Volume 6, Issue 6, June 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Experimental Study of Applying Multiclass Classifier Algorithms for Image Classification

Sanskruti Patel*
Assistant Professor, Faculty of Computer Science & Applications, CHARUSAT,
Gujarat, India

Abstract— Background/Objectives: To extract an information from a digital image often depends on two phases: Image segmentation which breaks down the image into homogenous regions and Image Classification which assigns these regions into particular classes. In the area of Computer Vision, Image segmentation is a specified process to represent an image into something more meaningful and easy to analyze so that it can be used for further processing. Data mining techniques, on the other hand, useful to identify hidden patterns from large amount of data. Methods: A data set of image segmentation is collected from UCI – repository of machine learning. To formulate the correct image type from featured data set of image segmentation, an experiment is performed on four classification algorithms. They are namely Naïve Byes, J48, SVM and K-Nearest Neighbor. Findings: An experiment is performed on four classification algorithms using weka data mining tool. From the result, it is observed that, J48 works better than other with the highest accuracy of 96.9264%. It correctly classifies the images from segmented image data.

Keywords— classification, image segmentation, machine learning.

I. INTRODUCTION

Computer vision is a field that automates the process of extraction, analysis and understanding of useful information from a single or a sequence of images[1]. It basically works on the given properties of the structure present in the images. It aims to model, replicate and exceed human vision using computer hardware and software. It mainly related to the area of Artificial Intelligence and Pattern Recognition. Before classification of a digital image, image needed to be segmented. To extract an information from a digital image often depends on two phases: Image segmentation which breaks down the image into homogenous regions and Image Classification which assigns these regions into particular classes. Image segmentation is a specified process in the area of computer vision. The basic aim of image segmentation is to represent an image into something more meaningful and easy to analyze and it is also to be used for further processing[2]. Image segmentation is used to locate objects and boundaries in images by partitioning a digital image into pixel sets. It is an essential step to in image analysis and other image processing tasks[3].

The role of segmentation is crucial in most tasks requiring image analysis. The success of the image analysis and classification tasks is often a direct consequence of the success of segmentation4. The applications of image processing include Satellite image processing, Medical diagnostic, Optical character recognition (OCR), Tracking of Objects etc.[4]

Data mining, sometimes also called as knowledge discovery, is a process of analyzing vast amount of data from different perspectives and summarizing it into useful information[5]. Traditional data analysis methods are no longer proficient to handle large data sets. Data mining is involving methods at the intersection of statistics, database systems, machine learning and artificial intelligence. It extracts information from a data set and transform it into understandable form for further use. It is widely used in diverse areas including Medical Diagnostics, Retail, Financial, Banking, Surveillance, Education, Network, Image Processing and many more[6]. Various algorithms and techniques like Clustering, Classification, Regression, Decision Trees, Nearest Neighbor, Artificial Neural Networks, Association Rules, Genetic Algorithm, etc. are available for mining the data[7].

As to classify an image, image segmentation is required and to uncover the hidden pattern from segmented data, data mining techniques are needed. Therefore, in this paper, an experimental study is performed on image segmentation data to classify the various images. Various classification algorithms like Naïve Byes, J48, SVM and K-Nearest Neighbor are tested to develop classifier for classification of image type. For experiment, a data set with 19 features is taken from the site of UCI repository. The entire experimental work is carried out with WEKA open source software under Windows 10 environment.

II. RELATED WORK

In various domain area, many researchers applied various data mining techniques for image segmentation. While applying data mining in image segmentation, the most important function of the mining is to generate all significant patterns without prior information of the patterns[8].

K. Fukuda et al[9] generated a decision tree classifier in WEKA. Their research work is intended for defoliation in aerial forest imagery. Walid MOUDANI et al[10] used decision tree technique for Image Classification. They have taken skin detection and face detection as problem statement and applied the data mining technique for classification. Kun-Che Lu

et al[11] proposed an efficient and effective model for image mining and image segmentation using decision tree. Content based tissue image mining was proposed by Gholap et al[12]. If the tissue images are indexed, stored and mined on content, tissue image mining is much faster and resourceful. Sanjay et al[13] carried research for applying an image mining technique using wavelet transform. The author proposed an image mining approach using wavelet transform. Image mining approach using clustering and data compression techniques was projected by Pattnaik et al[14].

