysis, Medical diagnosis, etc. Image segmentation is the process of partitioning an image into multiple segments, so as to change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation means assigning a label to each pixel in the image such that pixels with same labels share common visual characteristics. It makes an image easier to analyze in the image processing tasks. There are many different techniques available to perform image segmentation. Our motive is to implement almost the same concept as we humans try to implement, while understanding the image which we visualize. In human vision, the complex image is immediately segmented into the simple objects on the basis of color, texture, patterns, shapes, etc. This same thing is constructed with the help of the image segmentation techniques in the computer vision system[22]. All this would be done with the help of GUI.

RELATED WORK

In this section,we first outline research related to different machine learning techniques for structured and unstructured dataset,and then review literature related to how to combine different techniques to make a more generalized application to perform these techniques to any dataset.

Satish Kumar et. Al[12], in his survey explained the various applications that uses the concept of the image segmentation which includes, computer vision, medical, scanning, recognition, etc.

P. Sravani et al.. [13], in their survey, an overview of different segmentation methods and clustering are

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 06, June-2022

studied. Though many techniques are developed, not all types are useful for all types of images. Segmentation segments the image and clusters according to some similarity. Distance metric is a similarity measure and has direct impact on the clusters formed. In this, Fuzzy is powerful unsupervised clustering method which is widely used for robust segmentation of real time images. Traditional FCM and many other algorithms use Euclidean Distance metric.

H. P. Narkhede[14], in his review of image segmentation study, has described various methodologies and issues regarding to digital image processing used in various recognition patterns.

PunamThakare[17], in her paper describes the various image segmentation techniques and discusses in detail the edge detection techniques and their evaluation. It gives an algorithm which is a combination of detection and evaluation of the edge detectors. The results show that the recognition rate depends on the type of the image and their ground truths.

PROBLEM STATEMENT AND DATA:

Problem Statement:

In this paper our goal is to Make Pre-processing easy so developers could concentrate on applying ML algorithms more. Our proposed system should be able to handle most of pre-processing methods by making use of GUI and combines all advantages of different approaches: ease of use , and repeatability. It should be Not easily usable by everyone even by beginners. Handle both text and image dataset for pre- processing. find easy way for developers for graphical Analysis.

Data:

Data we use for our proposed system for testing is the titanic dataset for text data pre-processing and graphical analysis of dataset which has the following data dictionary:

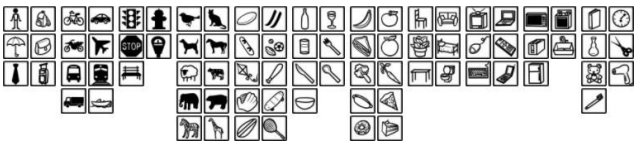
Variable	***Definition***	***Key***
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
Sex	Sex	
Age	Age in years	
Sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	Port of Embarkation C = Cherbourg, Q = Queenstown , S = Southampto n

And for Image segmentation we have used the COCO which is is a large-scale object detection, segmentation, and captioning dataset. COCO has several features:

Object segmentation Recognition in context
Superpixel stuff segmentation 330K images (>200K labeled)
1.5 million object instances 80 object categories
91 stuff categories

5 captions per image

250,000 people with keypoints

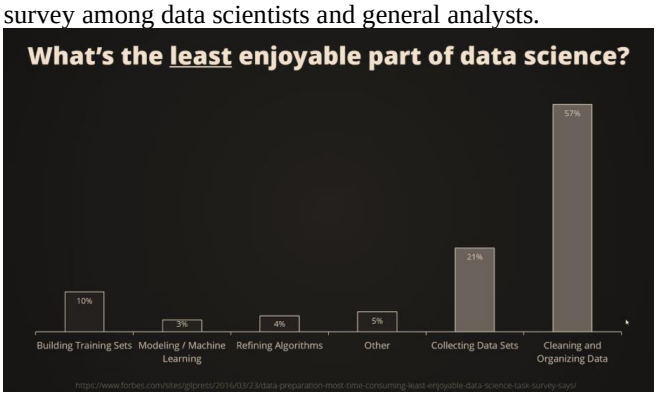


MACHINE LEARNING PRE-PROCESSING USING GUI:

