AN IMPROVED APPROACH USING CONSISTENCY OVER SAMPLING FOR HANDLING UNIFORM EFFECT IN UNSUPERVISED LEARNING

Shaik. Nagul[1]
Research Scholar,Department of Computer Science
Krishna University
Machilipatnam, Andhra Pradesh, India

R.Kiran kumar[2]
Senior Assistant Professor,Department of Computer Science
Krishna University
Machilipatnam, Andhra Pradesh, India

abstract:-To date, the K-means algorithm has been widely applied to diverse datasets in real world. The class imbalance datasets are one of the categories of datasets in which one of the class predominates the other class of data in terms of percentage of instances. The k-means algorithm tends to generate poor results when applied to imbalance data clustering. The above said effect is termed as "uniform effect" in k-means. In this paper, we proposed a novel clustering algorithm Consistency Over Sampling K-means (COS_K-means) for efficient handling of imbalance data and reducing the uniform effect. Motivated by general over sampling strategy, we proposed a unique oversampling strategy which uses consistency technique to identify effective features and instances in the data source. The fine-tuned instances are oversampled in the minority space for improved performance. The experimental results suggest that the proposed COS_K-means approach have performed better than the classical k-means approach and reduced the issues relating to uniform effect.

Keywords: Data Mining, Knowledge Discovery, Clustering, K-means, imbalance data, uniform effect, over sampling, COS_K-means.

1. INTRODUCTION

In knowledge discovery, data mining is one of the growing fields of research. In data mining, there are different techniques for knowledge discovery such as: classification, clustering, frequent pattern mining etc. In the above techniques, classification is one of the most popular and well practiced techniques for knowledge discovery due to its simplicity and insightful properties [1].

In classification, there exist different model building techniques for knowledge discovery such as decision trees, neural networks, support vector machines etc. As the name specifies, in decision tree approach a tree model is build by following the process of induction using the data source. Decision Tree models are very simple to understand and to interrupt for knowledge discovery. The decision tree formed contains root nodes and leaf nodes. Root nodes are the originating nodes of the decision tree. These root nodes as descend downwards as branches lead to the terminating nodes known as leaf node. The branch leading from the root node to a leaf node will provide a decision for a particular case using decision tree [2].

In clustering, the collections of instances are grouped into different clusters depending on the intrinsic properties of the instances. There are many popular clustering algorithms such as k-means, hierarchical, DBSCAN, k-medoids and fuzzy clustering etc. K-means is one of the most widely used clustering approaches in real world applicability.

Researchers have identified some data characteristics that may strongly impact the performance of K-means clustering, including the types and scales of data and attributes, the sparseness of data, and the noise and outliers in data [3]. K-means has been shown to perform as well as or better than a variety of other clustering techniques in text clustering, and has an appealing computational efficiency [4, 5, 6].

However, further investigation is needed to unveil how data distributions can make impact on the performance of K-means clustering. Wu. J [7] has conducted an organized study on the effect of imbalance data distributions on K-mean clustering. They illustrate that K-means tends to produce clusters in relatively uniform sizes, even if the input data have varying true cluster sizes. They named this phenomenon as "uniform effect". In this study, we investigated the causes and reasons for the uniform effect and proposed a novel algorithm to solve the uniform effect in k-means clustering.

The arrangement of paper is follows as. In Section 2, we present the recent literature on uniform effect in clustering. Section 3, describes the reason and rectifying technique for the problem of uniform effect and at last it laid the basis for the proposed technique COS_K-means. Section 4 presents the experimental set up and the validation measures used for results analysis. In Section 5, results of the proposed approach COS_K-means are presented with the k-means classical approach. Section 6 presents the conclusion with the extension of the future work.

2.CURRENT APPROACHES OF K-MEANS WITH IMBALANCE DATA LEARNING

K-means approaches with imbalance data is presented by many of the researchers, some of the contribution are review as the recent literature.

Aziah Ali et al [8] have applied data discovery techniques in health domain for investigation of retinal images for identification of blood vessels segmentation. The digital images captured are improved with the use of Gabor wavelet feature and k-means clustering. The improved digital images are analyses for efficiency using logical OR techniques. Ahmed Alkilany et al [9] have implemented the techniques for load shedding using k-means clustering and patter extraction. The proposed framework is used for predicting future load shedding in the process of semi clustering the regional loads.

