M. Tech (Software Engineering)

Performance Analysis of clustering algorithm on Arbitrary Shape and Density Varying

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT
FOR THE AWARD OF DEGREE
OF

MASTER OF TECHNOLOGY
IN
SOFTWARE ENGINEERING

Submitted By
Avadh Naresh
Kushwaha
2K19/SWE/03

Under the supervision of

Mrs. SONIKA DAHIYA
Assistant Professor
Department of Computer Science and Engineering
Delhi Technological University

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(FORMERLY DELHI COLLEGE OF ENGINEERING)
SHAHABAD, DAULATPUR, BAWANA ROAD, DELHI – 110042

JUNE, 2021

Department of Computer Science and Engineering
Delhi Technological University
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, Avadh Naresh Kushwaha, 2K19/SWE/03, student of Master of Technology (Software Engineering), hereby declare that the Major Project-II Dissertation titled **"Performance Analysis of clustering algorithm on Arbitrary Shape and Density Varying"** which is submitted by me to the Department of Computer Science and Engineering, Delhi Technological University, Delhi in partial fulfillment of requirement for the award of degree of Master Of Technology (Software Engineering) is original and not copied from any source without proper citation. This work has not been previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition

AVADH NARESH KUSHWAHA
2K19/SWE/03

Department of Computer Science and Engineering
Delhi Technological University
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled **"Performance Analysis of Clustering algorithm on Arbitrary Shape and Density Varying"** which is submitted by Avadh Naresh Kushwaha, (2K19/SWE/03) to the Department of Computer Science and Engineering, Delhi Technological University, Delhi in partial fulfillment of requirement for the award of the degree of Master of Technology, is a record of project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

SONIKA DAHIYA
Digitally signed by SONIKA DAHIYA
Date: 2021.07.31 18:49:50 +05'30'

Mrs. SONIKA DAHIYA

(**Supervisor)**

Assistant Professor(DTU)

ACKNOWLEDGEMENT

The successful completion of any task would be incomplete without accomplishing the people who made it all possible and whose constant guidance and encouragement secured us the success.

First of all, I would like to thank the Almighty, who has always guided me to follow the right path of the life. My greatest thanks are to my parents who bestowed the ability and strength in me to complete this work.

My thanks are addressed to my mentor **Mrs. Sonika Dahiya**, Department of Computer Science and Engineering who gave me this opportunity to work in a project under her supervision. It was her enigmatic supervision, unwavering support and expert guidance which has allowed me to complete this work in due time. I humbly take this opportunity to express my deepest gratitude to her.

Avadh Naresh Kushwaha
MTech (SWE)-4thSem
2K19/SWE/03

ABSTRACT

Clustering is the process that dividing the data into such group that contain similar data in one group and other data in another group. In simple way it separates the data with similar characteristic to make a cluster.

In this several algorithms used for that can group the data by partition, hierarchy algorithm, k-means algorithm, FCM algorithm, FCM sigma algorithm, standard FCM algorithm, Grid-based algorithm.

Most popular algorithm k-means and FCM algorithm are used to partition the data into group, this two-algorithm having different approaches in k-means data will be included in one particular cluster whereas in FCM a data can be included in all existing cluster, here k means and FCM by default uses Euclidean distance measure.

Here we using different distance measure to evaluate the performance analysis of k means and FCM algorithm using cosine distance measure, correlation distance measure, city block distance measure used in various dataset based on k-means clustering and FCM algorithm.

TABLE OF CONTENTS

Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ML	:	Machine Learning
AI	:	Artificial Intelligence
FCM	:	Fuzzy c means
KML	:	K means algorithm
KDD	:	Knowledge Discover data
ANN	:	Artificial Neural Network
FR	:	Face Recognition
NN	:	Neural Network
SVM	:	Support vector Machine
RL	:	Reinforcement Learning
RF	:	Random Forests
MLC	:	Multi-Level classification
LR	:	Logistic Regression
STING	:	Statistical information Grid

CHAPTER-1
INTRODUCTION

Data mining has widely used application due to wide ease of use of enormous number of data and storage as per need. Data mining method widely used in different organization to retrieve the essential information such as education field, data science, business organization etc.

Data mining system can be consider as belonging to a relevant database extract, types of information extract, the technique used or the application adapted. It can tackle complex situation and that uses machine learning technique, statistic, and Artificial Intelligence to mined information to estimate upcoming event probability. Data mining is used for fraud detection, scientific discovery, marketing etc. data/pattern analysis, knowledge extraction, data gathering is called knowledge discovery in data (KDD). Data mining used in various technique machine learning, classification, outlier detection, regression, sequential pattern, decision tree, statistical technique, clustering.

1.1 Overview

1.1.1 Machine learning

Machine learning is an artificial intelligence (AI) technology that allows you to use data to automatically learn from previous experience. The main focus of machine learning to learn computer without Human intervention.to develop a computer such way that can access the data and use own. Machine learning algorithm used in various application such as speech Reorganization, computer vision, email filtering. It developed model based on essential data known as training data to make prediction or decision without being explicitly programmed to do so.

There are three types of machine learning techniques as follows:

1.1.1.1 Supervised Learning

Supervised Machine learning is a type of method in which model is trained using labeled training data and on the basis of this information it predicts the output value. The labelled data means some input data is already tagged with the correct output.

In machine learning, well-trained data enables managers to teach the machine to work

Predict the outcome correctly. In supervised learning technique it provides input and desired output to the machine learning model. It developed mathematical model set of data that consist input and the desired output that is known as “training data”. It learns a general rule that maps input to output, the training process continues until

The learning algorithm can also compare its results with the expected correct results and find errors to change the model accordingly. Accuracy can be finding through loss function until error has been minimized.

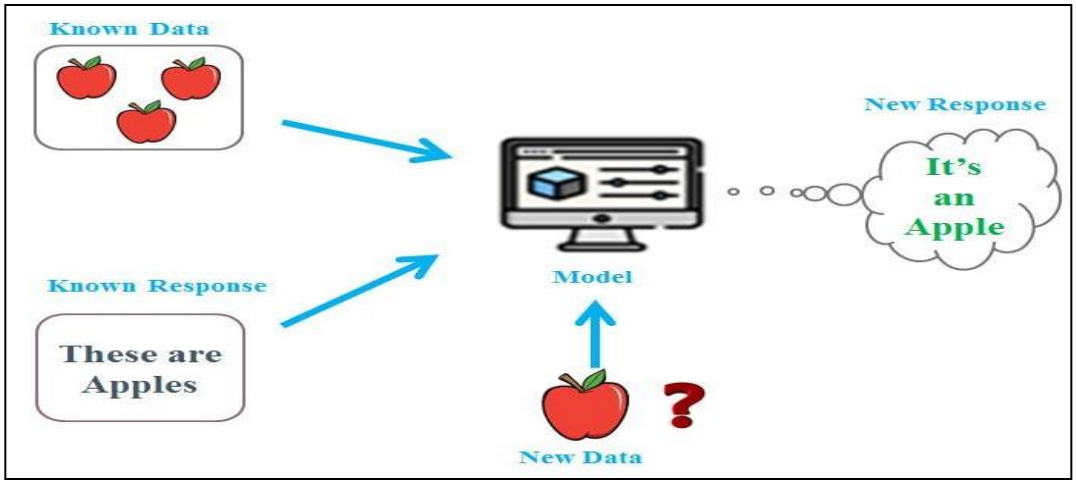


Figure 1.1: Schematic of Supervised Learning

Advantage of supervised algorithm:

- In ML supervised learning gathered data and produce output from previous experience.
- It will minimize the performed based on previous experience.
- It helps you to solve complex or real-world problem.
- In supervised before giving the data for training, you can find out exactly how many classes are there.
- You know the exact classes in the training data.
- It very helpful in classification problem.
- You cannot save your training data in memory after entire training completed you can keep the decision boundary as a mathematical formula.
- It predicts the numerical value for the training data.
- You can easily understand the supervised learning process.

Disadvantage of supervised algorithm:

- Supervise learning give only relevant information not any unknown information from training data like unsupervised learning do.
- It can’t handle some of the complex task in ML because of limited in a variety of sense.
- Classifying bid data can be a real challenge.
- The supervised apprenticeship training requires a lot of computing time.

Supervised can be portioned into two problems:

- **Regression**: It is a popular algorithm derived from linear regression class. A single value is used to predict one or more output variable, consider that each input value not relate to each other. When you consider more than one variable to get output known as multiple linear regression. This model relates the relationship between the given feature and output which is the limitation.

Figure 1.2 linear Regression

- **Classification**: Classification is the process of dividing data into different categories. Sequence data and unstructured data can be executed. the whole process starts with predicting the group of given data point.
- **Face Recognition**: FR Reading facial images and then identifying them or mapping them with the database, is also an application of supervised learning.

Various Algorithm and technique are used in supervised machine learning process.

- **Neural Network**: A neural network is a sequence of algorithm generally focus on solving complex signal process or pattern recognition problem. It refers to system of neuron either organic or artificial in nature. Neural network can modify to changing the input so that it can generate most feasible result without needing to redesign the output criteria.

  It having three-layer input layer, hidden layer, output layer the input layer carry input pattern the output layer has output signal which input pattern map. For example, a pattern can consist of a set of technical indicators for security.

  It is used in a variety of financial services applications from forecasting and market research to fraud detection and risk assessment.

  Neural network totally similar to human brain neural network. In neural

network neurons is a mathematical function that collect the data and categorize information according to a certain architecture.

Applications of neural network are:

1. Financial operation
2. Trading
3. Business analytics
4. Product maintenance

- **Naïve Bayes:** Naïve Bayes method is a technique of supervised algorithm based on Bayes theorem. Given the value of a class variable, there is a conditional independence between each pair of characteristics. This means that the presence of the characteristic does not affect another characteristic in the probability of a given outcome, and every predictor having equal outcome. There are three type of Naïve Bayes theorem

    1. **Multinomial Naïve Bayes**
    2. **Bernoulli Naïve Bayes**
    3. **Gaussian Naïve Bayes**

- **Logistic Regression:** Logistic regression is a branch of machine learning it used when the dependent variable is categorical, i.e. ‘true’ and ‘false’ or ‘yes’ and ‘No’. Logistic regression is basically used to solve binary classification problem for example span identification.
- **Support Vector machine (SVM):** In ML SVM is a SL mode that monitoring the data for categorization and regression analysis it is very best prediction method based on statistical learning framework.

  In the SVM algorithm, we draw each element as an n-dimensional space with specific coordinate values. We do the experiment by finding out which hyperplane can really differentiate the two SVM groups - these are just the same observation systems. It can solve linear or non-linear complex problem and work for many real words problem. This algorithm draws a line or hyper-line which distributes the data into classes.

  Two type of Support vector machine:

    - Linear SVM: Linear SVM is used for data that are linearly separable

i.e. dataset can be arranged into two groups by using single straight line, and then such data is called linear SVM.

- Non-Linear SVM: Used for non-linear separation, i.e. since the data set cannot be organized in a straight line, this data type is referred to as non-linear SVM.

1.1.1.2 Unsupervised Learning

Unsupervised learning is a machine learning technique in which model learn pattern from untagged data. Untagged data given to the model it can find the pattern of similar data it basically behaves like human brain and learning new things. Unsupervised algorithm cannot directly apply to the regression or classification in supervised algorithm having the input and corresponding output. Suppose the unsupervised algorithm we are given the dataset which contain image of different type of cat and dogs. Model is not trained upon the dataset, that means doesn't have the idea regarding the dataset. The task of unsupervised algorithm to identify the image on their own. Algorithm will perform this task by clustering the image dataset into group according to similarities between images.