In this section,We provide an overview of our proposed System which takes a raw dataset and provides various pre- processing tools using check and click functions and provides various tools for graphical analysis for the user to find various relationship between attributes and gives a final pre-processed dataset according to the developers need. We basically handled two types of dataset 1.text dataset 2.image dataset.in which we can perform Data Description ,Handling NULL Values ,Encoding Categorical Data,Feature Scaling ,Univariate analysis ,Bivariate analysis ,Image segmentation. And after the developer is finished and he get the desired pre- processed dataset he can download the dataset to perform ML algorithm.

a) WHY AUTOMATE PRE-PROCESSING AND ANALYSIS:

In the real world data are generally incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data. Noisy: containing errors or outliers. Inconsistent: containing discrepancies in codes or names. 80% of our work as machine learning experts and data scientists is preparing the data. I personally feel it's actually much higher. But here are some data statistics from somebody who did a survey among data scientists and general analysts.



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 06, June-2022

We need to analyze the data for the following reasons:

- Identifying dataset distribution
- Choosing the right Machine Learning algorithm
- Extracting Right features
- Evaluate our ML algorithm and presenting our results

In image recognition system, segmentation is an important stage that helps to extract the object of interest from an image which is further used for processing like recognition and description. Image segmentation is the practice for classifying the image pixels.
And by automating these functions we can save developers lot of time.

b) PRE-PROCESSING METHODS:

- Handling null values:

There are some instances where a particular element is absent because of various reasons, such as, corrupt data, failure to load the information, or incomplete extraction. Handling the missing values is one of the greatest challenges faced by analysts, because making the right decision on how to handle it generates robust data models.

PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

1. Deleting Rows
This method commonly used to handle the null values. Here, we either delete a particular row if it has a null value for a particular feature and a particular column if it has more than 70-75% of missing values. This method is advised only when there are enough samples in the data set.

2. Replacing With Mean/Median/Mode
This strategy can be applied on a feature which has numeric data like the age of a person or the ticket fare. We can calculate the mean, median or mode of the feature and replace it with the missing values. This is an approximation which can add variance to the data set. But the loss of the data can be negated by this method which yields better results compared to removal of rows and columns. Replacing with the above three approximations are a statistical approach of handling the missing values. This method is also called as leaking the data while training.

3. Assigning An Unique Category
A categorical feature will have a definite number of possibilities, such as gender, for example. Since they have a definite number of classes, we can assign another class for the missing values. Here, the features Cabin and Embarked have missing values which can be replaced with a new category, say, U for 'unknown'. This strategy will add more information into the dataset which will result in the change of variance.

Every dataset we come across will almost have some missing values which need to be dealt with. But handling them in an intelligent way and giving rise to robust models is a challenging task.

- Encoding categorical data:

We use this categorical data encoding technique when the features are nominal(do not have any order). In one hot encoding, for each level of a categorical feature, we create a new variable. Each category is mapped with a binary variable containing either 0 or 1. Here, 0 represents the absence, and 1 represents the presence of that category.

These newly created binary features are known as Dummy variables. The number of dummy variables depends on the levels present in the categorical variable.

Index	Animal
0	Dog
1	Cat
2	Sheep
3	Horse
4	Lion

One-Hot code

Index	Dog	Cat	Sheep	Lion	Horse
0	1	0	0	0	0
1	0	1	0	0	0
2	0	0	1	0	0
3	0	0	0	0	1
4	0	0	0	1	0

- Feature scaling:

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here's the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Normalization equation

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Here's the formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 06, June-2022

c) UNIVARIATE ANALYSIS:

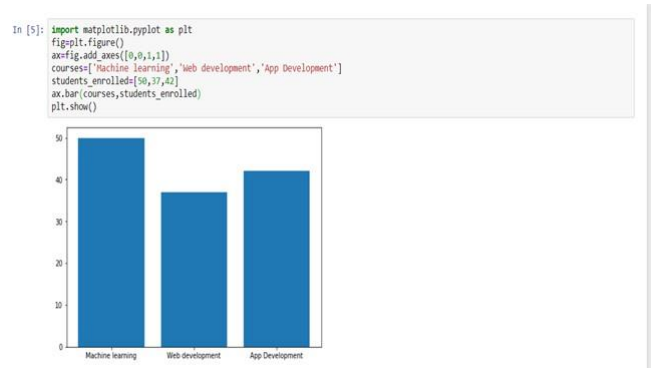
Uni means one and variate means variable, so in univariate analysis, there is only one dependable variable. The objective of univariate analysis is to derive the data, define and summarize it, and analyze the pattern present in it. In a dataset, it explores each variable separately. It is possible for two kinds of variables- Categorical and Numerical.

