International Research Journal of Computer Science (IRJCS)
Issue 08, Volume 4 (August 2017)
ISSN: 2393-9842
www.irjcs.com

# SVM-KNN: A NOVEL APPROACH TO CLASSIFICATION BASED ON SVM AND KNN

**Rithesh R N**
Student, CSE Department,
R.V. College of Engineering, Bengaluru, India
rithesh1000@gmail.com

**Manuscript History**
Number: **IRJCS/RS/Vol.04/Issue08/AUCS10088**
**DOI: 10.26562/IRJCS.2017.AUCS10088**
Received: 09, August 2017
Final Correction: 14, August 2017
Final Accepted: 19, August 2017
Published: August 2017
**Citation: Rithesh, R. N. (2017). SVM-KNN: A Novel Approach to Classification Based on SVM and KNN.IRJCS:: International Research Journal of Computer Science, IV, 43-49. doi: 10.26562/IRJCS.2017.AUCS10088**
Editor: Dr.A.Arul L.S, Chief Editor, IRJCS, AM Publications, India


**Abstract**— Classification is one of the most predominant tasks for wide range of applications such as Sentiment analysis in text, voice recognition, image recognition, genetic engineering, data classification etc. Though many efficient classification algorithms have been introduced in the past few decades, due to the drastic increase in the amount of data generated across industry and academia there is a demand for classification algorithms with very high accuracy and robustness. This paper presents a new approach to enhance the accuracy of the classifier by combining Support Vector Machine (Classification algorithm) with K-Means Clustering algorithm and, finally using K Nearest Neighbours to make optimal choice on the classification problem .Experiments have shown that this new methodology has increased the accuracy of the classification problem and thus serves the intended purpose.

**Keywords**— Support Vector Machine, K-Means Clustering, K Nearest Neighbours, Clustering, Classification.

## I. INTRODUCTION

In today's age of globalization the world has witnessed a remarkable growth in the fields of Industry, Agriculture, Medicine, Scientific Research, and Business etc. The advancements in these domains has produced large amount of data. It is of very high importance to make sense out of these data for better understanding of all the related activities and also for decision making in the future. Two most important tasks that can be carried out with data are Prediction and Classification. Prediction is concerned with making a decision (i.e. predicting the outcome) for new data item based on the nature of available data. Some of the existing algorithms for predicting are: Linear Regression, decision tree etc. Classification is concerned with classifying the available data based on the characteristic features of each individual data instances. Some of the algorithms for classifications are: SVM, Nearest Neighbour etc. Classification has two phases, First phase is the training phase where the actual classification is done and the Second phase is the prediction phase where a new data instance is classified into one of the available classes. Support Vector Machine (SVM) is one of the most efficient algorithms for classification [1]. In SVM, given a set of training instances, each belonging to one of the available classes, an SVM model would construct a hyper plane which acts as a Decision boundary for classification of new data instances. SVM performs well for both Linear and Non-linear problems.

SVM is used for face recognition [2][3], text classification, Image classification, Hand written character recognition etc. One of the drawbacks of SVM is that it requires full labelling of input data. K-Means clustering algorithm basically originated from Signal Processing. In k-means clustering, 'n' data instances is grouped into 'k' clusters in which each data instance belongs to the cluster with nearest mean. The initial value of 'k' is still a matter of concern because it influences the overall accuracy of model [4].Unlike SVM, In k-means the data instances need not have to be marked with the respective classes. K-means clustering is mainly used in market segmentation, computer vision, astronomy etc. But, the problem with this algorithm is different initial partitions can result in different final clusters and also, it is difficult to predict k-value.

K Nearest Neighbour algorithm is a non parametric method which can be used for both regressions as well as for classification. In this method, a new data instance is classified as belonging to a particular class, say Z, if among the k nearest neighbours majority of the neighbors belong to the same class, i.e. Z. It can be proved that the KNN rule becomes the Bayes optimal decision rule as k goes to infinity[5].KNN is primarily used in recommendation systems. One of the drawbacks in this method is choosing the optimal value for k and also, in choosing what should be the nature of distance that has to be considered between the data instances to obtain optimal results. By taking into account all the three methods i.e. SVM, K-means and KNN a new approach to solve classification problem has been proposed in this paper. The rest of the paper is organised as follows: II .Related work, III. Proposed Methodology, IV Experimentation and V conclusion.