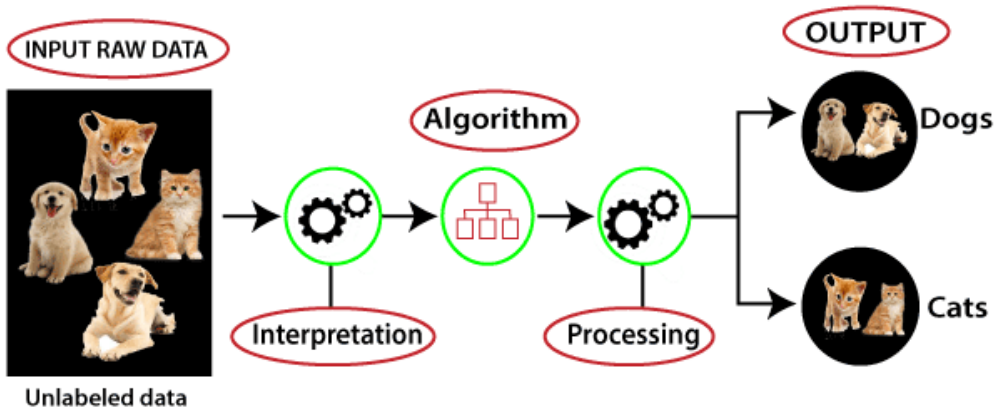


Figure 1.3: Schematic of Unsupervised Learning

Below is the list of algorithms used in unsupervised algorithm:

- K-means algorithm
- KNN
- Apriority algorithm

- Hierarchal algorithm
- Clustering
- Independent component analysis
- Principal component analysis
- Singular value decomposition

Advantage of unsupervised algorithm:

- In this algorithm applied for complex problem because data is unlabeled
- Recommend unsupervised learning because it is not easy to label Data and tagged data

Disadvantage of unsupervised algorithm:

- Unsupervised learning fundamentally hard than supervised algorithm Because it does not have corresponding output.
- The result of unsupervised learning algorithm is less correct as compared to supervised because data is unlabeled

1.1.1.3 Reinforcement Learning

Reinforcement learning is looked upon by many as learning from mistakes. When this learning algorithm is placed in a new environment, initially it will make many errors. When we provide feedback signals to the system for its output, we can reinforce our algorithm as required. Eventually, our learning algorithm learns to make less error as before.

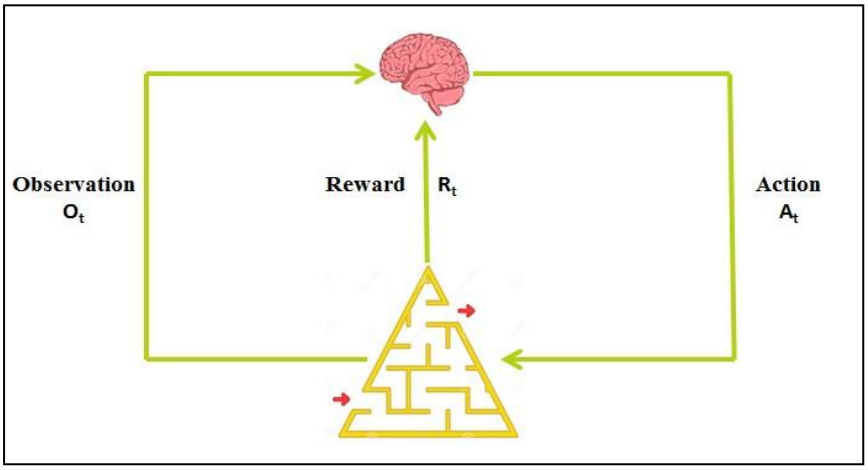


Figure 1.4: Reinforcement Learning Process

Application:

- **Video Games**: Learning to play videos games is one of the areas where reinforcement learning is exploited.

1.1.2 Classification

Classification is a predictive data mining technique which is used for data analysis task. In this model used to distinguish data classes and concept. In classification problem it identifies dataset belongs which categories a new observation belong training set data continues observing whose categories membership is known.

Classification is a method of predicting category labels such as "safety" and "risk" for project adoption. the output value of classification is a category not a value such as ‘red’ or ‘orange’, ‘fruit’ or ‘animal’ the main focus of classification algorithm to identify the category of a given dataset, and this algorithm used to predict the output for a category data.

Example of classification

1. It classifies it is spam or not
2. Provide handwritten character and classify it is one of the known characters.

In classification modeling perspective, it requires a well-formed dataset with many input and output examples for training. The model uses a completed training data set to calculate the best example of mapping input data, assign a category label, and this result-based

evaluation model. We can judge the performance of the model by the accuracy of the classification of the label in the popular metrics class.

Classification accuracy is not ideal, but it is a good starting point for many sorting tasks.

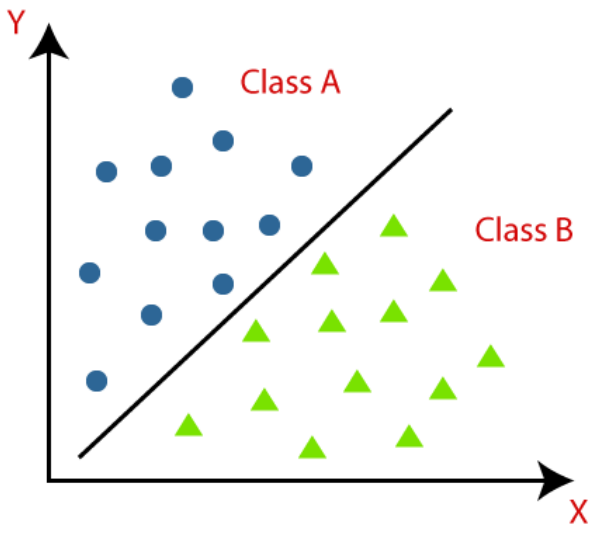


Figure 1.4 Schematic of classification

It is a two-step process:

1. **Learning step (training phases):** When creating a classifier model, test the model using the available tutorial set and use a variety of algorithms to create the classifier. The model to be trained to predict accurate results.
2. **Classification steps:** This model is used to predict low-category constructed category labels and test models to evaluate the accuracy of classification rules.

There are four type of classification task that you may encounter them are:

- **Binary classification:** Binary classification refers two cases one class that is normal state and another class that is the abnormal state.

  for example, one is ‘spam’ that is abnormal state and ‘Not Spam’ is the normal state another example is cancer not declared is the normal state of a task that involves a medical test and caner detected is the abnormal state. A class in normal state is defined as a class label of 0, and a class in abnormal state is defined as a class label of 1. It is a very common model to construct a binary classification that predicts the Bernoulli probability distribution for each example.

  The Bernoulli distribution is a discrete probability distribution it will cover case with binary output either 0 or 1.

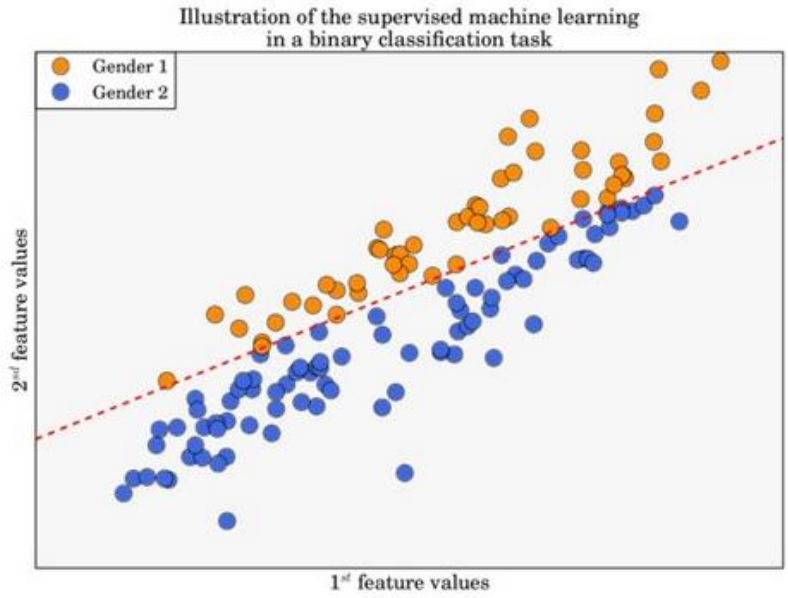


Figure 1.5 SML in a binary classification

Algorithms that can be used for binary classification include.

1. Logistic regression
2. K-Nearest Neighbors
3. Decision tree
4. Support vector machine
5. Naïve Bayes

Most of the algorithm specially outline for binary classification and not support more than two classes' example SVM, Logistic regression.

Example

1. Email spam
2. Churn prediction
3. Conversion prediction

- **Multi class classification:** In the case of multi-class classification, these are classification problems with more than two class names. In the case of multi-class classification, the classification as normal or abnormal is not taken into account. For some problem the number of classes label is very huge on some problem.

For example, a model predicts a photo as belonging to one among thousand or tens of thousands of faces in a face recognition system.

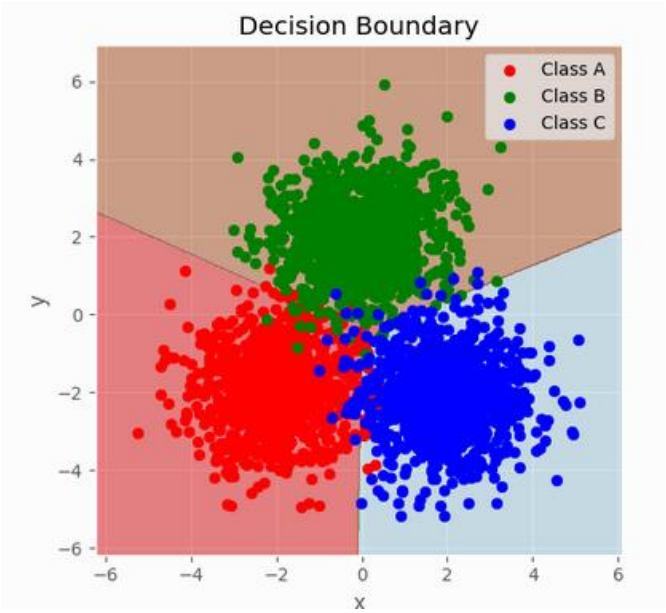


Figure 1.7 Decision Boundary 3 Cluster

Many binary classification algorithms can be used to classify multiple classes. Popular algorithms that can be used to categorize into multiple classes include:

1. K-Nearest Neighbors
2. Decision tree
3. Naïve Bayes
4. Random forest
5. Gradient Boosting

Algorithms developed for binary classification can be adapted to solve multiclass problems. This requires the use of multiple binary classification model strategies applicable to each class vs.-all other classes (called one-vs.-vest) or one model for each pair of classes (called on-vs.-one).

- one-vs.-rest: fit one binary classification model for each class vs. all other class
- one-vs.-one: Fit one binary classification model for each pair of classes

Binary classification algorithms that can use these strategies for multi-class classification include:

- Logistic regression
- Support vector machine

- **Multi-label classification**: MLC refers to that classification task that considers two or more class label, for each example more than one class may be predicted.

  In general, multi-label classification problems are created using models that predict multiple outputs, and each output is predicted as a Bernoulli probability distribution. Basically, it is a model that can make multiple binary classification predictions for each example.

  You can use the classification algorithm for binary or multi-class and multi-class classification, which specializes in standard classification algorithms, including the so-called multi-label version of the algorithm.

  - Multi-label decision tree
  - Multi-label random forests
  - Multi-label gradient boosting

  You can use the do multi-label Classification function () to generate a synthetic multi-label classification dataset.

- **Imbalanced classification:** Imbalanced classification is a classification task in which the number of classes is unevenly distributed. The unbalanced classification problem is a binary classification problem, in which most examples of the training data set belong to the normal class, and some belong to the abnormal class.

  Example includes:

  - Fraud detection
  - Outlier detection
  - Medical diagnostic test

  These problems belong to binary classification task these also required specialized technique. Specialized technique is a process to change the construction of sample in the training dataset by understanding the majority class or up sampling the minority class.

