Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 8, August -2017

Comparative Study of K-Nearest Neighbor Classification and J48 Decision Tree Algorithm with and without Clustering Considering Different Data Parameters

Vaibhav Sharma[1], Shrwan Ram[2]

[1]*Department of Computer Science and Engineering MBM Engineering College, Faculty of Engineering, Jai Narain Vyas University, Jodhpur (Rajasthan*
[2]*Department of Computer Science and Engineering MBM Engineering College, Faculty of Engineering, Jai Narain Vyas University, Jodhpur (Rajasthan*

Abstract— *Data mining is an area where computer science, machine learning and statistics meet and where the goal is to discover and extract information such as relations and patterns that's hidden inside the data. The volume of data is increasing exponentially and analyzing such a large volume of data has become one of the big challenges for IT industries. The data has become the asset of every enterprise. The mining of such large volume of data provides the valuable information regarding to the specific field for which data are collected. There are many types of data mining techniques available and used to extract the valuable hidden patterns from the large volume of data. The patterns extracted from the data become the part of knowledge base for the decision support system. The main goal of the data mining is to find out the relevant and more valuable information from the data and building the knowledge base. In this paper we are considering the K-means Clustering algorithm for classifying the data on the basis of similarity. This is one type of the unsupervised machine learning technique. The Clusters produced by the K-means clustering are further classified using Supervised machine learning techniques, Such as K-nearest Neighbour method and decision tree algorithm.*

Keywords— *Data mining, hidden patterns, knowledge base, decision support system, K-means Clustering algorithm, Unsupervised machine learning, Supervised machine learning, Knearest Neighbour method, decision tree algorithm.*

I. INTRODUCTION

Data are playing the major role in every field nowadays; the drastic changes had been accrued in the field of information technology. The volume of data is increasing exponentially in the field of banking systems, social networks and internet logs. The data is too complex and analysis of data is become one of the challenging task. Initially, with the advent of computers and means for mass digital storage, we started collecting and storing all sorts of data, counting on the power of computers to help sort through this amalgam of information. Unfortunately, these massive collections of data stored on disparate structures very rapidly became overwhelming. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The figure 1.1 as shown below depicts the data mining as a step in an iterative knowledge discovery process [10].

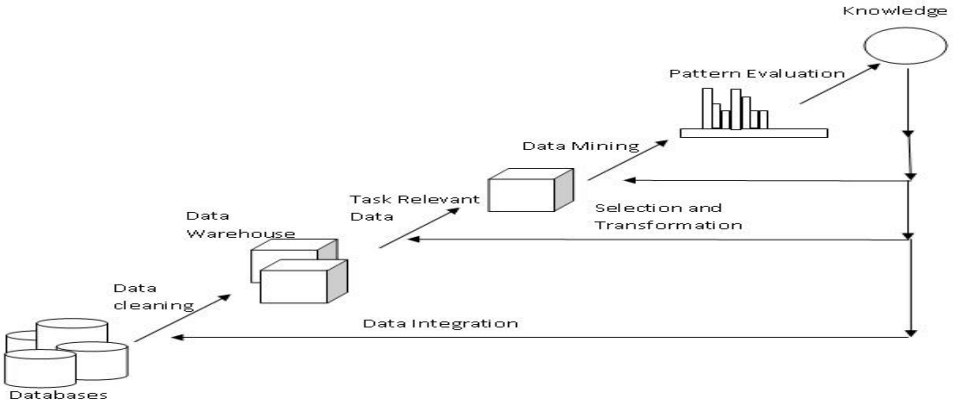

Figure: 1. Data Mining Knowledge Discovery Process [10]

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

- **Data cleaning:** also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection [1].
- **Data integration:** at this stage, multiple data sources, often heterogeneous, may be combined in a common source [1].
- **Data selection:** at this step, the data relevant to the analysis is decided on and retrieved from the data collection [1].
- **Data transformation:** also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure [1].
- **Data mining:** it is the crucial step in which clever techniques are applied to extract patterns potentially useful [1].
- **Pattern evaluation:** in this step, strictly interesting patterns representing knowledge are identified based on given measures [1].
- **Knowledge representation:** is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results [1].