Preeti Arora et al [10] have reviewed large scale 10k dataset from KEEL repository on two classical benchmark clustering techniques k-means and k-mediods. The clusters are grouped based on the provided inputs of the distributed data from various sources. Nameirakpam Dhanachandra et al [11] have investigated the process of digital image improvement using different techniques of subtractive clustering. The generation of centeriod is performed using the existing potential data points in the source. The medial filter technique is presented for elimination of noisy or outlier regions from the generated digital images.

Jeyhun Karimov et al [12] have discussed various centeriod selection techniques in k-means algorithm using meta-heuristic searching. The initial assigned centeriod is moved to better candidate media centeriod for improved results generation. T. Santhanam et al [13] have applied K-Means algorithm for improvement in feature selection and further investigation is carried using support vector machines as the base classifier.

3.The proposed Consistency Over Sampling K-means (COS_K-means)

This section presents the proposed algorithm Consistency Over Sampling K-means (COS_K-means), whose main characteristics are depicted in the following sections.

The different scenarios for the existing of various class imbalances are categorized by various researchers in the form a range of diverse sources [7]. The degradation of clustering performance is the respective clustering algorithm is linked to the inbuilt technique used in the clustering approach. This problem of uniform effect in k-means algorithm for non uniform data i.e class imbalance data is already evident from many studies [7]. In this paper, we propose to use the concept of unique oversampling to solve the problem of uniform effect in k-means clustering on 22 imbalance datasets especially in the minority class to perform oversampling.

In a binary class imbalance data, the class with high percentage of instances is called as the majority class and the class with less percentage of instances is known as minority class. The oversampling of the instances in the minority class will reduce the problem of class imbalance. The improper oversampling of data in the minority class can lead to disaster of the data source with an irreparable state. A proper statistical or technical analysis is to be conducted before initiating the oversampling process. In our work, the concept of intelligent over sampling of consistence oversampling is used to study and identify different regions in the minority subset.
The identified regions are categories into safe, borderline and unsafe regions. An action should be initiated for borderline and unsafe regions whereas no action is required for the safe regions. The unsafe regions are the regions formed with the noisy or outlier instances. The removing of the unsafe regions will help to improve the quality of minority subset. The borderline regions are formed with a mix of instances which belong to majority and minority subsets. The improvement of these regions is required for refining of the minority subset.

The oversampling of the reaming stronger instances in the minority subset is initiated. The percentage of oversample can range from 10-100% depending on the intrinsic properties of the dataset. In the final stage, the improve minority subset and majority subset are combined and applied to a base algorithm for results simulation. Here, we have considered k-means clustering [14] as the base algorithm for our frame work.

4.Experimental Setup and Assessment Criteria

Experiments are conducted using twenty two datasets from UCI [15] data repositories. Table 1 summarizes the benchmark datasets used in the anticipated study. For each data set, S.no., Dataset: name of the dataset, Instances: number of instances, Attributes: Number of Attributes, IR: Imbalance Ratio are described in the table for all the datasets.

Table 1 UCI datasets and their properties

S.no.	Dataset	Inst	Attributes	IR
1.	abalone19	4174	9	129.43
2.	abalone9-18	731	9	16.40
3.	ecoli-0-1-3-7_vs_2-6	281	8	39.14
4.	ecoli4	336	8	15.8
5.	glass-0-1-6_vs_2	192	10	10.29
6.	glass-0-1-6_vs_5	184	10	19.44
7.	glass2	214	10	11.58
8.	glass4	214	10	15.46
9.	glass5	214	10	22.77
10.	page-blocks-1-3_vs_4	472	11	15.85
11.	shuttle-c0-vs-c4	1829	10	13.86

12.	shuttle-c2-vs-c4		129	10
	20.5			
13.	vowel0	988	14	9.97
14.	yeast-0-5-6-7-9_vs_4	528	9	9.35
15.	yeast-1-2-8-9_vs_7	947	9	30.56
16.	yeast-1-4-5-8_vs_7	693	9	22.1
17.	yeast-1_vs_7	459	8	14.3
18.	yeast-2_vs_4		514	9
	9.07			
19.	yeast-2_vs_8	482	9	23.1
20.	yeast4		1484	9
	28.09			
21.	yeast5		1484	9
	32.72			
22.	yeast6	1484	9	41.4