  It required more attention on minority class when filling the model on the training

dataset. For example, cost-sensitive machine learning algorithms can be misleading to minimize classification accuracy and may require alternative indicators.

Example:

- Random under sampling
- SMOTE oversampling

1.1.3 Outlier detection:

Outlier detection is a data mining technique used to define unusual pattern from a large dataset without prior knowledge of which object to look for.

This outlier detection provides information about local anomalies in the entire system. Therefore, it identifies processes worthy of outliers to provide information about the data set.

There are many detection method which remove outlier from a dataset. Outlier detection is also called anomaly detection used in much application like time series for casting, banking, health care, networking, capital markets, and more.

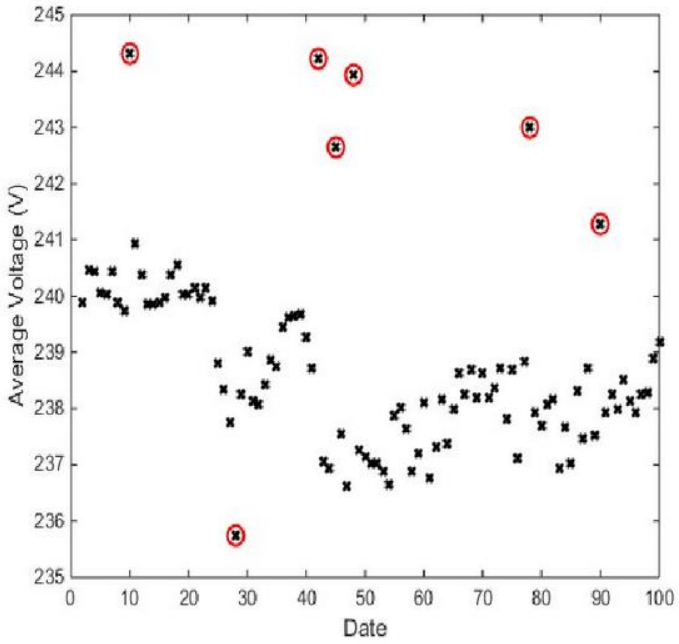


Figure 1.8 Outlier Analysis

Types of outlier:

- **Univariate data:** in the univariate data it consists one variable. this is very uncomplicated form of analysis since the data deal with only one amount that convert.it not relate with causes or relationship and the main purpose of the analysis is to relate the information and find pattern the belong it. Any how you can relate pattern found in

  Univariate data includes central tendency (mean, mode, median) and variance (range, variance, maximum, minimum, quartile, standard deviation). There are several ways to map one-dimensional data.

  For example, suppose that the length of six labour in a company is recorded, there is only one variable that is length and not relate with any cause or relationship. The description of pattern found in this type of information can be drawing conclusion using central tendency measures (mean, median and mode).

- **Multivariate data:** In this data involve three or more variable. It is defined under multivariate data it is similar to bivariate but contains more than one dependent variable. This type of analysis always performs with software. Multivariate analysis data can reduce the type I error.

  There is different way to performed multivariate analysis you can choose depend on the data you have and what yours goal are

  Types of analysis

  1. Multivariate analysis of variance
  2. Principal component analysis
  3. Factor analysis
  4. Canonical correlation analysis
  5. Redundancy analysis
  6. Corresponding analysis

  Multivariate analysis relates to any statistical technique it used to analysis complex data. Basically, you develop model that relate an actual product or process and optimize it using different method.

  There 4 Outlier detection methods:

1. **Numeric outlier:** This is very basic solution here non-parametric outlier detection technique used in a one-dimensional feature space. Here we calculate by mean of the IQR (inter quartile range).

2. **Z-score:** Z-score is a method for detecting parameter outliers in one-dimensional or low-dimensional feature space. This method takes into account the Gaussian distribution of the data. Migrants are far from the average because these are the data points at the bottom of the distribution.

3. **DBSCAN:** this is based on clustering technique it is same as numeric outlier it is also non-parametric. A density-based outlier detection method for all data that is defined as end points or noise points in the one-dimensional or multi-dimensional DBSCAN feature space.

   1-core point: data point has at least minimum point neighboring data point within a distance

   2- Border point: Border point is neighbors of a core point within the distance

   3- Isolation forest: isolation forest it used for huge dataset in a one multi-dimensional and it is non-parametric method. The isolation number is an major concept of these the isolation number splits needed to separate a data point. The number of splits is ascertained by following these steps

1.1.4 Regression

Regression analysis is a statistical method that establishes the relationship between dependent and independent variables. Regression analysis helps us how dependent values changes to Independent variable when another independent variable is stable. It predicts continuous value such as age, salary, price, temperature etc.

Let’s understand the example problem example suppose there is company B, who’s having different advertisement almost every year the data of company is the last 5 year and the corresponding sales

Advertisement	Sales
$90	$1000
$120	$1300
$150	$1800
$100	$1200
$130	$1380
$200	??

The company currently hopes to announce $ 200 in 2021 and wants to know the forecast. Therefore, if machine learning requires this type of prediction, regression analysis is required.

Regression helps us to find correlation between variable and active us to predict the outcome based on previous one or more predict variables.

It is used only prediction, forecasting, time series modeling, and determining the relationship between variable. Regression plots between variables are best suited for specific data points. Through this data machine learning make prediction about the data.

When the target variable and the explanatory variable have a linear or non-linear relationship, different types of regression analysis are used, and the target variable contain continuous values. It is a basic method to evaluate regression problem in ML using data modeling.

Different type of regression technique.

1. Linear regression
2. Logistic regression
3. Ridge regression
4. Lasso regression

- **Linear regression:** In linear regression is one of the popular techniques in machine learning this contain predictor variable and a dependent variable related linearly to each other. When your data contains multiple explanatory variables instead of linear regression, this is known as a multiple linear regression model.

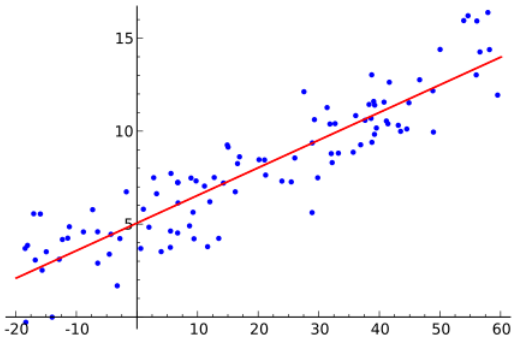


Figure1.9 straight regression

The predictor error represents the difference between the observed value and the predicted value. The value of machine learning gets selected such way that it gives the less predictor error.

- **Logistic regression:** LR is a type of regression analyst technique. Used when the discrete dependent variable is 0 or 1, true or false. This means that the variable can only have two values, and the sigmoid curve shows the relationship between the target variable and the independent variable.

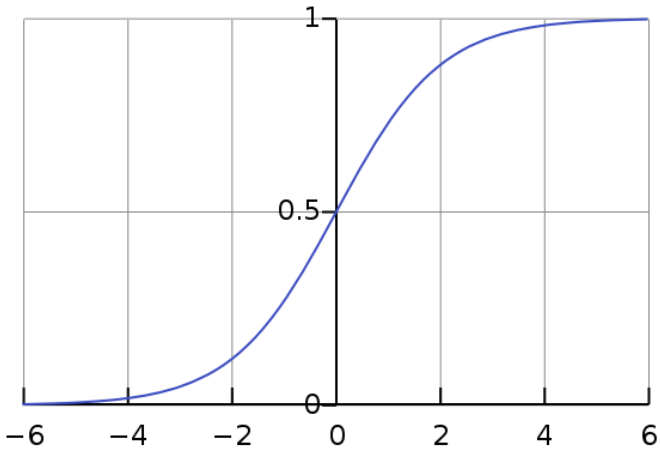


Figure 1.10 Logistic Analyst

If you choose logistic regression as your regression analysis method, your data should be large and the median occurrences of your target variables should be roughly the same.

- **Ridge regression:** This is another type of machine learning regression analysis used when there is a strong correlation between the independent variables. This is due to the multicollinearity of the data. The least squares estimate gives an unbiased

value. This is a powerful regression technique that makes the model less prone to overfitting.

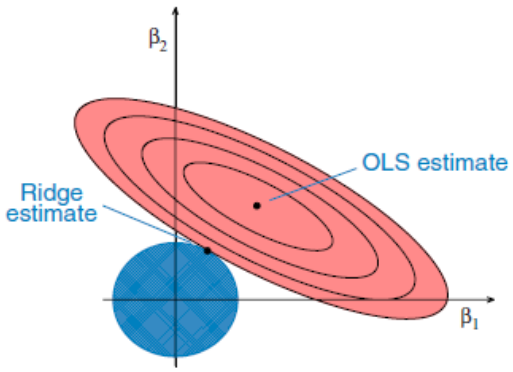


Figure 1.11 Ridge Regression

- **Lasso regression:** Lasso regression is a machine learning regression method that performs regularization. The value of this coefficient is close to zero, but is not displayed when the ridge regression.

Feature selection is used. In the case of recurring regression, only the functions required are used and the rest are set to zero

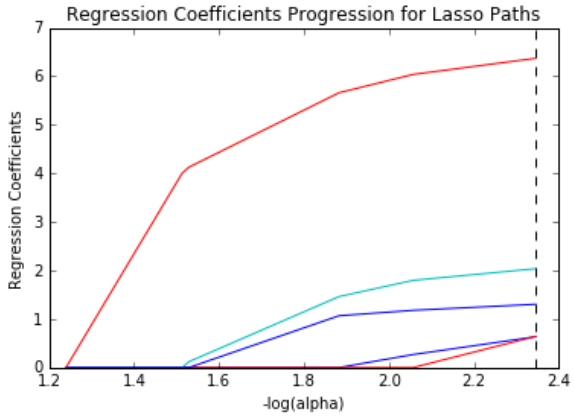


Figure 1.12 Lasso Regression

1.1.5 Sequential Pattern:

Data mining contain some information which stored in the database to recognize the data and/or take decision. There are some basic data mining taken are clustering, outlier, and pattern mining different type of unexpected pattern in database available such as

Frequent item sets, sub graph, sequential rules, and periodic pattern.

Sequence pattern mining is a data mining technique that specifically analyzes sequence data to find consistent patterns. Finding pattern sequences in many real-world applications is a purchase, as the data is encoded as character sequences in many areas such as bioinformatics, e-learning, shopping cart analysis texts, web traffic analysis examples, and more. It is run by an individual customer.

SID	Sequence
1	$\langle\{a,b\},\{c\},\{f,g\},\{g\},\{e\}\rangle$
2	$\langle\{a,d\},\{c\},\{b\},\{a,b,e,f\}\rangle$
3	$\langle\{a\},\{b\},\{f,g\},\{e\}$
4	$\langle\{b\},\{f,g\}\rangle$

The database contains 4 sequence every sequence represents the item purchase by customer at different items a database sequence contains order list of item sets. The first SID1 sequence is that after the customer purchases item c, items a and b are purchased together, then items f and g are purchased together, item g is purchased, and finally item e is purchased. bought. Please indicate what you have purchased.

to perform sequence pattern mining user should provide sequence database and provide relevant parameter known as minimum support threshold it specify minimum no of sequence in which pattern must appear to do consider frequent, and to show to the user. For example, if a user set the minimum support threshold to 2 sequence then the sequence pattern finds all the subsequence pattern less than 2 sequence of the input database sequence pattern mining also applied to time series e.g. stack data. For example, below figure showing a time series on left on the right a sequence is shown representing the same data. After applying a transformation. Different transformation may be transferring a time series to a sequence such as the popular SAX transformation. After completion of transformation any sequential pattern mining algorithm can be applied.