## II. RELATED WORK

### A. Support Vector Machine

In classification problems which are linearly separable or binary classifiers the objective of SVM is to construct a Decision boundary- which is a line in case of 2D data instances and a hyper plane in case of 3D data instance. A hyper plane is chosen as the decision boundary if it is away from both the classes to a maximum possible extent i.e. decision boundary is the maximum separating margin. Consider the training set X={x1,x2,x3...xn} and corresponding target vector Y={y1,y2,y3...yn}.Here each of the data instance, xi, belongs to R-dimensional space and yi ={+1, -1} Let 'w' and 'b' be the weight vector and bias respectively of the optimal hyper plane. Then the equation of the decision boundary can be written as formula 1.

$$w \cdot x + b = 0 \quad (1)$$

Furthermore formula 1 can be extended to frame equations for separating functions for the classes, which is formula 2 and formula 3.

$$w \cdot xi + b >= +1 \text{ where } yi = +1 \quad (2)$$

$$w \cdot xi + b <= -1 \text{ where } yi = -1 \quad (3)$$

Two parallel hyper planes forms the boundary and is parallel to decision boundary, the equation for the same shown in formula 4

$$w \cdot x + b = +1, \; w \cdot x + b = -1 \quad (4)$$

Every point in the R-dimensional space should satisfy the following constraint:

$$yi \, (w \cdot xi + b) >= 1 \quad (5)$$

By geometry, we get the value of classifying margin as $2/||w||$ .In order to get better classification results this value should be maximized, which in turn means minimizing the value of $||w||$ or $||w||^2$ . This constrained optimization problem can be solved using Lagrange multipliers. The following Lagrange function is shown in formula 6, where α is Lagrange multiplier

$$L(w, b, \alpha) = \tfrac{1}{2} w^T w - \sum \alpha[y(w. x + b) - 1] \quad (6)$$

Finally we get the following formula 7

$$w = \sum \alpha y x, \quad \sum \alpha y = 0 \quad (7)$$

By substituting formula 7 in formula 6 we get formula 8, where $L_d$ dual lagrangian:

$$L_d(\alpha) = \sum \alpha - \tfrac{1}{2} \sum\sum \alpha_i \alpha_j y_i y_j x_i^T x_j \quad (8)$$

In order to get optimal hyper plane formula 8 has to be maximized with respect to the following inequality constraint:

$$\alpha_i >= 0 \text{ for all } i, \sum \alpha_i y_i = 0 \quad (9)$$

Graphical representation of the SVM can be seen in the Figure 1.

### B. K-Means Clustering

K-Means is a centroid based clustering technique[6] . In this method 'n' data instances is grouped into 'k' clusters based on the distance between the data instance and the centroids of the 'k' clusters. The algorithm for the same is as follows:

Let X = {x1, x2, x3... xn} be the set of data points and C = {c1, c2, c3... ck} be the set of k centres.


 

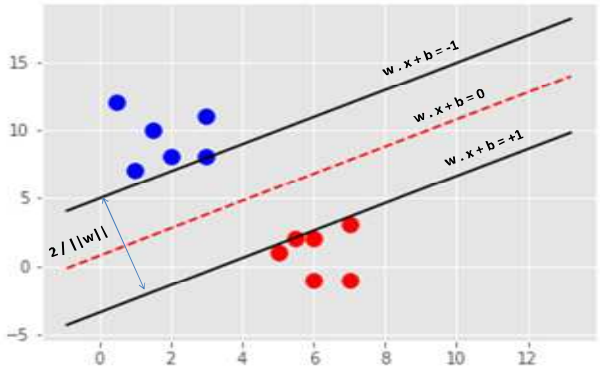


Fig. 1. Shows the SVM model for binary classification with red line showing the decision boundary

**K-Means Algorithm:**

*Input:* k : Number of desired clusters and X = {x1, x2, x3... xn}

*Output:* k distinct clusters

**Method:**

1. Randomly select k cluster centres.
2. Calculate the distance between each data instance and cluster centre.
3. Assign a data instance to a cluster whose distance from the cluster centre is minimum of all the cluster centres.
4. Recalculate the new cluster centre using formula 10
5. Recalculate the distance between each data instance and newly obtained cluster centres.
6. If no data instance was reassigned then stop else repeat from step 3.

$$c_i = (1/k_i) \sum x_i \quad (10)$$

The working of K-Means is pictorially represented in figure 2.