We performed the implementation of our new algorithms COS_K-means within the Weka [16] environment on windows 7 with i5-2410M CPU running on 2.30 GHz unit with 4.0 GB of RAM. The validation of the results is done using 10 fold cross validation, in which the dataset is split into 10 subsets and in each run nine subset are used for training and the remaining subset is used for testing. In 10 runs, the testing subset is altered and average measures for the 10 runs are generated. The evaluation metrics used in the paper are detailed below,

The Area under Curve (AUC) measure is computed using the below equation (1) or (2),

$$AUC = \frac{1 + TP_{RATE} - FP_{RATE}}{2} \quad \text{----------} \quad (1)$$

Or

$$AUC = \frac{TP_{RATE} + TN_{RATE}}{2} \quad \text{----------} \quad (2)$$

The Precision measure is computed using the below equation (3),

$$\Pr ecision = \frac{TP}{(TP)+(FP)} \quad \text{----------} \quad (3)$$

The Recall measure is computed using the below equation (4),

$$\Re call = \frac{TP}{(TP)+(FN)} \quad \text{----------} (4)$$

The F-measure Value is computed using the below equation (5),

$$F-measure = \frac{2 \times \Pr ecision \times \Re call}{\Pr ecision + \Re call} \quad \text{----------} (5)$$

5. Results

The main aim of the experiment setup is to verify whether the identified categories of data anomalies constitute a different degree of difficulty for the learning algorithms, and whether removal of the constituted categories improve the performance of the COS_K-means clustering approach.

In this experiment we focus on studying classical k-means clustering approach, there are more elements which could be influenced by data characteristics for other classifiers ad the uniform effect may not be visible in specific context. We have decided to K-means clustering algorithms, which have been often considered in related works and which represent benchmark in clustering.

In the first step of experiments, we compare the performance of COS_K-means on all the 22 imbalanced datasets. Again we can present the details of selected experiments in due course of discussion.

Table 2-5 presents the AUC, Precision, Recall and F-measure respectively. The datasets are sorted in the same way as in Table 1 – from the alphabetical order distributions. Relating the results to the labelling of the datasets presented in Table 2, we can observe how with the increasing difficulty of the dataset distribution, the performance of COS_K-means is performed better than k-means. For datasets where safe majority examples prevail (ecoli and glass variant datasets), COS_K-means clustered the majority class quite well – they recognize 70-90% of the majority examples. Only on abalone and yeast variant datasets, COS_K-means algorithm works less effective. In the clustering results generated we can clearly observe that a good improved is gained in the dataset which contains noisy, borderline or outlier instances in the 40-70% of the minority class,

When many rare and/or outlying examples are observed, the recall measure is also improved. Table 2 presents results in terms of AUC; the proposed approach COS_K-means has win on 15 datasets and loss on 7 datasets.

Table 2 Results of AUC on all the datasets with summary of tenfold cross validation performance

Datasets	K-means	COS_K-means
abalone19	0.414±0.136○	0.407±0.125
abalone9-18	0.479±0.120○	0.471±0.154
ecoli-0-1-3-7_vs_2-6	0.660±0.190●	0.820±0.111
ecoli4	0.533±0.173●	0.726±0.083
glass-0-1-6_vs_2	0.489±0.074●	0.489±0.050
glass-0-1-6_vs_5	0.621±0.249●	0.698±0.196
glass2	0.495±0.047●	0.501±0.055
glass4	0.785±0.191○	0.636±0.175
glass5	0.603±0.242●	0.681±0.185
page-blocks-1-3_vs_4	0.587±0.155○	0.406±0.154
shuttle-c0-vs-c4	0.685±0.191●	0.999±0.001
shuttle-c2-vs-c4	0.563±0.233○	0.705±0.247
vowel0	0.486±0.090●	0.690±0.137
yeast-0-5-6-7-9_vs_4	0.499±0.162●	0.728±0.089
yeast-1-2-8-9_vs_7	0.533±0.174○	0.533±0.143
yeast-1-4-5-8_vs_7	0.548±0.139○	0.525±0.103

yeast-1_vs_7	0.638±0.130●	0.656±0.097
yeast-2_vs_4	0.802±0.095●	0.869±0.059
yeast-2_vs_8	0.501±0.173●	0.824±0.111
yeast4	0.768±0.089●	0.802±0.077
yeast5	0.844±0.064●	0.902±0.015
yeast6	0.802±0.088●	0.843±0.056

● Bold dot indicates the win of Proposed COS_K-means approach;

Table 3 presents the comparison of the precision results, in which the proposed approach COS_K-means has win on all 22 datasets. Table 4 presents the comparison results of recall, in which the proposed approach COS_K-means win on 15 datasets and loss on 7 datasets. Table 5 presents the comparison results of F-measure, in which the proposed approach COS_K-means has win on all the 20 datasets and loss on 2 datasets.