Sheela & Shanthi[15] described the image mining approaches for categorization and segmentation of brain MRI data. Victor & Peter[16] applied a new minimum spanning tree based clustering algorithm for image mining. The said algorithm is proficient of detecting clusters with irregular boundaries. Hemalatha & Devasena[17] proposed a research to find out the accurate images while mining an image (multimedia) database. They explored an innovative method for mining images by means of LIM dependent image matching method by integrating neural networks. Eman M. Ali et al[18] presents a survey on various image mining techniques. These techniques were proposed earlier by researchers for the better development in the field of content-based image retrieval. A feature subset harmony search based fuzzy discernibility classifier (HSFD), hybrid wavelet based radial basis function (RBF) and neural network (WRBF) approaches are proposed as a data mining technique for image segmentation based classification by Mrutyunjaya Panda et al[19]. In the field of patient survival, Andrew Kusiak et al[20] used two different data mining algorithms for extracting knowledge in the form of decision rules. Their proposed method reduces the cost and effort of selecting patients for clinical studies.

III. DATA AND METHODOLOGY

The data set used for experimental purpose is downloaded from university of California of Iravin (UCI), Machine Learning Repository. Details of data set and attributes are given in Section 3.1.

A. Data Set Description

In the data set, Image data described by high-level numeric-valued attributes. The instances were drawn randomly from a database of 7 outdoor images. The images were handsegmented to create a classification for every pixel. Each instance is a 3x3 region. The data set has 210 instances of training data and 2100 instances of test data. Total 19 continuous attributes are present in the data set. The detail of data set is shown in following table 1. Also, there are seven classes namely sky, path, foliage, brickface, window, grass and cement. 30 instances per class for training data and 300 instances per class for test data are available in data set.

Table I : Attributes of Data Set and its possible Value

Attribute	**Value**
region-centroid-col	center pixel of the region column
region-centroid-row	center pixel of the region row
region-pixel-count	the number of pixels in a region = 9.
short-line-density-5	the results of a line extraction algorithm that counts how many lines of length 5 (any orientation) with low contrast, less than or equal to 5, go through the region.
short-line-density-2	same as short-line-density-5 but counts lines of high contrast, greater than 5.
vedge-mean	measure the contrast of horizontally adjacent pixels in the region – mean.
vegde-sd	measure the contrast of horizontally adjacent pixels in the region – standard deviation.
hedge-mean	measures the contrast of vertically adjacent pixels – mean.
hedge-sd	measures the contrast of vertically adjacent pixels – standard deviation.
intensity-mean	the average over the region of (R + G + B)/3.
rawred-mean	the average over the region of the R value.
rawblue-mean	the average over the region of the B value.
rawgreen-mean	the average over the region of the G value.
exred-mean	measure the excess red:(2R - (G + B)).
exblue-mean	measure the excess blue:(2B - (G + R)).
exgreen-mean	measure the excess green:(2G - (R + B)).
value-mean	3-d nonlinear transformation of RGB
saturatoin-mean	
hue-mean	

B. Data Mining Technique & Algorithms

Several techniques of data mining like classification, clustering, association rule, prediction, neural networks etc. are available to find useful patterns from large amount of data7. Among these, classification is the most frequently applied data mining technique that assigns items in a collection to target categories or classes. Classification maps data into predefined groups or classes. It classifies the collection of data into groups. The members of the groups are having one or

more characteristic in common. Various classification algorithms are available to build classifiers. Among of them, Naïve Byes, J48 and Support Vector Machine (SVM) have been studied and the results of this algorithms are presented in this paper.

1. Naive Byes

Naive Byes is one of frequently used classification algorithm in data mining. It represents a supervised learning method as well as a statistical method for classification. It is generally using when the dimensionality of the input is high. It applies Bayes' theorem with strong (naive) independence assumptions between the features. It requires a small amount of training data to estimate the parameters[23].

2. Support Vector Machine (SVM)

Support Vector Machine (SVM) is one of the method for supervised machine learning technique based on statistical learning theory. It is used for learning classification and regression rules from data. SVMs have frequently been found to provide maximum classification accuracies than other widely used pattern recognition techniques[21].

3. J48

J48 is the Java implementation of the C4.5 decision tree learner in the Weka data mining tool. C4.5 is a program that creates a decision tree based on a set of labelled input data. This algorithm was developed by Ross Quinlan. The algorithm uses a greedy technique to induce decision trees for 20 classifications and uses reduced-error pruning[22].