Some patterns that can be easily identified with univariate analysis are Central Tendency (mean, mode and median), Dispersion (range, variance), Quartiles (interquartile range), and Standard deviation.

Univariate data can be described through:

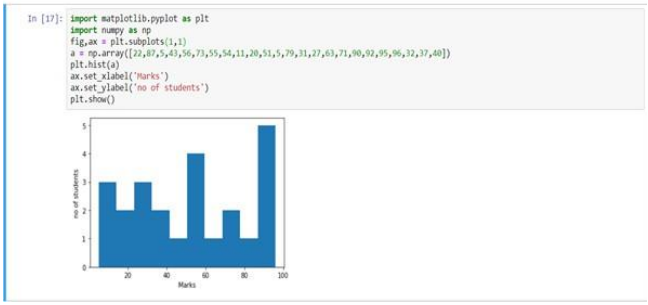
- Bar Charts:

The bar graph is very convenient while comparing categories of data or different groups of data. It helps to track changes over time. It is best for visualizing discrete data.



- Histograms:

Histograms are similar to bar charts and display the same categorical variables against the category of data. Histograms display these categories as bins which indicate the number of data points in a range. It is best for visualizing continuous data.

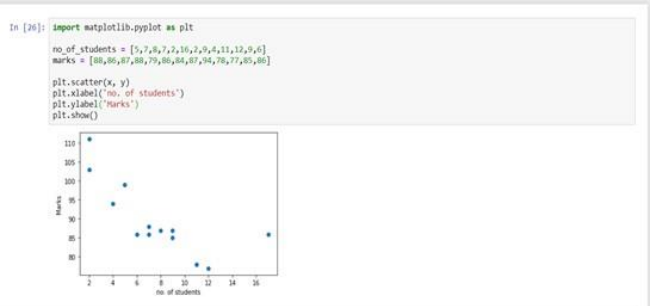


d) BIVARIATE ANALYSIS:

Bi means two and variate means variable, so here there are two variables. The analysis is related to cause and the relationship between the two variables.

- Scatter Plot:

A scatter plot represents individual pieces of data using dots. These plots make it easier to see if two variables are related to each other. The resulting pattern indicates the type (linear or non-linear) and strength of the relationship between two variables.



- Linear Correlation:

Linear Correlation represents the strength of a linear relationship between two numerical variables. If there is no correlation between the two variables, there is no tendency to change along with the values of the second quantity.

$$r = \frac{Covar(x,y)}{\sqrt{Var(x)Var(y)}}$$

$$Covar(x,y) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n}$$

$$Var(x) = \frac{\Sigma(x-\bar{x})^2}{n}$$

$$Var(y) = \frac{\Sigma(y-\bar{y})^2}{n}$$

r : Linear Correlation

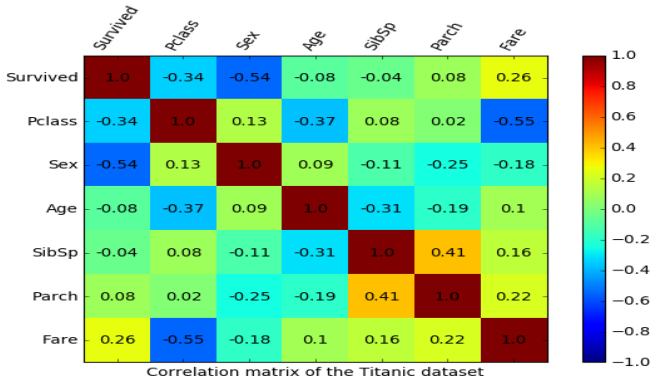
$Covar$: Covariance

Var : Variance

- Co-relation Matrix:

A correlation matrix is simply a table which displays the correlation coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values in a table. It is a powerful tool to summarize a large dataset and to identify and visualize patterns in the given data.

A correlation matrix consists of rows and columns that show the variables. Each cell in a table contains the correlation coefficient.