Table 3 Results of Precision on all the datasets with summary of tenfold cross validation performance

Datasets	K-means	COS_K-means
abalone19	0.003±0.005●	0.004±0.009
abalone9-18	0.050±0.037●	0.084±0.080
coli-0-1-3-7_vs_2-6	0.062±0.172●	0.186±0.203
ecoli4	0.042±0.063●	0.191±0.060
glass-0-1-6_vs_2	0.013±0.069●	0.023±0.072
glass-0-1-6_vs_5	0.098±0.140●	0.255±0.226
glass2	0.007±0.036●	0.024±0.086
glass4	0.217±0.178●	0.250±0.224
glass5	0.073±0.104●	0.182±0.131
page-blocks-1-3_vs_4	0.078±0.044●	0.112±0.233
shuttle-c0-vs-c4	0.324±0.372●	0.990±0.018
shuttle-c2-vs-c4	0.065±0.170●	0.410±0.494
vowel0	0.087±0.033●	0.185±0.069
yeast-0-5-6-7-9_vs_4	0.101±0.086●	0.353±0.109
yeast-1-2-8-9_vs_7	0.035±0.023●	0.066±0.041
yeast-1-4-5-8_vs_7	0.052±0.033●	0.089±0.049
yeast-1_vs_7	0.096±0.044●	0.196±0.048
yeast-2_vs_4	0.279±0.092●	0.619±0.106
yeast-2_vs_8	0.067±0.167●	0.958±0.140
yeast4	0.092±0.022●	0.198±0.092
yeast5	0.094±0.023●	0.240±0.029
yeast6	0.069±0.019●	0.158±0.024

● Bold dot indicates the win of Proposed COS_K-means approach;

Table 4 Results of Recall on all the datasets with summary of tenfold cross validation performance

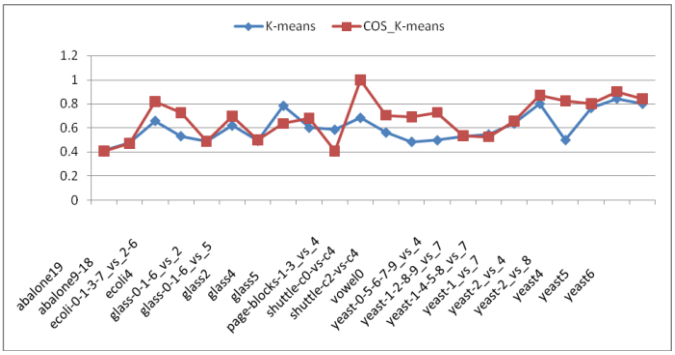
Datasets	K-means	COS_K-means
balone19	0.131±0.257○	0.101±0.218
abalone9-18	0.324±0.243○	0.322±0.325
ecoli-0-1-3-7_vs_2-6	0.390±0.490●	0.930±0.236
ecoli4	0.330±0.462●	0.923±0.243
glass-0-1-6_vs_2	0.025±0.131●	0.031±0.096
glass-0-1-6_vs_5	0.380±0.488●	0.580±0.387
glass2	0.020±0.098●	0.038±0.137
glass4	0.760±0.399○	0.457±0.349
glass5	0.370±0.485●	0.580±0.374
page-blocks-1-3_vs_4	0.633±0.320○	0.223±0.297
shuttle-c0-vs-c4	0.495±0.326●	1.000±0.000
shuttle-c2-vs-c4	0.200±0.402●	0.410±0.494
vowel0	0.431±0.164●	0.682±0.284
yeast-0-5-6-7-9_vs_4	0.440±0.289●	0.812±0.115
yeast-1-2-8-9_vs_7	0.547±0.350○	0.487±0.319
yeast-1-4-5-8_vs_7	0.443±0.296○	0.395±0.223
yeast-1_vs_7	0.700±0.309●	0.750±0.168
yeast-2_vs_4	0.862±0.186○	0.862±0.102
yeast-2_vs_8	0.465±0.350●	0.650±0.222
yeast4	0.828±0.189●	0.859±0.192
yeast5	0.970±0.171●	1.000±0.000
yeast6	0.898±0.183●	0.929±0.108