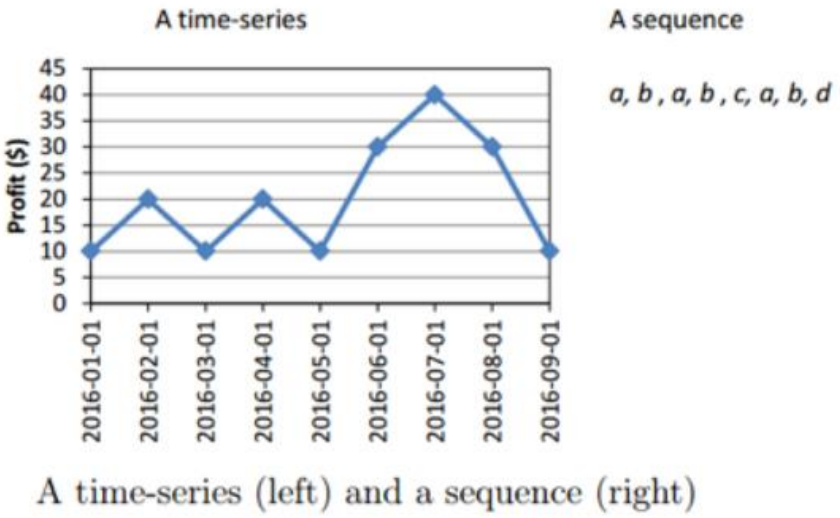


Figure 1.12 Sequence Pattern

Different type of algorithm is proposed in sequential pattern mining. Some of the basic algorithm prefix span, SPAM, Spade and GSP. However recently more efficient algorithm has been proposed such as CM-SPADE and CM-SPAM FClosm and FGensm (2017).

1.1.6 Decision Tree:

A Decision tree can used to build for predictive model which have application in machine learning, data mining, and statics called as decision tree learning.

In this decision tree node represent the data rather than decision. This type of model known as classification tree each branch node contain attribute that relate with a particular class label, which is in the leaf node.

Each information helps the model more accurately predict which of a finite set of value. After that this information can be used as input in a larger tree for making model.

Not every time but some time predicted variable will be a real number, such as price decision tree with continuous, infinite possible outcome are called regression tree.

Multitree also used in ensemble tree

- **Bagging**: building multiple trees by gathering Initial data aggregation, also known as packing, is a machine learning algorithm that uses statistical classification and regression.

- **Random forest classification**: It is used to increase the classification. It is a supervised learning algorithm for classification and regression.

Advantage of decision tree in machine learning:

- Works for either category or numerical data.
- It uses a white box model.
- Tree is verified and tested.
- Model problem with multiple outputs.

1.1.7 Clustering:

Clustering is the process that group similar group object. In the cluster data are divided into different group, in this step it partitioned the set of data into group with same data and after that group is assigned to the data.

It is used in many applications such as data analysis pattern recognition. It is used in biology, by deriving animal and plant taxonomies, identifying same DNA with same data. It is important topic of data mining .it used to divide several groups with similar data based on same rule which makes same group of data. It is method to study with logical and physical relation of data.

The result of cluster is not relevant the inherent relation and difference between data but provide further data analysis and some basic knowledge. There are many uses of data clustering analysis such as image processing, market research, binary data, categorical and internal-based and many more. Classification can also do on the basic of data analysis. Data clustering used in large data analysis and small data also many various types of data used with clustering algorithm of clustering. The data can be like binary data, categorical and internal-based data.

Application of cluster analysis:

- Used in various filed form image processing, data analysis, pattern recognize.
- Base on clustering classification of information also done.
- It is also used in detection application; face reorganization fraud in a credit card also detected, and pattern reorganization.

Requirement of clustering in data mining:

- **Interpretability:** the experiment result of cluster should be used further and easy to understand and express the possibility to translating one into the other.
- **High Dimensional:** in the data cluster it is also capable to deal with high number of coordinates needed specify along with data of small size.
- **Clustering scalability:** the data base is very big to handle the clustering should be changed in size or scale database, so it needs to be scale.

Clustering algorithm can be defined as follows:

- **Hierarchical method:**

Hierarchical clustering is the important technique of machine learning. It works like grouping data into various group data converted into tree of cluster. It first treated data pint as a separate cluster after that it repeat subsequent step.

- It will find the two closet cluster and combine them
- Combine the two closest clusters which are comparable to each other. We combine these steps until are single data cluster combined together.

In the hierarchical clustering, the main focus to build a cluster in nested form is known as Dendrogram basically it is showing graphical representation of hierarchy and in a reverse tree that describe the sequence order in which factor are merged bottom to view or break the cluster top to bottom.

Hierarchical cluster is also called HCA it is an unsupervised algorithm which create tree base on top to bottom or bottom to top. For example, all the file and folder are stored in hard disk in a hierarchical manner.

Basic two important about hierarchical clustering are:

- This technique has been implemented above using bottom up approach. It also follows top-down strategy.
- To merging two cluster based on closeness of these cluster.

Hierarchical clustering algorithms are divided into two type of algorithm:

- **Agglomerative clustering:** In agglomerative clustering algorithm work in a bottom-up manner. In this each data is treated as single nodes (leaf-node) at every process of the method, the two similar groups which are most common

are combined together into a new larger cluster. And this method iterator until all the data point combined/merged into one single big cluster (root). The combined outcome is a tree which can be plotted as a dendrogram. Steps that include in AHC (Agglomerative hierarchical clustering).

- Create single data as a cluster let understand there are p data point so number of clusters in p hierarchical in machine learning
- Take two data point and merge into single data point (cluster) then it should be p-1 cluster.
- Again, take the two nearest clusters and combine them and make it one cluster. Then it should be p-2 cluster.
- Repeat 3 process until it became single cluster
- After the entire cluster combined into big cluster, build the dendrogram divide cluster as per given problem.

- **Division hierarchical clustering:** this hierarchical clustering algorithm work in a top-down manner. Reverse process of AGNES it starts these processes from root node, and all the object in single cluster at every steps of iteration, the most different cluster is partition into two cluster. It performed until all the data in a single cluster.

  to find distance measure between two cluster as we know that the nearest distance between two cluster in hierarchical clustering, there are different method are used for measuring distance between two cluster. This method called linkage method there are some popular method are given below.

  - **Complete linkage:** this is very popular technique and this technique is farthest distance between two points of two different clusters. It is very tighter cluster than single-linkage. It produces small and easy cluster it is popular technique compares to single linkage cluster.

  - **Single linkage clustering:** this technique used for shortest distance between the closest parts of the cluster.

- **Average linkage:** it will consider all pairwise two dissimilar cluster and consider the average of these dissimilarities as the distance between the two clusters.

- **Centroid linkage clustering:** This method calculates the distance between the centers of gravity of the cluster.

- **Density-based method:** density-based clustering is a technique used in data mining for retrieving the data pattern from the data set. The reason behind density-based clustering is to detect cluster of non-spherical or arbitrary shapes.it is unsupervised learning method that find distinct group in the information this is used for clustering of very large database. Some of the common technique is DBSCAB, OPTICS, VDBSCAN, DVBSCAN, DBCLASD, and ST-DBSCAN AND DENCLUE.

1. **DBSCAN:** density based spatial clustering of application with noise (DBSCAN) it is very oldest clustering method. It will find high density region in spatial data base with noise and build cluster out of them.

   Advantage of DBSCAN:

   - It can detect arbitrary cluster
   - There should not any pair knowledge required for number of clusters
   - In this, object not belong to any other else
   - They don’t know how to ordering of the point in the database

   Disadvantage of DBSCAN:

   - Actually, deciding the first value of the parameter E and mini-pts. is difficult
   - For p data object time complexity n2 without spatial indexing
   - Noise point will not detect if density is varying.

2. **VDBSCAN:** varied density based spatial clustering of application with noise different type of density detected in the VDBSCAN algorithm. And it will select various value of input parameter Eps for different densities k value also

generated automatically based on the essential quality of dataset.

- It partitions the k-dist. plot for each project and calculate and store.
- It scans the dataset and cluster various density using corresponding epi.
- Actual cluster and different density will be display

Advantage:

- Different type of density and cluster can be detected.
- It can select various input parameter epsi for varied density.

Disadvantage:

- Time complexity is very high
- For p data object time complexity n log n when we using spatial indexing is used

3. **DVBSCAN:** Density-based algorithm for identifying different density groups in large spatial data sets. It will handle local density variation with in the cluster. Minimum object as an input parameter used. The algorithm can find the average density and density variation of the underlying object. Below the threshold also corresponds to the cluster similarity index.

Advantage of DVBSCAN:

- It takes care about density variation with in the cluster
- It displays DBSCAN specially in case of local density

Disadvantage of DVBSCAN:

- Time complexity is high for a set of p object

4. **DBCLASD:** Distribution-based clustering algorithms for finding large spatial databases are used to detect arbitrary shapes, it does not need any input data for further DBCLASD used for large spatial database.

Algorithm defines as:

- a point added to a cluster without consider whole database this is incremental algorithm
- The group of voting groups was formed using environmental questions

related to the Spatial Access Scheme.

- Use the chi-square test to test the hypothesis that the amount of distance from the closest neighbor of the group matches the expected distance distribution.

Advantage of DBCLASD:

- It not considers any input value
- It is used for continuous distributed point

- **ST-DBSCAN:** spatial temporal density-based cluster in this algorithm temporal data such as geographical involving system, medical image, and weather forecasting

  Advantage of ST-DBSCAN:

  - in this data can be performed based on spatial and temporal attribute and non-spatial
  - in a varied density it also detects noise point by assigning density factor so each cluster

- **OPTICS:** A cluster analysis technique called optics has been proposed to identify cluster structure algorithms that reduce the complexity of a set of global parameters in cluster analysis. Optic does not explicitly group records, but instead creates a grouped order.

  Advantage of optics:

  1) It is not mandatory to provide density threshold

  Disadvantage of optics:

  1) The time complexity is very high for a database of p object.

- **DENCLUE:** density-based clustering algorithm this algorithm is based on density distribution function. It is a improved of DBSCAN and optics in term of density estimation.

  - Cluster is formed from these local density maxima value

- If local density is small then object of the cluster are discards as noise under these objects are added to a cluster through density attractor using a step wise hill-climbing procedure

Advantage:

- we can find the arbitrary shape of cluster
- method is invariant against noise
- sensitivity of density is removed

Algorithm	Varied Density	Primary Input Requirement	Time Complexity	Cluster Type	Type of Data
DBSCAN	No	Cluster Radius, Minimum number of Objects	$O(n^2)$; O(nLogn) for spatial indexed data	Arbitrary Shaped	Spatial Data with Noise
VDBSCAN	Yes	Automatically Generated	$O(n^2)$	Arbitrary Shaped	Spatial Data with Varied Density
DVBSCAN	Yes	Two Input Parameters to be given by User	$O(n^2)$	Arbitrary Shaped	Spatial Data with Varied Density
DBCLASD	Yes	Automatically Generated	$O(n^2)$	Arbitrary Shaped	Spatial Data with Uniformly Distributed Points
ST-DBSCAN	No	Three Input Parameters to be given by User	$O(n^2)$	Arbitrary Shaped	Spatio-Temporal
OPTICS	Yes	Density Threshold	$O(n^2)$; O(nLogn) for spatial indexed data	Arbitrary Shaped	Spatial Data with Varied Density
DENCLUE	Yes	Radius	$O(n^2)$	Arbitrary Shaped	Spatial Data with Varied Density

Table 1.1 comparison of various density-based clustering algorithm

- **Grid-based method:** Grid-based clustering techniques use multi-resolution grid data structures. Using this method, the data space is formed into a limited number of cells that form a grid structure. All grouping operations performed in these grids are quick and independent of data object statistics such as STING (statistical grid), shaft grouping, and CLIQUE (grouping overview).

  The main advantage of grid clustering is that it significantly reduces computational complexity. This does not affect the data points, but it does affect the range of values that surround the data points. With basic fine steps required.