1.1 Knowledge Discovery in Data Mining

Classification and the clustering are the knowledge discovery process used for extracting hidden patterns. Data analysis is the process of organizing and classifying the data in given classes. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects. For example, after starting a credit policy, the Video Store managers could analyze the customers' behaviors vis-à-vis their credit, and label accordingly the customers who received credits with three possible labels "safe", "risky" and "very risky". The classification analysis would generate a model that could be used to either accept or reject credit requests in the future [11].

Prediction has attracted considerable attention given the potential implications of successful forecasting in a business context. There are two major types of predictions: one can either try to predict some unavailable data values or pending trends, or predict a class label for some data. The latter is tied to classification. Once a classification model is built based on a training set, the class label of an object can be foreseen based on the attribute values of the object and the attribute values of the classes. Prediction is however more often referred to the forecast of missing numerical values, or increase/ decrease trends in time related data. The major idea is to use a large number of past values to consider probable future values [11].

Clustering:- clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called unsupervised classification, because the classification is not dictated by given class labels. There are many clustering approaches all based on the principle of maximizing the similarity between objects in a same class (intra-class similarity) and minimizing the similarity between objects of different classes (inter-class similarity) [12].

II. DATA MINING TECHNIQUES

Data mining includes a variety of data analysis and extraction methods, these methods are in accordance with the different emphases of the mining tasks and can be divided into descriptive and predictive model. The most popular data mining algorithms used in various fields are association rules, classification and prediction, cluster analysis, sequence mode. Choosing the most suitable mining algorithm, the key point is that based on the structure of data and applications required to achieve the target, a combination of variety of algorithms can also be considered to achieve mining target [4].

2.1 **Machine Learning: Machine Learning** is a sub-field of data science that focuses on designing algorithms that can learn from and make predictions on the data. Machine learning includes Supervised Learning and Unsupervised Learning methods.

2.1.1 Supervised Learning: Most of the research in machine learning has focused on supervised learning. Supervised learning is provided with a training set which consists of a set of labeled examples $\{ xi + yi \}$ i = 1 to N , where $xi \in X$ denotes the input objects (or data points) and $yi \in Y$ denotes the output value associated with xi. This training set is used to learn a function $f : X \rightarrow Y$ whose output can be either continuous (regression), or can predict the label of the input object (classification). Supervised learning includes two categories of algorithms [3], e.g. Support vector machines, Neural networks, Decision trees and Nearest neighbors (*k*-NN)

2.1.2 **Unsupervised Learning:** A somewhat less explored area in Machine learning is unsupervised learning. The task is to analyze a set of input objects { xi } i = 1 to N for which no class labels yi are provided. This area includes a wide variety of different learning tasks such as data clustering, feature extraction, visualization, density estimation, anomaly detection, information retrieval etc. Approaches of unsupervised learning include [3].

III. RESEARCH METHODS

The engineering design process is based on the scientific approach to problem solving. The distinguishing characteristic of engineering is that it uses a systems perspective. It studies a problem environment to implement corrective solutions that take the form of new or improved systems. Engineering design process has six steps as below:

- **Identification of a need or opportunity:** The first step in problem-solving is the identification of a need or opportunity. For any research field, these needs and opportunities are extensive and varied. Data Analytics is a broad area of specialty among the engineering disciplines. Data mining can help engineering process and analyze information, deciding on the most effective data mining techniques and systems can be complicated.

- **Problem Definition:** Thus, a data mining methodology to meet the specific requirements of Knowledge Discovery is needed. Such a methodology should assist data analytics approach in selecting appropriate data mining tools and implementing data mining projects from a systems perspective.

- **Data Collection:** In this paper we have used Turkiye Student Evaluation Data Set for analysis. This data set contains a total 5820 evaluation scores provided by students from Gazi University in Ankara (Turkey). There are a total of 28 course specific questions and additional 5 attributes. Total 33 attributes are used in the dataset.