● Bold dot indicates the win of Proposed COS_K-means approach;

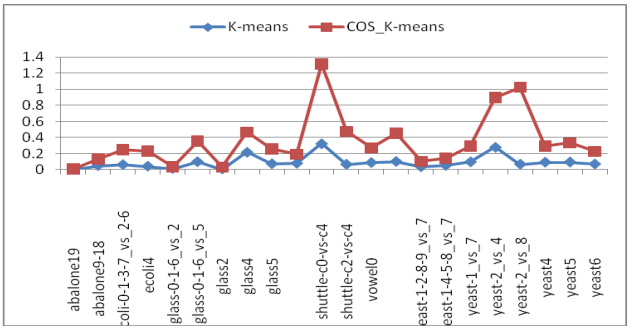
Table 5 Results of F-measure on all the datasets with summary of tenfold cross validation performance

Datasets	K-means	COS_K-means
abalone19	0.005±0.010●	0.008±0.017
abalone9-18	0.086±0.063●	0.132±0.127
ecoli-0-1-3-7_vs_2-6	0.089±0.180●	0.273±0.179
ecoli4	0.073±0.108●	0.316±0.094
glass-0-1-6_vs_2	0.017±0.089●	0.026±0.079
glass-0-1-6_vs_5	0.153±0.209●	0.335±0.251
glass2	0.010±0.052●	0.028±0.101
glass4	0.317±0.204○	0.301±0.230
glass5	0.119±0.167●	0.270±0.179
page-blocks-1-3_vs_4	0.138±0.076○	0.115±0.171
shuttle-c0-vs-c4	0.368±0.351●	0.995±0.009
shuttle-c2-vs-c4	0.090±0.205●	0.410±0.494
vowel0	0.144±0.054●	0.288±0.106
yeast-0-5-6-7-9_vs_4	0.162±0.125●	0.482±0.106
yeast-1-2-8-9_vs_7	0.066±0.043●	0.116±0.072
yeast-1-4-5-8_vs_7	0.092±0.059●	0.144±0.080
yeast-1_vs_7	0.169±0.076●	0.310±0.073
yeast-2_vs_4	0.416±0.117●	0.715±0.088
yeast-2_vs_8	0.089±0.115●	0.754±0.189
yeast4	0.165±0.039●	0.310±0.070
yeast5	0.170±0.040●	0.386±0.038
yeast6	0.128±0.034●	0.269±0.038

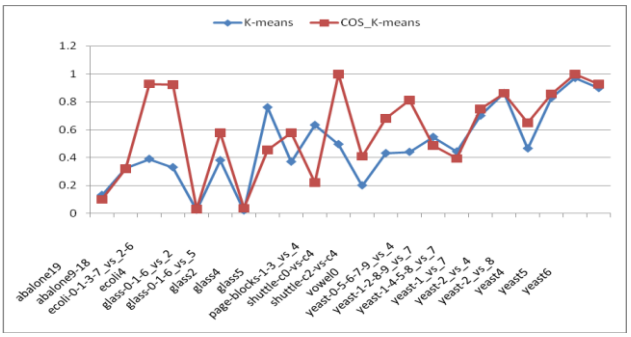
● Bold dot indicates the win of Proposed COS_K-means approach;



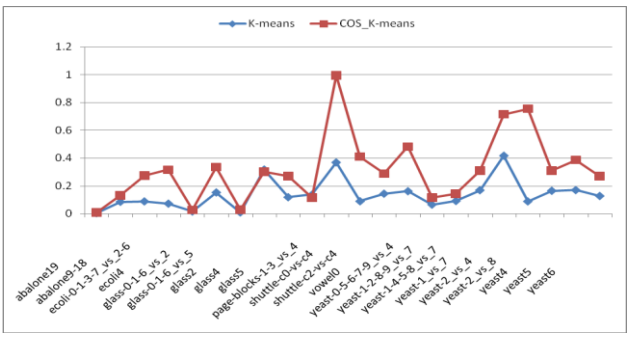
1 (a)



1(b)



1 (c)



1 (d)

Fig. 1 (a) AUC (b) precision (c) Recall (d) F-measure - Trends for K-means versus COS_K-means on imbalance data sets

The pictorial representations of AUC, precision, recall and F-measure against K-means verses COS_K-means are presented in Figure 1(a), 1(b), 1(c) and 1(d) respectively. Therefore, the overall results conclude that the proposed COS_K-means algorithm efficiently reduced the quality of the minority and majority data space. There by increasing the quality of the results provide the evidence and novelty of the algorithm on class imbalance learning datasets.