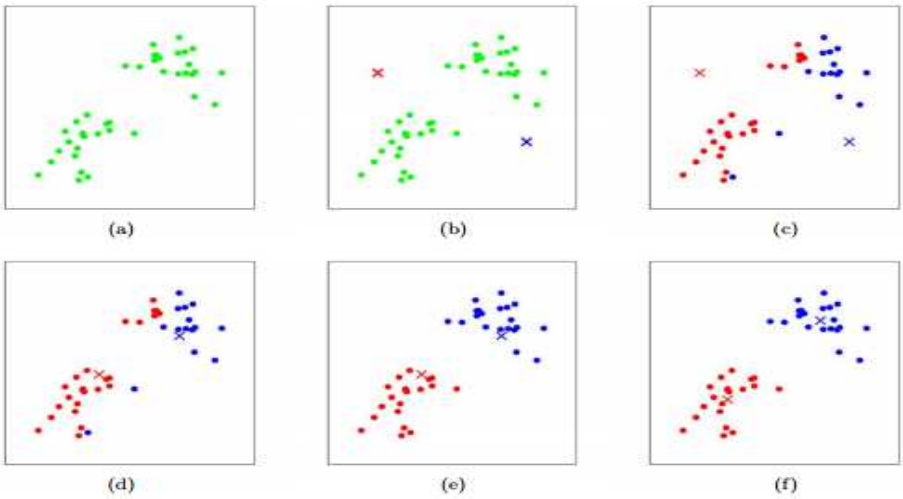


Fig. 2 . Depicts the different phases of the K-Means algorithm for k=2

**K Nearest Neighbor**

The training data instances are vectors in a multi dimensional feature space each with the respective Target vector i.e. the class labels. The training phase of this algorithm only consists of storing feature values and target vectors of training data instances [7]. During the classification phase a new data instance is classified to one of the available classes based on the classes of the k nearest neighbours of the new data instance i.e. among the k nearest neighbours the maximum number of data points belonging to the same class is given prominence and the corresponding class with maximum nearest neighbours is assigned to the new data instance. Given a data instance, the k nearest neighbours is determined using the Euclidean distance formula. For two data instances x1 and x2 in R-dimensional feature space the formula to determine the distance between them is depicted in formula 11. The impact of the right value of k can be better understood by the figure 3. In figure 3, the data instance marked by red coloured star is classified to Class B if k=3 and the same data instance is classified to Class A if k=6.

$$d(x_1, x_2) = || x_1 - x_2|| = \text{sqrt}( \sum (x_1 - x_2)^2 ) \quad (11)$$

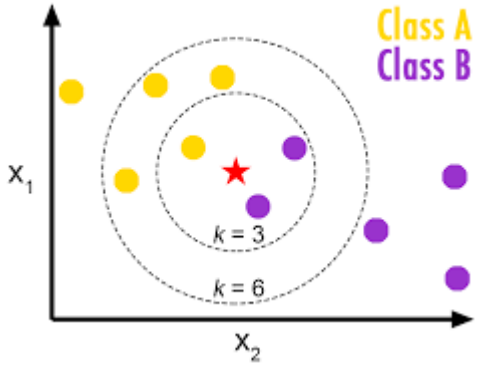


Fig. 3. Depicts the classification of KNN algorithm for k=3 and k=6

## III. PROPOSED METHODOLOGY

Every model discussed so far had implicit drawbacks in the system which restrained the system from performing at higher accuracies. So, a better model would be the one taking the positive aspects of all these systems for decision making. In the proposed model SVM-KNN, all the three models i.e. SVM, K-Means and K Nearest Neighbor algorithms are combined to produce results with high accuracy. The main objective of the SVM-KNN algorithm is to achieve high accuracy and also to guarantee the generalization performance. SVM-KNN is a comprehensive model, whose decision making is decentralized i.e. classification is made at multiple levels and finally based on the results thus obtained the final decision is made. SVM-KNN has two phases of working which are Training phase and Predicting phase. In the Training phase, like any other classification model the set of data instances along with their target vectors is recorded and corresponding distinct classes are formed. And during the Predicting phase, to determine the class of the new data instance, prediction is made solely in the SVM model and then solely in the K-Means algorithm and the corresponding results are compared. If both SVM and K-Means algorithm gives the same result i.e. if both the methods decide the new data instance to belong to same class then the new data instance is marked with that corresponding class. Else if, SVM and K-Means algorithm give two different results then K Nearest Neighbor algorithm is called with parameter k in the range [2, n],where n is the number of data instances accounted in the model. Initially k is set to 2, later in the future, during every iteration k is multiplied with 2 and then KNN is called for the corresponding k value. This way KNN algorithm is called multiple times and if we get the same class as the preferred class for at least two consecutive times then the corresponding class is marked for the respective new data instance. This ensures higher chances of associating the new data instance with the right class. With this, the number of misclassifications can be reduced by a significant factor.