- **Analysis of Alternatives**: There are several different data mining methodologies, but there is no one standard methodology for applying data mining. Consequently, several vendors have created their own proprietary methodologies. These have some drawbacks. Software vendors have designed approaches that are strongly correlated with the design of their own solutions and software packages. There are many data mining tool available for analyzing the data sets.

IV. DATA MINING TECHNIQUES USED FOR DATA CLASSIFICATION

4.1 Data mining techniques used for experimental purpose:

As a data mining function, clustering can be used for distribution of data, to observe the characteristics of each cluster, and to focus on a particular set of clusters for further analysis. Clustering is one of the most fundamental issues in data recognition. It plays a very important role in searching for structures in data. It may serve as a pre-processing step for other algorithms, which will operate on the identified clusters. In general, clustering algorithms are used to group some given objects defined by a set of numerical properties in such a way that the objects within a group are more similar than the objects in different groups.

4.1.1 **k-means clustering:** The k-means clustering algorithm uses the Euclidean distance to measure the similarities between objects. Both iterative algorithm and adaptive algorithm exist for the standard k-means clustering. K-means clustering algorithms need to assume that the number of groups (clusters) is known a priori. An important step in clustering is to select a distance metric, which will determine how the Similarity of two elements is calculated.

Basic Euclidean Distance Formula used in K-means algorithm

Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

i. Select 'c' cluster centers randomly.

ii. Calculate the distance between each data point and cluster centers using the Euclidean distance metric as follows:

Euclidean Distance:

$$D_{Euclidean}\ (x_i, x_j) = \sqrt{(x_i - x_j)^2} = \sqrt{\sum_{k=1}^{d} (x_{ik} - x_{jk})^2} \quad (1)$$

iii. Data point is assigned to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
iv. New cluster center is calculated using:

$$v_i = (\frac{1}{Ci}) \sum_1^{Ci} x_i \qquad (2)$$

where, c_i denotes the number of data points in i^{th} cluster.
v. The distance between each data point and new obtained cluster centers is recalculated.
vi. If no data point was reassigned then stop, otherwise repeat steps from iii to v.

4.1.2 J48 Decision Tree

This algorithm generates the rules for the prediction of the target variable. With the help of tree classification algorithm the critical distribution of the data is easily understandable. J48 is an extension of ID3. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In the WEKA data mining tool, J48 is an open source Java implementation of the C4.5 algorithm.

Basic Steps in the Algorithm:
(i) In case the instances belong to the same class the tree represents a leaf so the leaf is returned by labelling with the same class.
(ii) The potential information is calculated for every attribute, given by a test on the attribute. Then the gain in information is calculated that would result from a test on the attribute.
(iii) Then the best attribute is found on the basis of the present selection criterion and that attribute selected for branching.

- **Counting Gain**
  This process uses the "Entropy" which is a measure of the data disorder. The Entropy of is calculated by

$$\text{Entropy}(\vec{y}) = -\sum_{j=1}^{n} \frac{|y_i|}{|\vec{y}|} \log\left(\frac{|y_i|}{|\vec{y}|}\right) \qquad (3)$$

$$\text{Entropy}(j|\vec{y}) = \frac{|y_i|}{|\vec{y}|} \log(\frac{|y_i|}{|\vec{y}|}) \qquad (4)$$

And the Gain is

$$\text{Gain}(\vec{y}, j) = |\text{Entropy}(\vec{y}) - \text{Entropy}(j|\vec{y})| \qquad (5)$$

The objective is to maximize the Gain, dividing by overall entropy due to split argument by value j.