  - Creating the Grid structure
  - Calculating the cell density for each cell

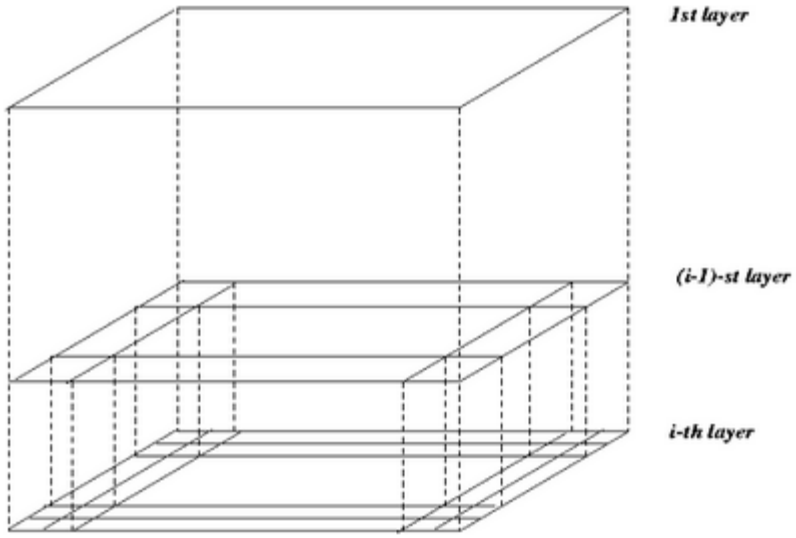
- Maintaining the order of cell according to their density
- Identify the cluster centers
- Traversal of neighbor cells

STING:

- A statical information grid approach
- Spatial area is partition into rectangular cell
- various levels of cell at different level of decision
- High level cell is partitioned into several lower level cell
- Statically attribute is stored in cell
- Mean, maximum, minimum
- delete the irrelevant from further consideration

Sting Algorithm:

- decide a layer to begin with
- For each cell in this layer, calculate the confidence interval (or range of estimates) for the probability that this cell is associated with a query.
- Marks cells as related or unrelated from a previously calculated range.
- This layer is the bottom layer. Go to step 6. Otherwise, go to step 5.
- Move down one level in the hierarchy, and then go to step 2 of the cells that make up the corresponding top-level cell.



- If the required specifications are met, proceed to step 8. Otherwise, go to step 7.
- Take this data, classify it into related cells, and perform further processing and go to step 9
- Find the appropriate cell range. Return the area that meets the query requirements. Proceed to step 9. Find the corresponding area of the cell. Returns the area that meets the query requirements. Skip to step 9
- stop

Advantage:

- query- independent approach
- computational complexity is order off n
- It's easy to parallelize query processing algorithms with this structure.
- incremental update
- the time clustering of generating cluster I order n where n is the total numbers
- This structure facilitates the parallelization of query processing algorithms.
- the method efficiency is a major advantage
- High speed processing because the statistical information stored in each cell

is a summary of the data.

Disadvantage:

- a multi-resolution clustering approach that applies wavelet transport to feature space
- Wave conversion is a unique processing technology that divides a signal into different frequency sub-bands
- it uses feature of bid grid-based and density-based clustering
- Wave groups that can handle up to 20-dimensional data

Input parameter: number of the grid unit is the wavelet of each dimension and is the number of times the wavelet transform is applied.

How to apply wavelet transform to find cluster:

- Aggregate data by imposing a multidimensional grid structure on the data space.

  These multidimensional spatial data objects are represented in n dimensions.

  Characteristic space
- Apply a wave transformation to a feature space to get a dense area in a space feature.
- When wavelength conversion is repeatedly applied groups of different scales are created

Why is wavelet transformation useful for clustering?

- unsupervised clustering
- Use a spatial filter to select an area with a set of points, but at the same time capture weak information around that edge
- effective elimination of outlier
- multi-resolutions
- cost efficiency

In the grid –based clustering it is different from the standard clustering algorithm it is not anxious with date but value space surrounded the data points.

CLIQUE: You can think of it as density and grid based

- Divides the m-dimensional data space into non-overlapping rectangular blocks.
- If the proportion of the total number of data points contained in a unit exceeds the parameters of the input model, the unit is said to be dense.
- Groups are the largest collection of dense units connected in a subspace.

- **Model-based method:** model based-method is a popular technique used to represent different constraint and geometric properties of the covariance matrices. Model based cluster use Optimize adaptation between data and model. In the model-based approach the data parameter considers as mixture of probability distribution for these each cluster consider as different component. The concept of grouping to maximize expectations, various methods using neural network methods, it finds the number of clusters automatically based on standard statistics, it therefore yields robust clustering method.

  Model based clustering also known as mixture model. Model based clustering is used to design for unknown distribution mixture of simpler distribution also called basic distribution many advantages of model-based clustering over method.

  K means clustering is a part of model-based clustering, where all the distribution assumed to be Gaussians with equal variance.

  - We repeat every possibility distribution, and uncertainly generated the input
  - We iterate the process until all the input distribution converge
  - For each gene, now calculate the probability that the gene pattern was generated for every of the distribution
  - Calculate the parameter of each distribution and maximize the expression data given the probability that each gene was generated from the distribution
  - Give each gene to distribution which generate the gene's

  Each cluster is representing mathematically by a parametric probability distribution

  - Component distribution

- Data is a combination of these distribution
- Mixture density model
- To estimate parameter of probability distribution

Expectation maximization:

- Iterative refinement algorithm: used to find parameter estimates
- Extension of k mean assign an object is grouped by weight, representing the probability of belonging
- Initial estimate of parameter
- Iteratively reassign scores
- Simple easy to implement
- Complexity depend on feature, object and iteration

The Three examples of clustering based on the extended k-means partition algorithm are:

1) Expectation maximization

2) Conceptual clustering

3) Neural network approach to clustering

There are two groups, each of which follows a normal or Gaussian distribution and has its own mean and standard deviation

Conceptual clustering- COBWEB

- Concept grouping is a form of grouping in machine learning
- Generates a classification scheme for objects by specifying a set of unlabeled objects
- It essentially identifies groups of similar objects, then finds a description of the characteristics of each group to represent concepts or classes
- Hence conceptual clustering is a two-step process
- Clustering is performed first, followed by characterization

- Most concept grouping methods use probability-based statistical methods
- COBWEB is a simple and general way to group incremental concepts
- Input characteristics are described by categorical attribute / value pairs
- COBWEB creates a hierarchical group as a classification tree
- COBWEB uses a metric called Category Utility to determine the stress of the tree
- Category utility (cu) id defined as

Limitation of COBWEB:

- The assumption of probability distribution as separate attribute is It remains true that they are statistically independent of each other
- It is quite expensive to update and store the cluster
- The time and space complexity are more, it depends not only on the number of attributes, but also on the value of each attribute
- Classification trees are not very well balanced

CLASSIT:

- CLASSIT is an extension of COBWEB for the continuous incremental grouping of data
- Use the modified classification helper, which is an integer number of continuous attributes, not the sum of discrete attributes like COBWEB
- It is not suitable for clustering large dataset data
- However, this method is not suitable for large data sets

- **Constrained-based method:** while mining, Users indicate intuition or expectations as limitations in defining the search space. This strategy is called constraint extraction. Constrained base modeling is a scientific proven mathematically approach their output of All solutions are limited by the minimum and maximum ranges, and all variable value types with similar constraints must have a solution value in the constraint.

Constraint modeling is used primarily in optimization techniques. For example, use mixed and linear integers to minimize the objective function.

Constraint-based modeling is not hypothetical. Analysts using constraint-based models should not be able to predict the behavior of the entire model ahead of time. The solution that makes this possible is not true constraint-based modeling. In real-world stress-based modeling, surrogate values for stress solutions are also calculated. Not only does the administrator notify that the machine's power is used 40 hours a week, but there is no additional production. The best constraint-based modeling methods provide this information, but also identify opportunities that affect earnings.

- **Partitioning method:** partitioning is a data mining technique in which it divided the object in multiple groups. The data in each group consist similar characteristic where the object of various cluster have dissimilar characteristic in term of dataset attribute. We can identify similarity or dissimilarity with the help of distance measure to improve the partitioning technique from one object to another it uses iterative relocation technique. The main focus of partitioning method is to fix the number of cluster and reduces the square root error. In the partitioned algorithm contain a single cluster instead of cluster structure such as a dendrogram partitioning clustering algorithm split the data into k different partition. In these papers I am using k-mean algorithm and fuzzy c mean algorithm and different type of distance matrix to represent and compare the result. If each part behaves like a cluster and is divided by functions, these functions reduce the least squares calculation.

$$\boldsymbol{E} = \in\in (||\boldsymbol{p} - \boldsymbol{m_i}||)^{\mathbf{2}} \qquad \mathbf{1}$$

Where p is the point in a cluster and mi is the mean of the cluster. the cluster contain two properties.

  - Each cluster contain one object at least
  - Every object belongs to one group it must

Drawback:

  - Points near the center of another group will not work well due to overlapping data points

There are different types of segmentation methods that use k-resource, medoid, and

PAM (segmentation).

Advantage:

- Relatively scalable and simple
- This is good for database that compact spherical cluster that total separated.

Disadvantage:

- Bad cluster descriptors
- Very high noise and outlier

K-Medoids: k- medoids is fall under the categories of unsupervised machine learning. It is totally different from other clustering algorithm in the term of how to select the cluster centers. K-medoids is different from the k median because it chooses the medians instead of center.

Working steps of k-medoids approach:

- It selects random p point from the input data. The value we choose p can be assessed by method such as Silhouette method.
- Every data point assigned to the group of clusters nearest medoids belong.
- Every data point of cluster I, computed the distance from other and added
- Step 2 and 3 are repeated until convergence is reached i.e.

Advantage of k-medoids:

- Very easy to understand
- It is very fast in term of fixed number of steps
- PAM is less sensitive to outlier than another partitioning algorithm
- K-medoids is better than in term of robust to outlier and noise
- In k-medoids is most central element of the cluster

Disadvantage of k-medoids:

- In the k-medoids it’s not suitable for non-spherical clustering group of objects
- It minimized the distance between medoids and non-medoids object

Partitioning Around Medoids: PAM is known as partitioning around medoids in this algorithm is used to find sequence of object known as medoids that are present in center of cluster. The main focus of this algorithm to reduce/minimize the dissimilarity object to the closest selected object. PAM allows clustering with respect to any specified distance matric based on experience PAM does have problem of small clustering.

These algorithms consist of two steps:

- BUILD, gathering number of p object chooses for a first set s.
- SWAP, we exchanging chooses object with UN-chooses object to improve the quality of cluster.

CHAPTER-2
LITERATURE REVIEW

This module discusses about the work being conducted by various researchers in the field of k-means algorithm and FCM (fuzzy c-means algorithm). In these we comparing different types of distance measure; researchers have done various researches on k-means by using different techniques. It also discusses about the existing work of some researches that will help us to know about the k means and FCM algorithm and their different measure in depth.

2.1 Background Work

Clustering is “the process that groups objects into subclasses so that those subclasses are entirely related or have some similarities and the different clusters are altogether dissimilar from each other”. It is used in many applications such as data analysis pattern recognition. It is used in biology, by deriving animal and plant taxonomies, identifying same DNA with same data. It is important topic of data mining .it used to divide several groups with similar data based on same rule which makes same group of data. It is method to study with logical and physical relation of data.

Clustering algorithms can be divided into four groups: segment clustering, hierarchical algorithms, density-based algorithms, and lattice algorithms.

Partitioning clustering, the data objects are directly divided into clusters of a pre-defined number. "The checking for all possible clusters is computationally impractical; certain greedy heuristics are used in the form of iterative optimization of a cluster. The partitioning clustering consists of several approaches such as K-means Clustering, K-medoid Clustering, Relocation Algorithms, and Probabilistic Clustering." We introduce the k-mean Algorithm, and FCM (fuzzy c-mean Algorithm) k-mean depends on data to detect anomaly detection's-mean Continuous search a user 'defines' many cluster middles and relate data to the nearest cluster middle such that the intra-cluster variance is minimized.