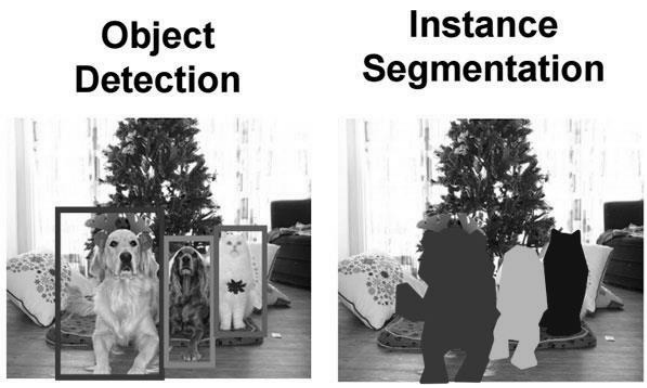


International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 06, June-2022

e) IMAGE SEGMENTATION:

We can divide or partition the image into various parts called segments. It's not a great idea to process the entire image at the same time as there will be regions in the image which do not contain any information. By dividing the image into segments, we can make use of the important segments for processing the image. That, in a nutshell, is how image segmentation works.

An image is a collection or set of different pixels. We group together the pixels that have similar attributes using image segmentation.



Region-based Segmentation:

One simple way to segment different objects could be to use their pixel values. An important point to note – the pixel values will be different for the objects and the image's background if there's a sharp contrast between them.

In this case, we can set a threshold value. The pixel values falling below or above that threshold can be classified accordingly (as an object or the background). This technique is known as Threshold Segmentation. If we want to divide the image into two regions (object and background), we define a single threshold value. This is known as the global threshold.

Edge Detection Segmentation:

What divides two objects in an image? There is always an edge between two adjacent regions with different grayscale values (pixel values). The edges can be considered as the discontinuous local features of an image.

We can make use of this discontinuity to detect edges and hence define a boundary of the object. This helps us in detecting the shapes of multiple objects present in a given image.

CONCLUSION:

Data pre-processing is an integral step in Machine Learning as the quality of data and the useful information that can be derived from it directly affects the ability of your model to learn. Data is said to be unclean if it is missing attribute, attribute values, contains noise, or outliers, and duplicate or wrong data.

Therefore, it is imperative that you pre-process your data before feeding it into your model. But data pre-processing is often considered time consuming and tedious by many Machine Learning developers.

In this paper, a study based representation of different pre- processing approaches is defined. Throughout this study of the various techniques, we concluded out that by automating these methods can make pre-processing easy and beginner friendly so that developers can concentrate on the actual algorithm or model.

REFERENCES:

[1] Muzamil Bhat. (2014, January). "Digital Image Processing". International Journal of Science & Technology Research. Volume 3 (issue 1), ISSN 2277-8616.

[2] Pushmeet Kohli, Stefanie Jegelka, (2013). "A Principled Deep Random Field Model for Image Segmentation".

[3] Nikita Sharma, Mahendra Mishra, Manish Shrivastava. (2012, May). "Color Image Segmentation Techniques and Issues: An Approach". International Journal of Science & Technology Research. Volume 1 (issue 41), ISSN 2277- 8616.

[4] D.Sasirekha, Tamilnadu, India, Dr.E.Chandra, Dr.SNS Rajalakshmi. (2012, September). "Enhanced Techniques for PDF Image Segmentation and Text Extraction". International Journal of Computer Science and Information Security (IJCSIS). Volume 10 (issue 9).

[5] Rajeshwar Dass, Priyanka, Swapna Devi. (2012, January-March). "Image Segmentation Techniques". IJECT. Volume 3 (issue 1), ISSN: 2230-7109 (Online) | ISSN: 2230- 9543 (Print).

[6] Krishna Kant Singh, Akansha Singh. (2010, September). "A study of Image Segmentation Algorithms for Different Types of Images". IJCSI International Journal of Computer Science Issues. Volume 7 (issue 5). ISSN (Online): 1694-0784. ISSN (Print): 1694-0814.

[7] Jifeng Ning, LeiZhang, DavidZhang, ChengkeWu. (2010). "Interactive image segmentation by maximal similarity based region merging". journal homepage: www.elsevier.com/locate/pr, Pattern Recognition 43 (2010) 445 -- 456

[8] Salem Saleh Al-amri, N.V. Kalyankar and Khamitkar S.D. (2010, May). "Image Segmentation by Using Threshold Techniques". Journal of Computing. Volume 2, ISSUE 5. [Online].

[9] N. Senthilkumaran and R. Rajesh. (2009, May). "Edge Detection Techniques for Image Segmentation – A Survey of Soft Computing Approaches". International Journal of Recent Trends in Engineering. INFORMATION PAPER. Volume 1 (issue 2).