4.1.3 k-Nearest Neighbor: K-Nearest neighbor is instance based lazy learner. Because it stores all the training instances and delays the process of model building until test is given for classification. It is used for prediction. For describing training tuples, n-numeric attributes are used. N-dimensional space is used to store all training examples. When a test sample is available, K nearest neighbor algorithm finds the k training samples that are closest to that test sample. Different distance functions are used to search difference between training and testing samples which is describe in next section. K nearest neighbor algorithm steps is given below:

i. Determine k nearest neighbors and D set of training example.
ii. for each test example xi do Calculate d (xi , yi) based on distance measure.
iii. Select the k closest training examples yi to test example xi.
iv. Use majority voting to classify the test examples
v. End step ii

Distance measures are very essential to find the similarity and dissimilarity between data points. Similarity is the measure of two objects is alike. And dissimilarity is two objects are different. Euclidian Distance K-Nearest Neighbor can be calculated by using Euclidian distance. It gives efficiency and productivity. It is a distance between two points in Euclidian space. It computes the root of square differences between co-ordinates of pair of data points. The distance can be calculated using equation 1 as shown above.

V. EXPERIMENTAL SETUP AND RESULTS

The basic concept of data extraction is to identify what kind of class a certain data point belongs to based on the features that the point possesses.

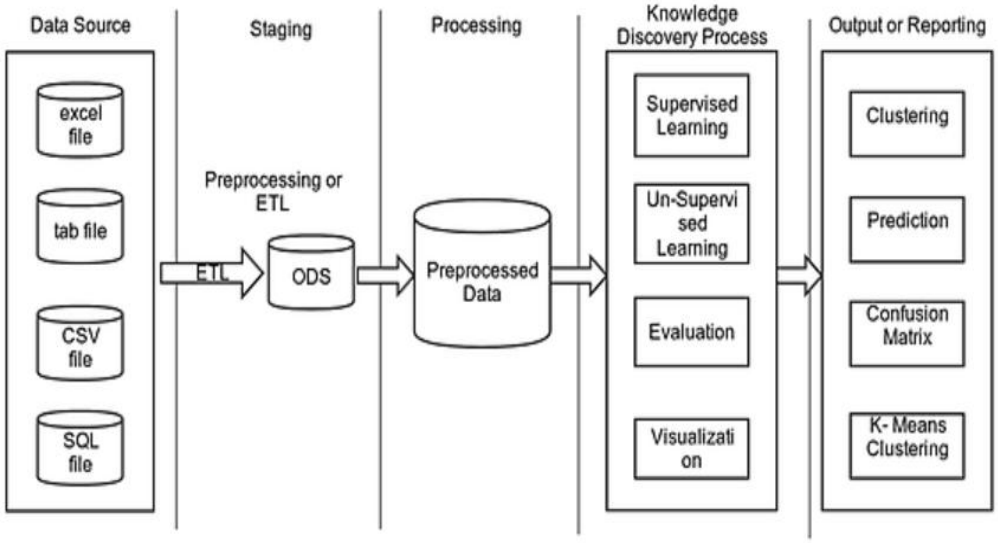

Figure: 2 Data Mining Technique and Data Extraction for Experiment

5.1 Data Set used for experimental purpose

Large non-generated data sets with usable data in the form of personal data records was hard to find for free, however during this thesis i received access to a relatively large registries to analyze. I had used Turkiye Student Evaluation Data Set for the thesis. This data set contains a total 5820 evaluation scores provided by students from Gazi University in Ankara (Turkey). There is a total of 28 course specific questions and additional 5 attributes.

Table 1. Data Set Description

Data Set Characteristic	Multivariant	**Number of Instances:**	5820	**Missing Values?**	N/A
Attribute Characteristic	N/A	**Number of Attributes**	33	**Associated Tasks:**	Classification, Clustering

5.2 Data Analytics tool (WEKA):-

There are many tools available for data mining and machine learning, but this thesis I choose to use the open source software suite WEKA which stands for Waikato Environment for Knowledge Analysis. The main reason why I selected to use WEKA was because of its versatility. WEKA is a popular tool used for data analysis, machine learning and predictive modeling that was developed by the University of Waikato in New Zealand using the programming language JAVA.

5.2.1 File Format used in WEKA:-

An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files have two distinct sections. The first section is the **Header** information, which is followed the **Data** information.