6. Conclusion

In this paper, we propose a novel clustering algorithm Consistency Over Sampling K-means (COS_K-means) for efficient handling of imbalance data. The proposed COS_K-means approach uses the uniform weight assigning technique with unique statistical computing for upgrading the dataset in terms of logical balancing. The experimental observation suggests that the proposed approach COS_K-means improves in terms of AUC, Precision, Recall and F-measure with the benchmark K-means on 22 imbalance datasets from UCI repository.

In future work, we will like to extend our system for high dimensional and complex datasets.

References

[1] Juanli Hu, Jiabin Deng, Mingxiang Sui, A New Approach for Decision Tree Based on Principal Component Analysis, Proceedings of Conference on Computational Intelligence and Software Engineering, page no:1-4, 2009.

[2] Shane Bergsma, Large-Scale Semi-Supervised Learning for Natural Language Processing, PhD Thesis, University of Alberta, 2010.

[3] Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Addison-Wesley, Upper Saddle River (2005)

[4] Krishna,K., Narasimha Murty,M.: Genetic k-means algorithm. IEEE Trans. Syst.Man Cybern. Part B **29**(3), 433–439 (1999).

[5] Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: Proceedings of the KDD Workshop on Text Mining (2000)

[6] Zhang, T., Ramakrishnan, R.,M.Livny: Birch: an efficient data clustering method for very large databases. In: Proceedings of 1996 ACM SIGMOD International Conference on Management of Data, pp. 103–114 (1996)

[7] Wu.J,"The Uniform Effect of K-means Clustering", J. Wu, *Advances in K-means Clustering*, Springer Theses, DOI:10.1007/978-3-642-29807-3_2, © Springer-Verlag Berlin Heidelberg 2012.

[8] Aziah Ali, Wan Mimi Diyana Wan Zaki and Aini Hussain," Blood Vessel Segmentation from Color Retinal Images Using K-Means Clustering and 2D Gabor Wavelet", K. Ntalianis and A. Croitoru (eds.), *Applied Physics, System Science and Computers*, Lecture Notes in Electrical Engineering 428, DOI 10.1007/978-3-319-53934-8_27.

[9] Ahmed Alkilany, Almahdi Ahmed, Hammad Said, Azuraliza Abu Bakar," Application of the K-means Clustering Algorithm to Predict Load Shedding of the Southern Electrical Grid of Libya", 978-1-4799-4233-6/14/$31.00 ©2014 IEEE.

[10] Preeti Arora, Dr. Deepali , Shipra Varshney," Analysis of K-Means and K-Medoids Algorithm For Big Data", Procedia Computer Science 78 (2016) 507 – 512.

[11] Nameirakpam Dhanachandra, Khumanthem Manglem and Yambem Jina Chanu," Image Segmentation using *K*-means Clustering Algorithm and Subtractive Clustering Algorithm", Procedia Computer Science 54 (2015) 764 – 771.

[12] Jeyhun Karimov, Murat Ozbayoglu," Clustering Quality Improvement of k-means using a Hybrid Evolutionary Model", Procedia Computer Science 61 (2015) 38 – 45.

[13] T. Santhanam, M.S Padmavathi," Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis", Procedia Computer Science 47 (2015) 76 – 83.
[14] MacQueen, J. B. (1967). *Some Methods for classification and Analysis of Multivariate Observations*. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. **1**. University of California Press. pp. 281–297. MR 0214227. Zbl 0214.46201. Retrieved 2009-04-07
[15] HamiltonA. Asuncion D. Newman. (2007). *UCI Repository of Machine Learning Database* ±School of Information and Computer Science, Irvine, CA: Univ. of California [Online].Available: http://www.ics.uci.edu/ ~mlearn/MLRepository.html
[16] Witten, I.H. and Frank, E. (2005) Data Mining: Practical machine learning tools and techniques. 2nd edition Morgan Kaufmann, San Francisco.

Sk.Nagul M.Tech (Ph.D)
Research Scholar,
Department of Computer Science
Krishna University,Machilipatnam
Mobile:9666724212,
nagulcse@gmail.com

Dr.Reddi Kiran Kumar M.Tech,Ph.D
Senior Assistant Professor,
Department of Computer Science,
Krishna University, Machilipatnam.
Mobile : (91)9440872455
kirankreddi@gmail.com