4. K-Nearest Neighbour

K-Nearest Neighbours algorithm is a non-parametric method used for classification and regression in the field of data mining. It basically used for pattern recognition. It is a type of lazy learning, where the function is only approximated locally and all computation is deferred until classification. It is one of the simplest machine learning algorithm.

IV. EXPERIMENTAL RESULT

The work represented in this paper is implemented in Weka. Weka is a collection of machine learning algorithms for data mining tasks. It offers a user interface for experiment various algorithms and also it can be integrated in our own JAVA code. It is an open source software and freely available under General Public License (GNU) agreement. The screen shot of WEKA during preprocessing stage is depicted in Figure 1.

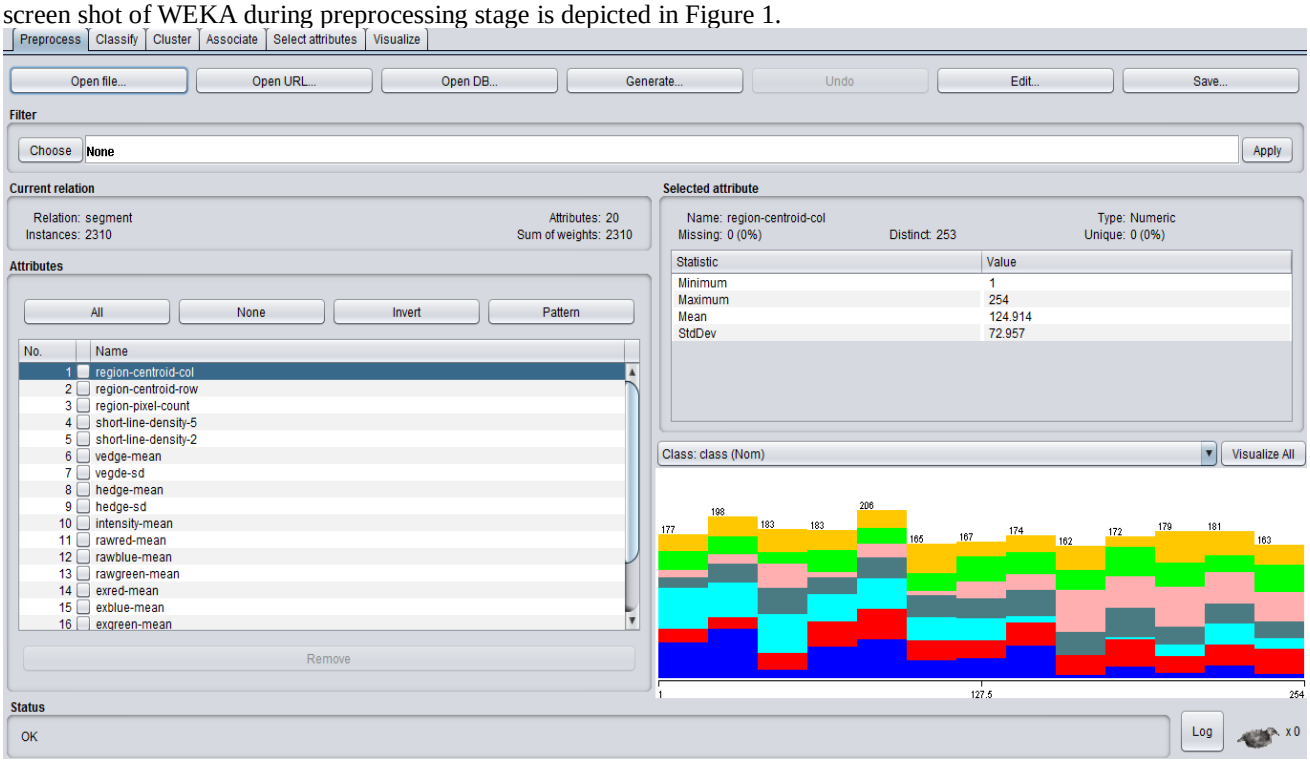

Figure 1 Preprocessing Stage in Weka

A. Classification Accuracy

Accuracy

Accuracy is derived as correctly classified instances divided by the total number of instances present in the dataset [21].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \ldots\ldots(1)$$

Here, TP- True Positive, FP- False Positive, TN- True Negative, FN- False Negative