Hierarchical Grouping "creates a hierarchical breakdown of a particular set of data objects. Create a group hierarchy called a dendrogram. Get clusters of non-overlapping clusters by cutting the tree to the required level. Use hierarchical grouping to find different grades of

Data.

Density-based algorithms "may get a set of unexpected random shapes. The objective density function is used to group objects. Through these grouping methods, the number of adjacent elements. The group will continue to grow until it is restricted.

Grid-based clustering "quantifies the model space of a limited number of cells by creating a grid structure. After calculating the grouping based on density, sort the cells by density. The group location is located and all neighboring cells intersect. In the article described here, we used different types of distance metrics and compared the k-means and FCM algorithms based on the distance metrics.

2.2 Distance measurement in k-means algorithm and FCM:

In the k-means algorithm we are finding the distance between every point of the data to every centroid initialized. Whatever the value found we allocate to the centroid with minimum distance.

Distance matric is a very popular method in the clustering algorithm. There are different techniques we use to calculate the distance between two points. So, our focus to deal with proper techniques and apply to get the minimum distance to select such technique we have to take care some property of the data and dimensional of dataset. These experiments consider city block, correlation, cosine, Euclidean. This technique used to calculate distance in the k-means algorithm with different-different parameter used.

2.2.1. City-blocks distance

In the city-blocks measure it consider two-point a and b with k dimensional is defined as

$$\sum_{j=1}^{k} |a_j - b_j|$$

City blocks is also known as Manhattan let consider two-point x-y plane, we consider shortest distance between the two points is along the hypotenuse which is Euclidean distance. The distance between blocks is used to calculate the membership function of the fuzzy set.

Instead, the city block distance is calculated as x distance plus y distance. This is similar to how you travel in a city (for example, Manhattan) where you need to go through the building instead of going through it.

You can move only one dimensional of the space at a time in city-block by analogy. It does not affect distance in the space note also that many equal length paths exist between two point in city- block space.

Euclidean distance and city-block distance are special case (different value of k) of the metric in two dimensions.

$$\text{Distance}= \sqrt[k]{x^k + y^k}$$

Here x and y are distance in each of two dimensions generalizing this to P dimension and using the form of the equation of ED

2.2.2. Euclidean distance:

In Euclidean distance there are two-point a and b with k dimension is calculated

$$\sqrt{\sum_{j=1}^{k} \left(a_j - b_j\right)^2}$$

By default, k-resource uses Euclidean rules, and the distance is usually calculated based on raw data rather than standardized data. Distances are affected by the proportional difference between the calculated distance dimensions. The advantage of this method is that the distance between the two features is not affected when new features that may be outlets are added to the analysis.

For example, when one of the dimensions given in millimeters and then you converted into centimeters, the resulting Euclidean can be affected and consequently Grouping results may vary. The distance between two astigmatism objects is usually defined as the minimum distance between a pair of points on the two objects. A known formula for calculating the distance between different types of objects, such as the distance from a point to a straight line. The concept of distance is generalized to abstract metric spaces and non-Euclidean distances are studied. Many statistical and optimized applications use Euclidean squares instead of the area itself.

One dimension -

- The distance between points in the real line is the absolute value of the difference in the coordinate values. Therefore, if p and q are two points on the real line

$$d(p,q) = \sqrt{(p-q)^2}$$

2.2.3. Cosine distance:

Cosine length is the sharp length of Euclidean that normalizes data to units of length. The cosine distance between two points is not the cosine of the angle between the points (it is considered a carrier). m (1 row in n columns) row vector x1, x2 ... ... In the data matrix X processed as xm with m rows and n columns, the chord distance between the vectors xs and xt is defined as follows:

$$d_{st} = 1 - \frac{x_s x_t}{\sqrt{(x_s x_s')(x_t x_t')}}$$

2.2.4. Correlation Distance

Distance correlation is a measure of the relationship between random vectors. Consider the x matrix of the given m × n data as the row vector m (1 × n) x1, x2 ... xm. The correlation distance between the vectors xs and xt is defined as

$$d_{st} = 1 - \frac{(x_s - x_s')(x_t - x_t')'}{\sqrt{(x_s - x_s')(x_s - x_s')'}\sqrt{(x_t - x_t')(x_t - x_t')'}}$$

2.3 FCM algorithm:

The fuzzy c mean clustering algorithm automatically finds various types of outliers by tracking multiple systems in real-time. Primary fig 1 below, In FCM, it is used to combine the concept of clustering. It dependent on normal and abnormal behaviour and is divided into as many groups as possible but does not overlap.

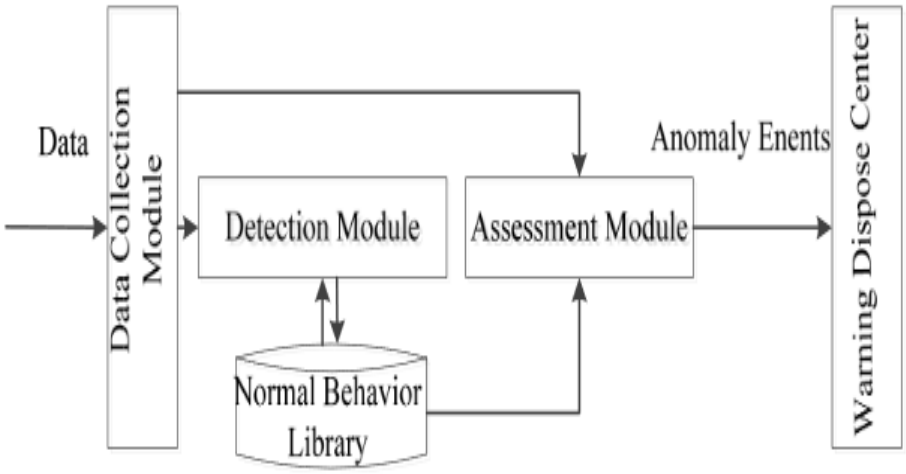


Fig. 1 Basic model of anomaly detection

FCM locate anomalies detection method are widely used in many technologies. In medical applications, the literature applies FCM-locate anomalies detection to segment MRI images of the brain using the FCM algorithm to correct intensity mismatch and improve the tissue segmentation algorithm's accuracy. The fuzzy c-mean algorithm locates the brain's segmentation anomalies, tumours using magnetic resonance imaging. An artificial bee colony algorithm is then introduced to reduce noise and detect brain tumours and effectively compensate for a single algorithm's limitation, thereby providing essential and accurate energy improvements for critical energy consumers.

2.3.1. Theory of FCM:

Fuzzy grouping in multiple technologies classifies goals and determines faint similarities based on characteristics, relationships, and goal similarities. A preliminary clustering survey can be broken into three ways:

- The number of classifications is unknown. This number that hey can be dynamically grouped just as a different obligation.
- Displays the number of classifications. The goal is to detect an excellent way to group the value. This procedure is based on a clustering function called Fuzzy C-Means (fuzzy C-mean) algorithm or ISODATA fuzzy clustering
- If there are significant violations, grouped by the fuzzy similarity matrix. This procedure is called fuzzy noise grouping

Fuzzy C- Mean Clustering is a division-based algorithm; this is a better resource-based algorithm. C-management algorithms are strict on data sharing, while FCM is flexible and

ambiguous. The FCM uses the membership for each statistic example according to the amount in the specific grouping. It splits the statistical example.

$$\boldsymbol{x = \{x_i / x_i \in R(i = 1, 2 \ldots n)\}}$$

FCM calculates the cluster centers of each category to minimize the cost of the mismatch classification matrix.

$$\boldsymbol{U = \{U_{ij} \mid i = 1, 2 \ldots n; j = 1, 2 \ldots k\}},$$

$$\boldsymbol{\sum_{j=1}^{k} U_{ij} = 1, \forall i = 1, 2 \ldots n} \quad (\mathbf{1})$$

We use fuzzy c mean for partitioning so that each can be determined according to membership between 0 and 1. The element of the matrix U value between 0 and 1 as follows:

$$\boldsymbol{J_{ij}(U, C) = \sum_{i=1}^{n} \sum_{j=1}^{k} U_{ij}^{m} d_{ij}^{2}(x_i, c_j)} \quad (\mathbf{2})$$

J can be looked at as the quadratic sum of the distance between each data instance and

$$\boldsymbol{C = \left\{ \frac{C_i}{C_j} \in I, j = 1, 2, \ldots . k \right\} \; and \; C_i \in I}$$

The cluster center in (2)

$$\boldsymbol{C = [C_1, C_2, \ldots . . C_k]}$$

$$\boldsymbol{U_{ij} = 1 / \sum_{i=1}^{k} \left(\frac{d_{ij}}{d_{i1}} \right)^{\frac{2}{m-1}}, \forall i} \; (\mathbf{3})$$

$$\boldsymbol{C_{ij} = \frac{\left[\sum_{i=1}^{m} U_{ij}{}^{m} X_j \right]}{\left[\sum_{i=1}^{m} U_{ij} \right]}, \forall j} \quad (\mathbf{4})$$

The 'variable m in the above formula' is scaling to power the categories' blur length. The more prominent m is, the extra blurred when m = 1, the FCM algorithm is simplified to the C-Means Clustering (FCM) algorithm. FCM clustering requires multiple iterations of the value function to get the minimum value.

2.3.2. FCM using anomaly detection:

From the above conclusion, we can see the FCM locate interference detection algorithm needed two variables: the no of clusters C and the variable meter. The number of clusters can use C as the clustering seed, and C is less than the total number of cluster samples. You can do the following to optimize detection:

Steps1: actuate the membership matrix U with a random number between 0 and 1, and satisfy the formula

$$\sum_{i=1}^{n} U_{ij} = 1, \forall j = 1, 2 \ldots . n.$$

Steps2: use $C_i = \left(\sum_{i=1}^{n} u_{ij}{}^{m} x_j\right) / \sum_{j=1}^{n} u_{ij}{}^{m}$ **to calculate the cluster center** $C_i, i = 1, 2 \ldots . k$.

Steps3: use $U_{ij} = 1 / \sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-j)}$ **To calculate the new membership matrix U.**

Steps4: calculate the value function according to

$J_m(U, C) = \sum_{i=1}^{N} \sum_{j=1}^{k} u_{ij}{}^{m} d_{ij}{}^{2}\left(X_i, C_j\right)$**. if it is smaller**

If the value exceeds a specific limit or is less than the function change, the results' clustering and output will stop. Or else, restore to step 2 to continue the production iteration of the algorithm. An N×K fuzzy separator matrix indicates the community of each sample related to each group.

Each sample can be identified using a matrix, including categories that meet the principle of full membership. The algorithm shows that clustering is better suited for normally distributed data and is more sensitive to an outlier.

2.3.3. Problem with FCM:

Cluster evaluation has become a critical data mining approach. Many cluster technologies are widely used, but all technologies have limitations and scalability issues. For fuzzy clustering algorithms, the FCM algorithm is most often used, but there are many problems.

The user has to pre-set the no of clusters, etc., and to choose the correct number of Clusters is an additional requirement for clusters, and the number of groups is challenging to determine.

The Fuzzy c means algorithm is susceptible to initialization because for local optimization methods, it is essential to use the mountaineering method to find the best iterative solution, and it's easy to reach a local minimum to find the best global solution.

2.3.4. Optimization of FCM algorithm:

FCM is an incursion detection method based on FCM grouping and is usually combined with other intrusion detection methods.

After all, there are many hybrid methods, such as combining black C-methods with a sound immune system or using intermediate information entropy, connecting FCM with support vector machines, fuzzy genetic algorithms, etc.

This research proposes how to increase the number of clusters in recent years and the many solutions.

2.3.5. Establishment of the number of clusters:

While many studies have focused on determining the number of Fuzzy C-Mean clustering algorithms and the first selection of clustering centers, related studies only assess the number of clusters or only select initial clustering centers.