### A. SVM – KNN Algorithm

Let X = {x1, x2, x3...xn} be the 'n' number of data instances in a R dimensional sample space and let Y = {y1, y2, y3...yk} be the number of classes and let 'z' be the test sample in R dimensional sample space whose class is to be determined.

1. Train the SVM model with the data set X.
2. Train the K-Means model with the data set X
3. Predict the class of z in SVM model and Predict the class of z in K-Means model, If both the models predict the same class then mark z as belonging to the corresponding class and exit.
4. If SVM and K-Means model predict z as belonging to different classes then call KNN algorithm.
5. KNN model is called with k(Initially k=2), the result is stored in **Temporary_Result**.
6. Update k = k*2 and verify if k is less than 'n', if True then repeat step 5. If two consecutive calls of KNN returns the same class then mark the respective class for the corresponding data instance z and exit.
7. If no two consecutive calls made the same decision on class, then return the class which was chosen maximum number of times in all the previously made calls to KNN and exit.
8. If all the previously made calls to KNN gave unique result in each iteration then return the class chosen by SVM and exit.(This is the worst case situation of the SVM-KNN algorithm)

This algorithm ensures to assign a new data instance to most favourable class. SVM-KNN necessitates multi stage verification and thus the chances of the new data instance falling into wrong class is hindered to a maximum possible extent. Also, most importantly in SVM-KNN an effort is made to correctly classify most of the data instances which are misclassified in existing classification methods. Consider a situation as shown in figure 4. There are two classes one represented by blue coloured cross symbols named as BLUE class and other by red coloured cross symbols named as RED class. The red coloured line is the hyper plane constructed by SVM algorithm and the green coloured line is a hypothetical line drawn depicting the decision boundary of K-means algorithm.

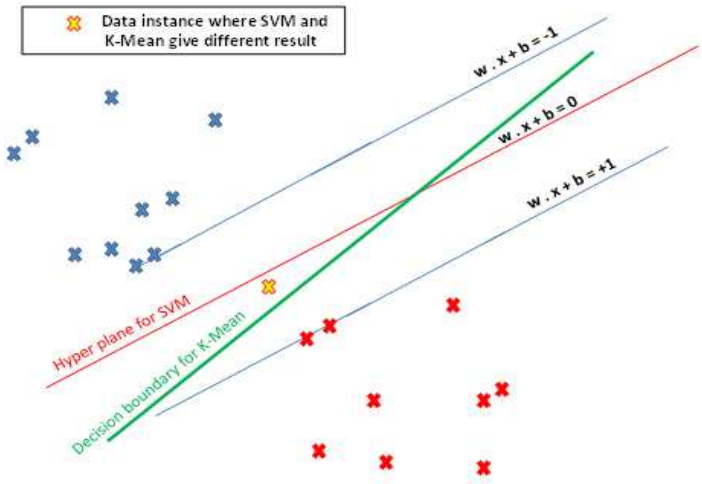


Fig. 4. Depicts the possible scenario where SVM and K-Means algorithm give different results

Consider a new data instance represented by yellow coloured cross symbol. The new data instance is classified as belonging to RED class by SVM algorithm and the same data instance is classified as belonging to BLUE class by K-Means algorithm. According to SVM-KNN, in this type of situation KNN algorithm is called for k values 2, 4, 8....n. From figure 5, 6 and 7, It can be observed that for k=2 SVM-KNN algorithm decides RED class for the new data instance and, for k=8 and k=16 SVM-KNN algorithm decides BLUE class. Furthermore when any two consecutive calls to KNN gives the same class as preferred class then the new data instance is marked with the corresponding class. In our example for k=8 and k=16 (two consecutive calls for KNN) gives BLUE as the preferred class and hence the new data instance is classified as belonging to BLUE class.