Precision

The correlation of number of modules correctly classified to the number of entire modules classified fault-prone is given as a precision. It is quantity of units correctly predicted as faulty [21].

$$TP\ Rate = \frac{TP}{TP + FP} \quad \ldots\ldots(2)$$

TP Rate

TP Rate is used to find the high true-positive rate. The true-positive rate is also called as sensitivity [21].

$$TP\ Rate = \frac{TP}{TP + FN} \quad \ldots\ldots(3)$$

F-Measure

F- Measure is the one has the combination of both precision and recall which is used to compute the score. In the field of Information Retrieval, the F-measure is habitually used in order to guesstimate the query classification performance [21].

$$F\ Measure = \frac{2 * Recall * Precision}{Recall + Precision} \quad \ldots\ldots(4)$$

B. Outcome of the Experiment

There are total 2100 test data and 210 training data in the data set. All the data are classified as brickface, sky, foliage, cement, window, path, grass images. The classified dataset is evaluated using 10-fold cross-validation and the results are compared in Table 2. Table 2 shows the comparison between three algorithms namely Naïve Byes, J48 and Support Vector Machine (SVM). Each row of the table represents Correctly Classified Instances, Incorrectly Classified Instances, Precision, Recall, F-Measure, FP-rate and TP-rate.

Table 2: Accuracy of various Classification Algorithms

Algorithm	Correctly Classified Instances (%)	Incorrectly Classified Instances (%)	Precision	Recall	F-Measure	FP-rate	TP-rate
Naïve Byes	80.2165	19.7835	0.819	0.802	0.780	0.033	0.802
J48	96.9264	3.0736	0.969	0.969	0.969	0.005	0.969
SVM	93.0736	6.9264	0.931	0.931	0.930	0.012	0.931
k-Nearest Neighbor	97.1429	2.8571	0.971	0.971	0.971	0.005	0.971

V. RESULTS & DISCUSSION

Table 2 shows the accuracy of various Classification Algorithms. Also, the correctly classified instances and incorrectly classified instances of these algorithms are represented in Figure 2. Also, comparison between several performance measures are represented in Figure 3. By seeing the above result, it is clear that, J48 performs better than any other algorithm. The obtained kappa statistics of J48 algorithm is 0.9641 which is considerably good.