To search the number of clusters in the FCM algorithm, a technique founded on average breast entropy uses the density function to get the original cluster centers.

The distribution of clusters is more reasonable, the data's entropy is smaller, and the information about the clusters' properties.

The clustering algorithm improves information disorder location, and the average data disorder is used as a standard to decide the number of clusters. The concept of the mean entropy of such information has already been mentioned.

$$\boldsymbol{H(K) = -\sum_{i=1}^{c}\sum_{j=1}^{n}\left\{\left[u_{ij} \times \log 2\left(U_{ij}\right) + \left(1 - U_{ij}\right) \times \log 2\left(1 - U_{ij}\right)\right]/N\right\}} \quad (5)$$

1st to define the scope of the no of the cluster [Cmin, Cmax] in (5) Uij indicate the extent of the sample j belongs to cluster j, Uij belongs to [0, 1] for all i,j. when K increases from C min, C max, it can be built as Cmax - Cmin+1 of Hk(x).

According to the rules, as long as the cluster owns the data, the information value will be the minimum value found using cluster number K=2, and the minimum value found using cluster number C will be Hk (x). An ambiguous C means that the algorithm (FCM) must determine the cluster number in advance. Each input sampling penalty is different because the Fuzzy membership function uses the support vector machine algorithm. The membership matrix obtained by injecting the SVM into the Fuzzy algorithm is used as the Fuzzy community's function. You can get a long-range hyperplane at a reasonable price.

If DSV <= min (ds1, ds2), the original dataset is Unable to classify, the assumption is incorrect. Otherwise, it will be installed, and the real data set can be split into at least two groups.

2.3.6. Acquire the global optimal solution:

This clustering technique based on genetic algorithms is suggest to aim that the sensitive initialization based on FCM quickly leads to the minimum value, and this method can globally convert to the maximum value with high probability.

To solve this problem, we used the 'clonal selection algorithm' (CSA) to improve the purpose of the automatic Fuzzy c-mean clustering algorithm.

The clone selection algorithm uses an antibody cloning mechanism to create a clone operative that combines advanced search, world search, potential search, and local search. Since the cloning operator is a CSA clustering strategy, you get the best overall solution with consistency and randomness.

They will soon unite. Simultaneously, the algorithm calculates the demand for a defensive response to overcome the premature growth trend. this article proposes an improved FCM algorithm for full recognition.

2.3.7. The enhanced intrusion detection algorithm located on (FCM):

People are strongly encouraged to conduct on-going research on how to enhance the finding efficiency of IDS. We group not only numeric data but also character data. If we are studying the mixed element of sample data, the distance is used for numeric data, therefore to solve the issue, a new logic is used, which we can describe the Distance between Xi, Xj, and k as fallow

$$d(x_{ik}, x_{jk}) = \begin{cases} 1, x_{ik} \neq x_{jk} \\ 0, x_{ik} = x_{jk} \end{cases} \quad (6)$$

Object statistics and have a p-value for Q and the specific property, and I think the distances

$$d(x_i, x_j) = d_n(x_i, x_j) + d_e(x_i, x_j) \quad (7)$$

In (7) dn (Xi, Xj), I not equal to j is the distance of the numerical attributes, and dc (Xi, Xj) Characters are separated. We can obtain the objective function from the mixed attribute data set. It can be changed to (2), namely:

$$j_m(U, C) = \sum_{i=1}^{N} \left\{ \sum_{j=1}^{k} {u_{ij}}^m \sum_{i=1}^{p} (x_{ik} - x_{jk})^2 + \lambda \sum_{j-1}^{k} {u_{ij}}^m \sum_{i-p-1}^{p-q} d^e(x_{ik}, x_{jk}) \right\}$$

In (8), force is used to stabilize a hybrid attribute database's property, and the relationship decides its value between the two types of properties. M> 1 vague container. Used to control the length of the American escape. Received

$${c_i}^n = \sum_{j=1}^{k} {u_{ij}}^m \sum_{i=1}^{p} (x_{ik}, x_{jt})^2 \quad (9)$$

$${c_i}^c = \lambda \sum_{j=1}^{k} {u_{ij}}^m \sum_{l=p+1}^{p+q} d_e(x_{ik}, x_{jk}) \quad (10)$$

As $C_i^n \; and \; C_i^c$ and nonnegative, we can minimize $C_i^n \; and \; C_i^c \;\; to \; make \; the \; j_m(U, C)$ minimize, at $u_{ij} = \left\{ \sum_{i=1}^{k} \left[\frac{d(x_i, x_j)}{d(x_i, x_j)} \right]^{\frac{2}{m-1}} \right\}^{-1}, \forall i \quad (11)$

Here we repeat the operation with 9 and 10 and 11, since M> 1, the algorithm is not compatible. Better FCM-based interference spotting algorithm, which is briefly described below

Step1: The member matrix U begins with an odd number from 0 to 1 and fulfils (1).

Step2: Use (9) or (10) to calculate the group center for various data attributes.

$$C_i^k \; i = 1,2 \ldots . k.$$

Step 3: we use 11 to compute the new community matrix U.

Step 4: compute the price functions just as to 8. If it is below a particular limit value or below the last modification in the values function, the grouping result is stopped and ended. Otherwise, return to step 2 and repeat the process.

When exploring sampled data using this method, you must consider both character and numeric data. Not only can it help you analyse data more fully to group it and reduce false positives and alarms, but it can also be used in conjunction with the FCM optimization algorithm. This can further improve detection efficiency

2.4 K-means algorithm:

K means algorithm cluster the different group of data into number of clusters in term of distance function. In k-mean algorithm different type of distance measure used but by default Euclidean distance used in k mean and in this small distance behave strong similarly long distance implies low similarity. Clustering is “the process that groups objects into subclasses so that those subclasses are entirely related or have some similarities and

the different clusters are altogether dissimilar from each other". Clustering algorithms can be divided into four different groups: segmentation, hierarchy, density-based, and grid-based clustering.

Here Euclidean distance calculation between two data point (x&y) with an object n dimensional space.

$$distance(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

The k mean algorithm is encapsulate as follows

- Randomly assigned k cluster
- Which having the nearest centroid assigned each object to the group. And after then Euclidean distance measure the distance between each object and each cluster centroid.
- Then calculate the centroid using formula $m_j = \frac{1}{n_{j \forall data_{p \in c_j}}} \sum data_p$,m_j represent centroid vector of the cluster j, nj, is the data vector cluster in cluster j, cj, is subset of the data vector from cluster j.
- Repeat step 2 until it's not reached maximum number of iterations. Until centroid do not change any more in the predefined number of iterations.

2.4.1. Partitioning clustering:

Partitioning clustering, the data objects are directly divided into clusters of a pre-defined number. "The checking for all possible clusters is computationally impractical; certain greedy heuristics are used in the form of iterative optimization of a cluster. The partitioning clustering consists of several approaches such as K-means Clustering, K-medoid Clustering, Relocation Algorithms, and Probabilistic Clustering. "The main focus of partitioning method is to fix the number of cluster and reduces the square root error. In the partitioned algorithm contain a single cluster instead of cluster structure such as a dendrogram partitioning clustering algorithm split the data into k different partition. In these papers I am using k-mean

algorithm and fuzzy c mean algorithm and different type of distance matrix to represent and compare the result. Where each partition behaves as cluster and partitioned is based on some function and these functions reduce the square error criterion which is computed as.

2.4.2. Hierarchical clustering:

hierarchical clustering "creates a hierarchical decomposition of the given set of data objects. It builds a cluster hierarchy, known as a dendrogram. Disjoint groups of clusters are obtained by cutting the tree at the desired level. Hierarchical clustering is used to find data on different levels of dissimilarity." Hierarchical clustering is the important technique of machine learning. It works like grouping data into various group data converted into tree of cluster. It first treated data pint as a separate cluster after that it repeat subsequent step.

- It will find the two closet cluster and combine them
- Combine the two closest clusters which are comparable to each other. We combine these steps until are single data cluster combined together.

In the hierarchical clustering, the main focus to build a cluster in nested form is known as Dendrogram basically it is showing graphical representation of Describes the order in which items are grouped to display or subdivide a group from top to bottom in a hierarchical reverse structure.

Hierarchical cluster is also called HCA it is an unsupervised algorithm which create tree base on top to bottom or bottom to top. For example, all the file and folder are stored in hard disk in a hierarchical manner.

Basic two important about hierarchical clustering are:

- This technique has been implemented above using bottom-up approach. It also follows top-down strategy.
- To merging two cluster based on closeness of these cluster.

2.4.3. Density-based:

Density-based algorithm “can find a cluster of random shapes unexpectedly. To group objects, it uses the density objective function.in these method clusters will increase until the number of the article in the nearby growing some limitation. The

reason behind density-based clustering is to detect cluster of non-spherical or arbitrary shapes.it is unsupervised learning method that find distinct group in the information this is used for clustering of very large database. Some of the common technique is DBSCAB, OPTICS, VDBSCAN, DVBSCAN, DBCLASD, and ST-DBSCAN AND DENCLUE.

2.4.4. Grid based:

Grid-based clustering "quantizes the object space into a finite number of cells that form a grid structure. After calculating density grid-based clustering, sort the cells according to their densities. Cluster centers are identified, and all neighbor cells are traversed."

2.4.5. The objective of algorithm:

Considering the d-dimension data set$\{x_i | x_i \in R^d, i = 1,2, \dots N\}$, the method to follow easy steps to classify given dataset, number of clusters $w_1, w_2, \dots . w_k$ to define K centroid $c = c_1, c_2, \dots . c_k$, one for each group,

$c_i = \frac{1}{n_i} \sum_{x \in wi} . x$ Where n_i Are various datasets in the cluster?

Select all the closest points in the dataset and do the closest one. The step is competition. Re-sour the centrifugal cluster obtained from the previous light

A new link has been created between the current record and the next new approach, the iteration cycle. This is because the center of gravity of this loop is slowly moving until there is no change. Finally, an algorithm that minimizes the objective function.

The objective function $J = \sum_{i=1}^{k} . \sum_{j=1}^{n_i} . d_{ij}(x_j, c_i)$ where $d_{ij}(x_j, c_i)$ Distance between the data point x_j and cluster center c_i.

2.4.6. Steps by steps algorithm:

Step1: (initialize) randomly choose K occurrence $c_1, c_2, \dots . c_k$ from the data set X and start cluster center of the clustering space.

Step2: (allotment) provide each occurrence to the nearest center: $d_{ij}(X_i, C_j) < d_{ij}(X_i, C_m)$ i.e., $j = 1,2 \ldots\ldots k$ & $i = 1,2 \ldots\ldots n$ $j \neq m$, $m = 1,2 \ldots\ldots k$ and then allocate x_i to c_j

Step3: (updating) Recalculate the centroid of the cluster$c_1{}^*, c_2{}^*, \ldots\ldots\ldots c_k{}^*$;

Step4: (loop) if $i = \{1 \ldots\ldots k\}$ $c_i{}^* = c_i$ then stop the algorithm and initial $c_1{}^*, c_2{}^*, \ldots\ldots\ldots c_k{}^*$ represent the end group, or else allocate $c_i = c_i{}^*$ & perform step 2 and 3 until there isn't any more modification

2.4.7. Advantage and disadvantage:

K-mean algorithm is essential for a considerable large dataset, and its time complexity O (tkn), where t is the number of loops, k is the number of clusters, and n is the data point in the dataset.

With the k-mean algorithm, the cluster number is required first, yet the number of clusters is usually determined later. It is sensitive to outliers.

It is not easy to predict k-value, the k-mean algorithm with global cluster does not work well, and it does not perform well with a different group and size.

CHAPTER-3

RESULT ANALYSIS

This module describes about the clustering with different type of distance measure used in k-mean to analysis the result to see the reaction of changes in cluster distance.

3.1. Analysis of k-means:

In this we are using k-mean with 8 different datasets to calculate distance measure each You can visually check the data points of each group and the result of grouping.