VI. RESULTS AND CONLUSION

The performance analysis is carried out on the turkey student data set. The classification of the data with and without clustering is also carried out in this experiment. So, there are different scenarios in which analysis of the data set is considered and those scenarios are as follows:

6.1 **Cluster Analysis of Data Set:**

K-means clustering is used to divide the data set into different clusters, the Euclidean distance (default) or the Manhattan distance can be used and we used the Euclidean distance for the clustering purpose. The below figure shows the result of K-means clustering applied to the Dataset. There are total eight clusters are used for dividing the dataset.

```
kMeans
======
Number of iterations: 15
Within cluster sum of squared errors: 5690.527369758503
Missing values globally replaced with mean/mode
Cluster centroids:
                        Cluster#
Attribute     Full Data         0         1         2         3         4         5         6         7
                 (5820)      (459)     (996)     (829)     (593)     (665)     (725)     (823)     (730)
=============================================================================================================
instr            2.4856    2.6863    2.4327         3    2.6914    2.6451    2.5462    2.3329    1.6466
class            7.2763    7.4815    7.0261    6.7708     7.317    7.3173    7.7117    7.3426    7.4849
nb.repeat        1.2141     1.207      1.26    1.2171    1.1855    1.2932    1.2193    1.2029     1.111
attendance       1.6756    2.2549    2.2319    1.7865    0.2445    1.6316    0.9628    1.9478    2.0301
difficulty       2.7835    3.1634    3.6476    2.8589    1.3558    2.8737    2.2828    2.7388    2.9055
Q1               2.9299     1.878    2.9297    3.6405    2.7909     1.806    1.1117    4.7497    3.6753
Q2               3.0739    2.4052     3.004    3.8552    2.9056    1.9008    1.0883    4.8651    3.8603
Q3               3.1787    3.0196    3.0331    3.8951    2.9696    2.1609    1.1034    4.8931     3.889
Q4               3.0825    2.4118    3.0241    3.8589     2.946    1.9639    1.0883    4.8505    3.8192
Q5               3.1058    2.4793    3.0241    3.9204    2.9528    1.9053     1.051    4.9064    3.9151
Q6               3.1074    2.5773    3.0432    3.8854    2.9815    1.9624    1.0469    4.9101    3.8041
Q7               3.0663    2.2745    3.0191     3.848    2.9612    1.8827    1.0566    4.9077    3.8247
Q8               3.0419    2.2135     2.989    3.8166    2.9309    1.8887     1.051    4.8834    3.7973
Q9                3.166    2.7691    3.0151    3.9156    3.0118    2.1534    1.1255    4.8931    3.8973
Q10              3.0907    2.4466      2.99    3.9059     2.946    1.8917    1.0276    4.9198    3.9041
Q11              3.1838    2.9782    3.0462    3.9119    2.9831    2.0782      1.08    4.9052    3.9932
Q12              3.0356    2.3137    2.9729    3.7853    2.9191    1.8917    1.0566    4.8566    3.7726
Q13              3.2428    3.5926    2.9859    4.0157    2.9831    2.0496    1.0497    4.9538    4.0425
Q14              3.2909    3.8083    3.0181    4.0603    3.0236    2.1684      1.04    4.9599    4.0575
Q15              3.2873    3.7669     3.006    4.0567    3.0236     2.212    1.0372    4.9635    4.0342
Q16              3.1696    3.1068    2.9418    3.9891    2.9882    1.9008    1.0193    4.9429    4.0288
Q17              3.3985    4.2745    3.1064    4.0639    3.0438    2.5474    1.1007    4.9599    4.0753
Q18              3.2225    3.4967    2.9669    3.9843    2.9545    2.0195    1.0372    4.9417    4.0795
Q19              3.2617    3.6841      2.99    4.0434    2.9831    2.0962    1.0441    4.9465    4.0699
Q20              3.2854    3.9259    2.9729    4.0881    3.0101    2.1143    1.0276    4.9611    4.0411
Q21              3.3074    4.0545    2.9739    4.0965    2.9983    2.1774    1.0524    4.9611    4.0521
Q22              3.3175    4.0174    2.9839    4.1001    2.9933    2.2361     1.051    4.9696    4.0808
Q23              3.2019    3.3464    2.9488    4.0072    2.9359    2.0045    1.0303    4.9563    4.0274
Q24              3.1668    3.1329    2.9307    3.9662    2.9444    1.9789     1.029    4.9283    4.0027
Q25              3.3125    3.9608     2.989    4.0941    3.0034    2.2872    1.0469    4.9587    4.0384
Q26              3.2222    3.4379    2.9408    4.0109    2.9949    2.1173      1.04    4.9235    4.0151
Q27              3.1548    3.1089    2.9147    3.9433    2.9933    2.0316    1.0303    4.8919    3.9219
Q28              3.3081    3.9564      2.99    4.0808    3.0253    2.2707    1.0469    4.9368    4.0411
```