[10]806 Nida M. Zaitoun and Musbah J. Aqel / Procedia Computer Science 65 (2015) 797 – 806 [10] Yi Yang, Sam Hallman, Deva Ramanan, Charless C. Fowlkes. (2009-2010). "Layered object Models for Image Segmentation".

[11] L'ubor Ladick ý , Chris Russell and Philip H.S. Pushmeet Kohli. (2009). "Associative Hierarchical CRFs for Object Class Image Segmentation".

[12] DR.S.V.KASMIR RAJA, A.SHAIK ABDUL KHADIR, DR.S.S.RIAZ AHAMED. (2005-2009). "Moving Toward Region-Based Image Segmentation Techniques: A Study". Journal of Theoretical and Applied Information Technology.

[13] Orlando J. Tobias, Rui Seara. (2002, December). "Image Segmentation by Histogram Thresholding Using Fuzzy Sets". IEEE TRANSACTIONS ON IMAGE PROCESSING, Volume 11(issue 12).

[14] Costantino Carlos Reyes-Aldasoro. (2001). "Image Segmrntaion with Kohonen Neural Network Self- Organazing Maps".

[15] Hai Gao, Wan-Chi Siu and Chao-Huan Hou. (2001, December). "Improved Techniques for Automatic Image Segmentation".

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 06, June-2022

IEEE Transactions on Circuits and Systems for Video Technology. Volume 11 (issue 12).

[16] Kamiya Motwani, Nagesh Adluru, Chris Hinrichs, Andrew Alexander, Vikas Singh. "Epitome driven 3-D Diffusion Tensor image segmentation: on extracting specific structures". {kmotwani, hinrichs, vsingh} @cs.wisc.edu. {adluru, alalexander2}@wisc.edu.

[17] John Paul Walters, Vidyananth Balu, Suryaprakash Kompalli, Vipin Chaudhary. "Evaluating the use of GPUs in Liver Image Segmentation and HMMER Database Searches".

[18] Sara Vicente, Vladimir Kolmogorov, Carsten Rother. "Graph cut based image segmentation with connectivity priors Technical report".

[19] Mustafa Özden, Ediz Polat. "Image Segmentation Using Color and Texture Features".

[20] Bo Peng, Lei Zhang, Jian Yang. "Iterated Graph Cuts for Image Segmentation".

[21] Yi Yang, Sam Hallman, Deva Ramanan, Charless C. Fowlkes. "Layered Object Models for Image Segmentation".

[22] Dorin Comaniciu, Peter Meer. "Robust Analysis of Feature Spaces: Color Image Segmentation"

[23] Chen, Z., Zhao, Z., Gong, P. and Zeng, B., (2006) A new process for the segmentation of high resolution remote sensing imagery. International Journal of Remote Sensing, 27 (22), pp. 4991-5001.

[24] Satish Kumar, Raghavendra Srinivas, "A Study on Image Segmentation and its Methods", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 9, September 2013.

[25] P. Sravani et al, " A Survey on Image Segmentation Techniques and Clustering", International Journal of Advance Research in Computer Science and Management Studies, Special Issue, December 2013.

[26] H. P. Narkhede, "Review of Image Segmentation Techniques", International Journal of Science and Modern Engineering (IJISME) ISSN: 2319-6386, Volume-1, Issue-8, July 2013.

[27] Rajeshwar Dass, Priyanka, Swapna Devi, "Image Segmentation Techniques", IJECT Vol. 3, Issue 1, Jan.-March 2012.

[28] Nikita Sharma, Mahendra Mishra, Manish Shrivastava, "COLOUR IMAGE SEGMENTATION TECHNIQUES AND ISSUES: AN APPROACH", International Journal of Scientific & Technology Research Volume 1, Issue 4, May 2012.

[29] PunamThakare, "A Study of Image Segmentation and Edge Detection Techniques", International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No. 2 Feb 2011.

[30] V. Dey, Y. Zhang, M. Zhong, "A REVIEW ON IMAGE SEGMENTATION TECHNIQUES WITH REMOTE SENSING PERSPECTIVE", July 5–7, 2010, IAPRS, Vol. XXXVIII, Part 7A.

[31] N. Senthilkumaran and R. Rajesh, "Edge Detection Techniques for Image Segmentation – A Survey of Soft Computing Approaches", International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009.