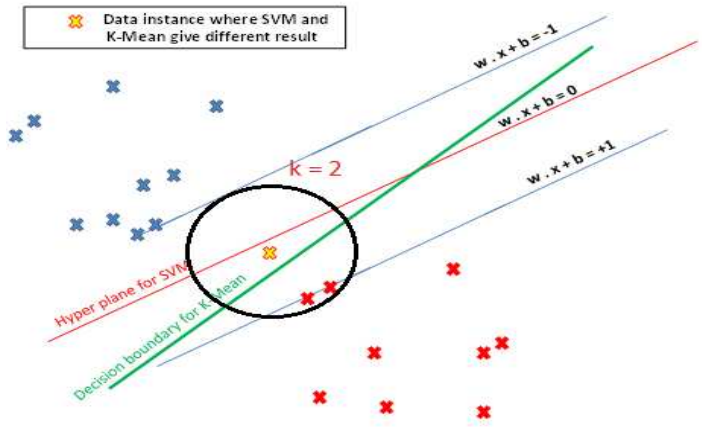


Fig 5. SVM-KNN for k=2

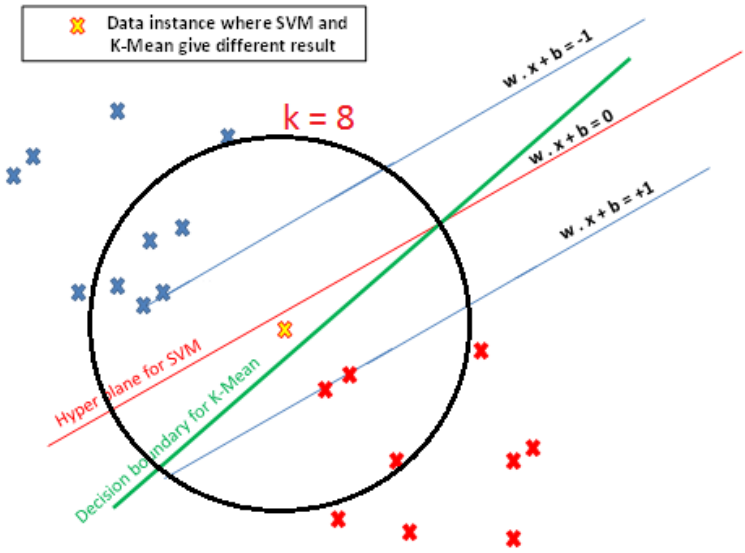


Fig 6. SVM-KNN for k=8

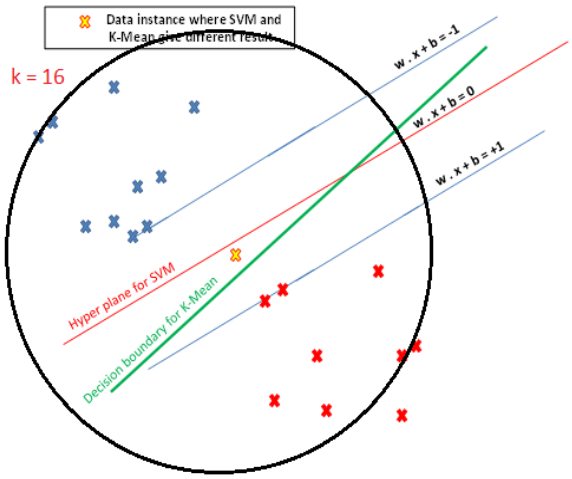


Fig 7. SVM-KNN for k=16

## IV. EXPERIMENTATION

The proposed algorithm SVM-KNN along with Support Vector Machine and K Nearest Neighbor was implemented in a PC which satisfy basic software and hardware requirements. Also, sk learn library and Ipython notebook was used for the implementation. In this experimentation four different data sets were used – Liver Disorder, Heart data, Diabetes data and Satellite data. The data for training and testing was obtained from Repository of machine learning databases of University of California at Irvine (UCI)[8]. The results of our experiment are shown in the Table 1. A comparison of accuracy of existing methods (SVM and KNN) and proposed method is depicted in the Table 1. From Table 1 It is clearly evident that the proposed methodology i.e. SVM-KNN has achieved better accuracy than existing methods. The graph of accuracy versus number of data instance for SVM, KNN and SVM-KNN is shown in Figure 8. Also, one important observation that can be made from the Table 1 is that, In the worst case SVM-KNN would give an accuracy same as SVM, as expected.