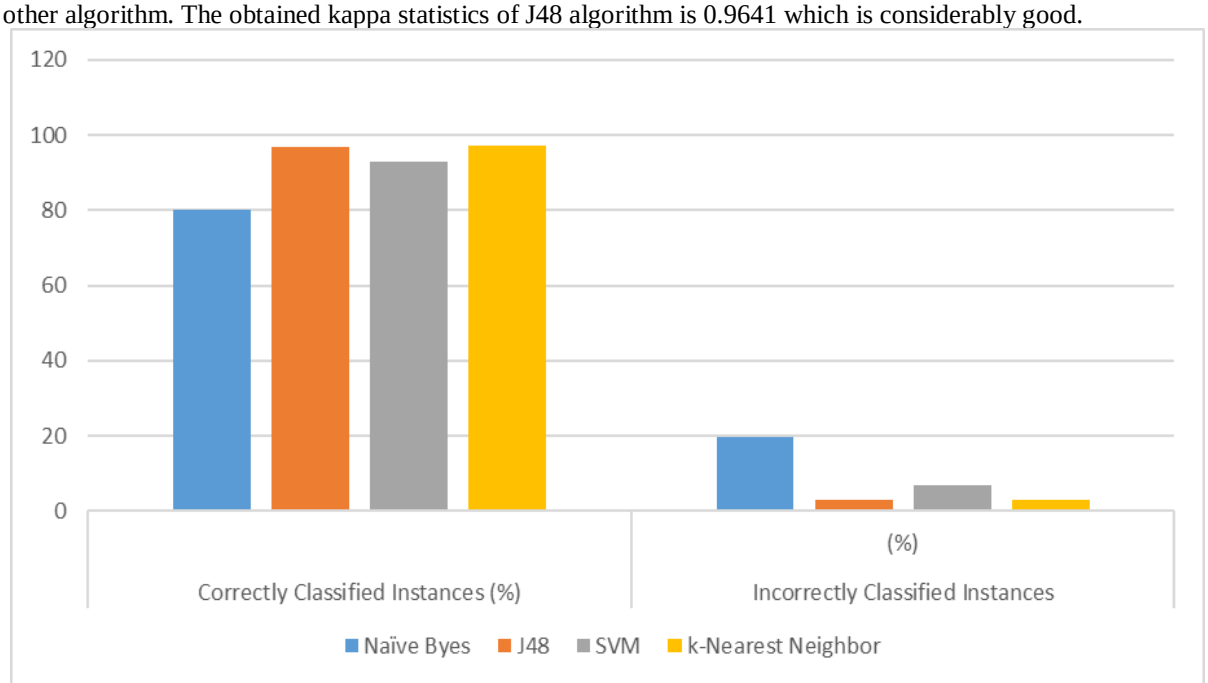

Figure 2 Accuracy of Classification Algorithms

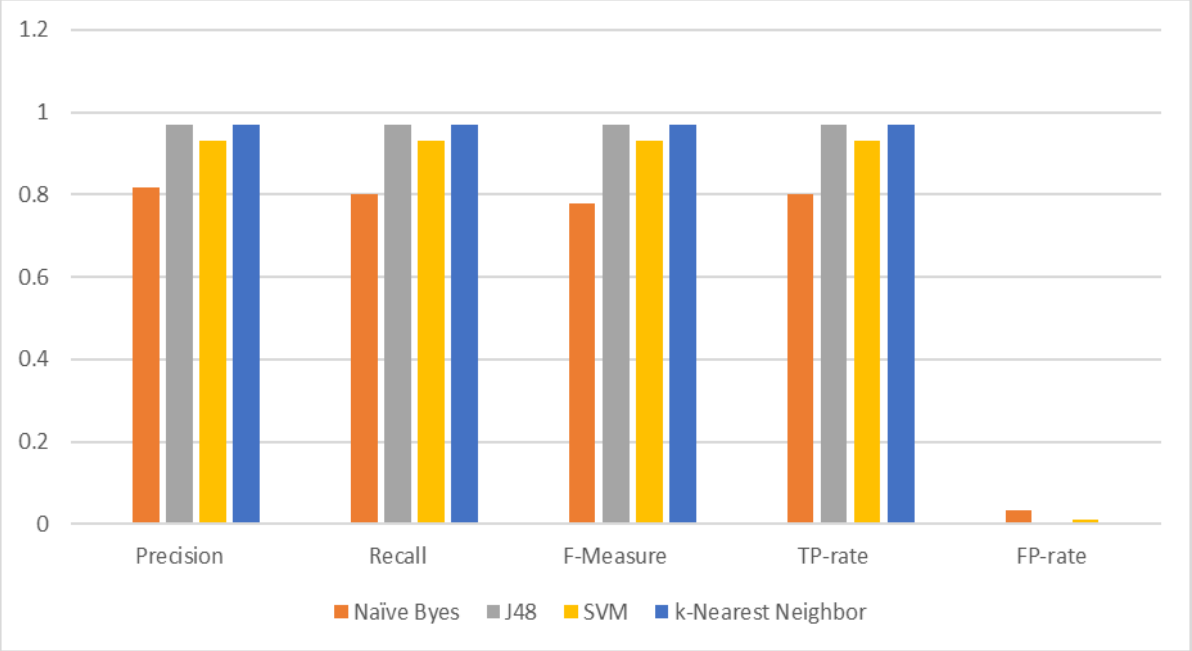

Figure 3 Performance Measures of Classification Algorithms

VI. CONCLUSION

Image segmentation is a specified process used to represent an image into something more meaningful and easy to analyze. To correctly classified images from these segmented data, data mining techniques are very useful. Several techniques and algorithms of data mining are available to find useful patterns from large amount of data. In this paper, four classification algorithms are selected for experimental purpose. From the result, J48 works better among all other classification algorithms. The result can be extended and enhanced too by variation in attributes and increasing number of records too.