Figure 3. Cluster Analysis of Dataset

6.2 Generating Decision Tree:

The Decision tree for the J48 is generated and result is analyzed. There are two scenarios which generate the J48 tree and comparison of two scenarios shown in below table:

Table 2. Decision tree J48 Result Comparison

Result without clustering	Number of Leaves	324
	Size of the tree	647
Result with clustering	Number of Leaves	265
	Size of the tree	529

6.3 K Nearest Neighbor Classifier:

It's examined that testing of K-nearest neighbour using ten folds of the algorithm has been done and it's found that in each fold the result for the following parameters are improved:

Table 3. Improved Attributes of KNN Algorithm

Correctly Classified Instances	Increased
InCorrectly Classified Instances	Decreased
TP Rate	Increased
FP Rate	Decreased
Precision	Increased
Recall	Decreased
FMeasure	Increased
ROC Area	Increased

6.4 Conclusion: It is found that accuracy of decision tree and K-Nearest Neighbour method can be improved by using the clustering techniques compared to classifying the data without clustering.

REFERENCES

[1]. Adeniyi, D. A., Z. Wei, and Y. Yongquan. "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method.",PP (90-108) Applied Computing and Informatics 12, no. 1 (2016).
[2]. Bhagat, Amol, Nilesh Kshirsagar, Priti Khodke, Kiran Dongre, and Sadique Ali. "Penalty Parameter Selection for Hierarchical Data Stream Clustering.",PP(24-31), Procedia Computer Science 79 (2016).
[3]. Ravneet Kaur, and Sarbjeet Singh. "A survey of data mining and social network analysis based anomaly detection techniques.",PP(199-214), Egyptian Informatics Journal (2015).
[4]. Kaur, Manpreet, and Shivani Kang. "Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining.", PP(78-85), Procedia Computer Science 85 (2016).
[5]. A. Naik and L. Samant, "Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime",pp.(662-668), Procedia Computer Science, 85,2016.
[6]. Baitharu Tapas Ranjan, and Subhendu Kumar Pani. "Analysis of Data Mining Techniques for Healthcare Decision Support System Using Liver Disorder Dataset.", P.P(862-870), Procedia Computer Science 85 (2016).
[7]. Anguera, A., J. M. Barreiro, J. A. Lara, and D. Lizcano. "Applying data mining techniques to medical time series: an empirical case study in electroencephalography and stabilometry.", PP(185-199), Computational and Structural Biotechnology Journal 14 (2016).
[8]. Bini, B. S., and Tessy Mathew. "Clustering and Regression Techniques for Stock Prediction.", PP(1248-1255), Procedia Technology 24 (2016).
[9]. Ristoski Petar and Heiko Paulheim. "Semantic Web in data mining and knowledge discovery: A comprehensive survey.", PP(1-22), Web Semantics: Science, Services and Agents on the World Wide Web 36 (2016).
[10]. Bahari T. Femina, and M. Sudheep Elayidom. "An efficient CRM-data mining framework for the prediction of customer behaviour.", PP(725-731), Procedia Computer Science 46 (2015).
[11]. Agrawal Shikha and Jitendra Agrawal. "Survey on Anomaly Detection using Data Mining Techniques.",PP(708-713), Procedia Computer Science 60 (2015).
[12]. Singh Sachin and Ashish Mishra. "Clustering analysis for large scale data sets." In Computing, Communication & Automation (ICCCA), 2015 International Conference on, PP(1-4), IEEE, 2015.