**TABLE I - EXPERIMENTAL RESULTS**

| Application | Feature | Training Data | Testing Data | Classes | Accuracy | | |
|---|---|---|---|---|---|---|---|
| | | | | | KNN(%) | SVM(%) | SVM -KNN(%) |
| Liver Disorder | 6 | 230 | 115 | 2 | 60.0000 | 63.4782 | 64.3478 |
| Heart Data | 13 | 260 | 70 | 2 | 75.7142 | 82.8571 | 82.8571 |
| Diabetes Data | 8 | 500 | 200 | 2 | 70.0000 | 80.5000 | 82.0000 |
| Satellite Data | 12 | 4435 | 2000 | 7 | 90.4000 | 91.8000 | 91.8000 |

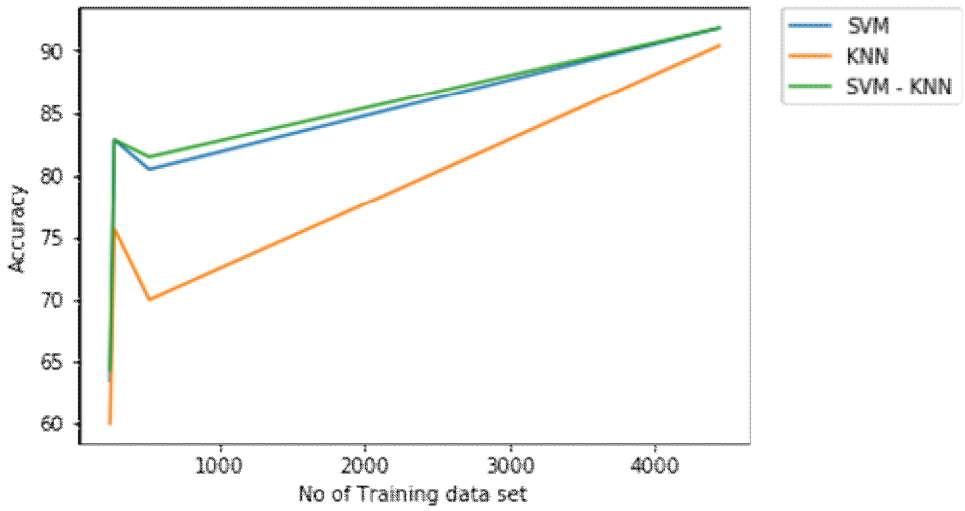


Fig.8. Graph of accuracy versus number of data instances

## V. CONCLUSIONS

In this paper, I intend to build a new model which offers better accuracy than the accuracy obtained by using existing methodologies for classification. And it is evident from the results obtained that my intended objective to maximize the accuracy is achieved.

By using SVM-KNN, the newly proposed model the accuracy of the model can be increased to considerable extent in most of the cases. In the future, SVM- KNN can be optimized to give better results for higher dimensional feature space. Also, other classification algorithms can be explored to increase the accuracy of the system.

## REFERENCES

1. Yukai Yao, Yang Liu, Yongqing Yu, Hong Xu, Weiming Lv, Zhao Li, Xiaoyun Chen, " K-SVM: An Effective SVM Algorithm Based on K-means Clustering", Journal of computers, pp.2632-2640, Vol.8 No.10, October 2013.
2. Jing Bai, Lihong Yang, Xueying Zhang, "Parameter Optimization and Application of Support Vector Machine Based in Parallel Artificial Fish Swarm Algorithm", Journal of Software, pp. 673-679, vol. 8, no. 3,2013.
3. Riadh Ksantini, Boubakeur Seddik Boufama, Imran Shafiq Ahmad, "A New KSVM + KFD Model for Improved Classification and Face Recognition", Journal of Multimedia, pp. 39-47, vol. 6,no. 1,2011.
4. Unnati R. Raval, Chaita Jani, "Implementing and Improvisation of K-means Clustering", International Journal of Computer Science and Mobile Computing, pp 72-76,Vol.4 Issue 11, November – 2015.
5. R.O.Duda and P.E.Hart, "Pattern Classification and Scene Analysis", New York:John Wiley & Sons, 1973.
6. Jyoti Yadav , Monika Sharma, "A Review of K-mean Algorithm", International Journal of Engineering Trends and Technology, pp 2972-2977, Vol .4 Issue.7, July-2013.
7. Jinho Kim, Byung-soo kim, Silvio Savarese, "Comparing Image Classification Methods: K-Nearest Neighbor and Support Vector Machine", Applied Mathematics in Electrical and Computer Engineering, pp.133-139.
8. C. Blake, C. Merz, "UCI Repository of Machine Learning Databases", Available: http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998[Online]


 