REFERENCES

[1] What is Computer Vision? www.bmva.org/visionoverview, retrieved on March, 2016

[2] Christophoros Nikou, Basic Methods for Image Segmentation, http://www.cs.uoi.gr/~cnikou/Courses/Digital_Image_Processing/Chapter_10_Image_Segmentation.pdf, retrieved on April, 2016

[3] Segmentation, http://www.cs.uu.nl/docs/vakken/ibv/reader/chapter10.pdf A. M. Khan, Ravi. S, Image Segmentation Methods: A Comparative Study, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-3, Issue-4, Pages: 84-92, September 2013

[4] Xiaojun Qi, Image Segmentation, http://digital.cs.usu.edu/~xqi/Teaching/REU05/Notes/DIPSegmentation.pdf, retrieved on April, 2016

[5] Data Mining: What is Data Mining? http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm, retrieved on April, 2016

[6] Dharminder Kumar, Deepak Bhardwaj, Rise of Data Mining: Current and Future Application Areas, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, Pages:256-260, September 2011

[7] Bharati M. Ramageri, Data Mining Techniques and Applications, Indian Journal of Computer Science and Engineering, Volume 1, No. 4, Pages: 301-305

[8] Ramadass Sudhir, A Survey on Image Mining Techniques: Theory and Applications, Computer Engineering and Intelligent Systems www.iiste.org ISSN 2222-1719, Vol 2, No.6, Pages:44-51, 2011,

[9] K. Fukuda and P. A. Pearson, Data mining and image segmentation approaches for classifying defoliation in aerial forest imagery, http://www.iemss.org/iemss2006/papers/w7/83_Fukuda_4.pdf, retrieved on April, 2016

[10] Walid MOUDANI, Abdel Rahman SAYED, Efficient Image Classification using Data Mining, International Journal of Combinatorial Optimization Problems and Informatics, ISSN: 2007-1558, Vol. 2, No. 1, Jan-April 2011, pp. 27-44

[11] Kun-Che Lu, Don-Lin Yang, and Ming-Chuan Hung, Decision Trees Based Image Data Mining and Its Application on Image Segmentation, http://arbor.ee.ntu.edu.tw/pakdd02/paper/P0198.pdf, retrieved on May, 2016

[12] Abhi Gholap, Gauri Naik, Aparna Joshi and CVK Rao,"Content-Based Tissue Image Mining", IEEE Computational Systems Bioinformatics Conference - (CSBW'05), Pages: 359-363, 2005

[13] J. Fernandez, N. Miranda, R. Guerrero and F. Piccoli, "Appling Parallelism in Image Mining," www.ing.unp.edu.ar/wicc2007/trabajos/PDP/120.pdf, 2007

[14] Sabyasachi Pattnaik, Pranab Kumar Das Gupta and Manojranjan Nayak (2008), "Mining images using clustering and data compressing techniques", International Journal of Information and Communication Technology, vol. 1, no. 2, Pages: 131-147, 2008

[15] L. Jaba Sheela and V. Shanthi, "Image Mining Techniques for Classification and Segmentation of Brain MRI Data," Journal of Theoretical and Applied Information Technology," vol. 3, no. 4, Pages: 115-121, 2007

[16] S. P. Victor and S. John Peter (2010), "A Novel Minimum Spanning Tree Based Clustering Algorithm for Image Mining," European Journal of Scientific Research, vol. 40, no. 4, Pages: 540-546, 2010

[17] M. Hemalatha and C. Lakshmi Devasena (2011), "A Hybrid Image Mining Technique using LIMbased Data Mining Algorithm," International Journal of Computer Applications, vol. 25, no.2, Pages: 1-5. 2011

[18] Eman M. Ali, Ahmed F. Seddik, Mohamed H. Haggag, Using Data Mining Techniques for Children Brain Tumors Classification based on Magnetic Resonance Imaging, International Journal of Computer Applications (0975 – 8887), Volume 131, No.2, Pages:36-42, December 2015,

[19] Mrutyunjaya Panda, Aboul Ella Hassanien, Ajith Abraham , Hybrid Data Mining Approach For Image Segmentation Based Classification , International Journal of Rough Sets and Data Analysis Volume 3,(Issue 2):Article 5, April 2016

[20] AndrewKusiak, Bradley Dixonb, Shital Shaha, Predicting survival time for kidney dialysis patients: a data mining approach, Elsevier Publication, Computers in Biology and Medicine 35, page: 311–327, 2005

[21] Dr. S. Vijayarani, Mr.S.Dhayanand,, Data Mining Classification Algorithms For Kidney Disease Prediction, International Journal on Cybernetics & Informatics (IJCI) Vol. 4, No. 4, August 2015.

[22] J. Ross Quinlan. C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[23] Naïve Byes, www.wikipedia.org, retrieved on April, 2016