The circular group in Figure 1 evaluates the effect of changing the distance between the two groups, simulating for the elliptical group. The distance between the two clusters is defined as:

$$R_{distance} = \frac{D}{r_A + \gamma_B}$$

Distance between two cluster is D radius of the circular cluster is r_A and γ_B is the simi minor-axis of the elliptic cluster. Both $r_A \, and \, \gamma_B$ are fixed ($r_A = 14.37 \, and \, \gamma_B = 9.97$) and D varies as the cluster move toward each other.

Where the distance ratio $R_{distance}$ is less than 1 the two-cluster overlap. The clustering algorithm should work efficiently with as few $R_{distance}$ as possible, using measures of various lengths to see the effect of cluster changes.

3.1.1 City block distance measure

In the city-blocks measure it consider two-point a and b with k dimensional is defined as

$$\sum_{j=1}^{k} |a_j - b_j|$$

City blocks is also known as Manhattan let consider two-point x-y plane, we consider shortest distance between the two points is along the hypotenuse which is Euclidean

distance. The distance between blocks is used to calculate the fuzzy membership function

The center of town.

Instead, the city block distance is calculated as x distance plus y distance. This is similar to how you travel in a city (for example, Manhattan) where you need to go through the building instead of going through it.

You can move only one dimensional of the space at a time in city-block by analogy. It does not affect distance in the space note also that many equal length paths exist between two point in city- block space.

Euclidean distance and city-block distance are special case (different value of k) of the metric in two dimensions.

$$\text{Distance}= \sqrt[k]{x^k + y^k}$$

Here x and y are distance in each of two dimensions generalizing this to P dimension and using the form of the equation of ED

In this two coordinate X-axis and Y-axis using with first numeric dataset in these D=48 and $R_{distance} = 1.99$

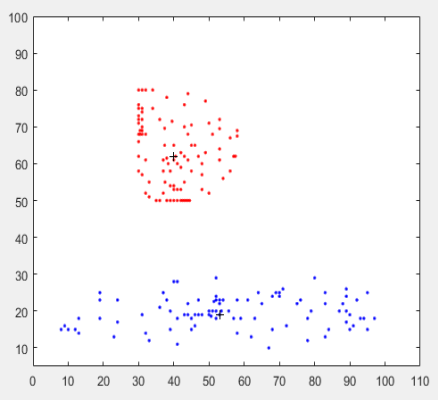


Fig1.1 cityblock

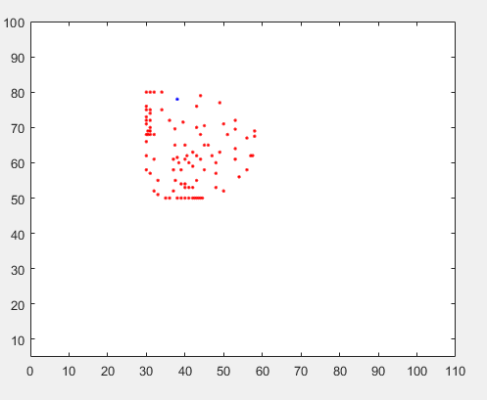
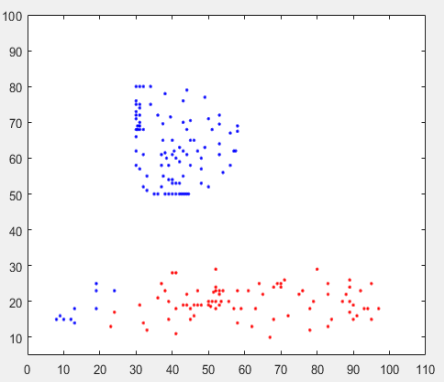


Fig1.2 correaltion



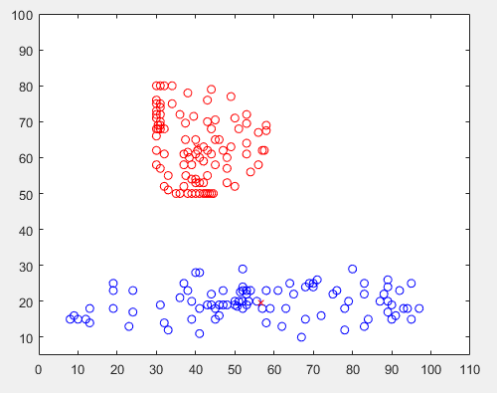


Fig 1.3 cosine

Fig 1.4 Euclidean

In the next dataset here D=42 and $R_{distance} = 1.74$

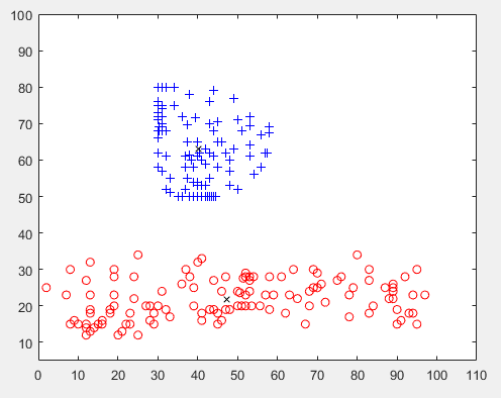


Fig 2.1Euclidean

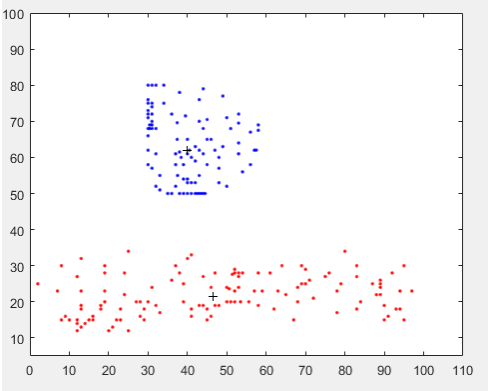


Fig 2.1 city block

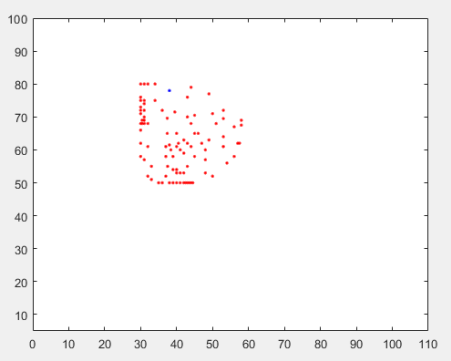


Fig 2.3 correlation

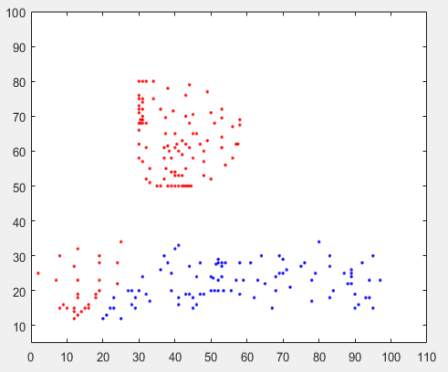


Fig 2.4 cosine

In the next dataset we using D=30 and $R_{distance} = 1.25$

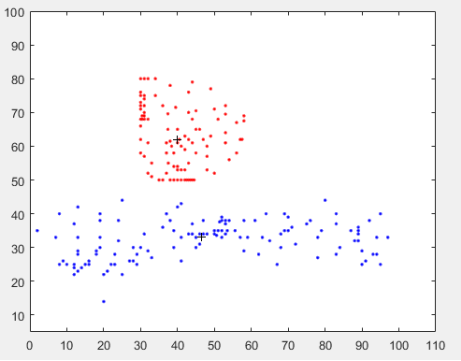


Fig 3.1 city block

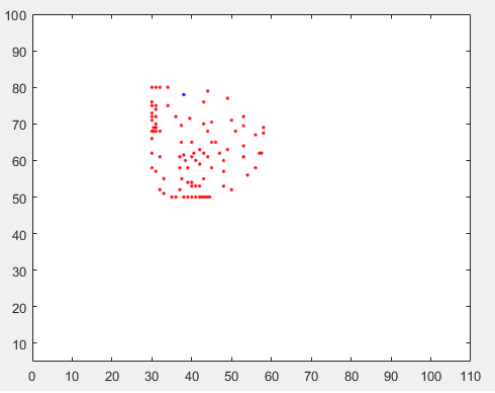


Fig 3.2 correlation

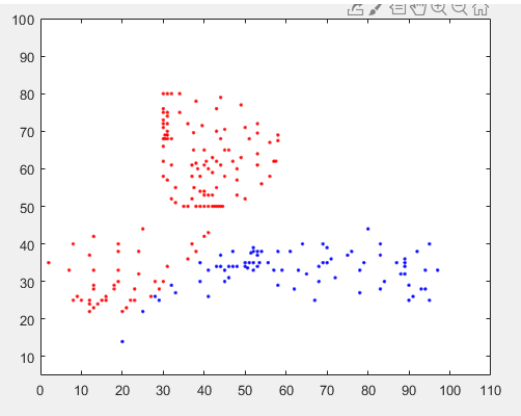


Fig 3.3 cosine

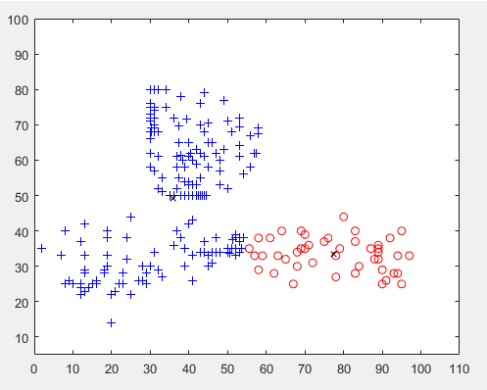


Fig 3.4 Euclidean

3.1.2 Euclidean distance

In Euclidean distance there are two-point a and b with k dimension is calculated

$$\sqrt{\sum_{j=1}^{k}(a_j - b_j)^2}$$

In k-Means, the Euclidean distance is used by default, and the distance is usually calculated based on raw data rather than standardized data. The distance is affected by the proportional difference between the dimensions of the calculated distance. One of the advantages of this approach is the distance between two features, which is not affected by adding new features that may be outliers to analyze.

For example, when one of the dimensions given in millimeters and then you converted into centimeters, the resulting Euclid can be manipulated, so the grouping results can be different. The distance between two objects that are not points is generally defined as the minimum distance between a pair of points on two objects. The formula for calculating the distance between different types of objects is well known, such as the distance from a point to a straight line, the concept of distance is expanded to include abstract metric space, and the study focuses on distance rather than Euclidean increase. Many statistics and optimized applications use the Euclidean square area instead of the area itself.

One dimension -

- The distance between the points on the solid line is the absolute value of the difference between the coordinate values. So, if p and q are two points on the solid line, the distance between them is determined by the expression

$$d(p,q) = \sqrt{(p-q)^2}$$

Fig display the clustering result of the clustering algorithm for the simulated dataset decreasing $R_{density}$ value from 1.00 to 0.10. to evaluate the effect of non-circular data cluster a circular cluster of fixed radius and an elliptic cluster with varying semimajor-axis are simulated the roundness measure of the non-circular cluster is defined as

$$R_{roundness} = \frac{\gamma_2}{\gamma_1}$$

In the first dataset $r_B = 25\ r_{density} = 0.48$

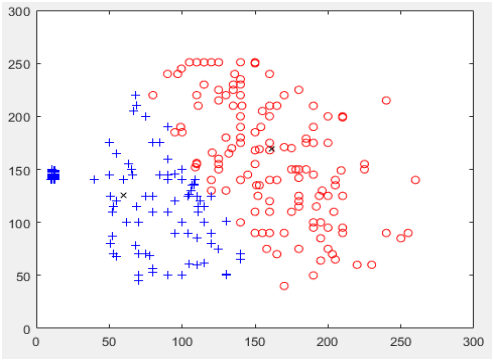


Fig 4.1 Euclidean

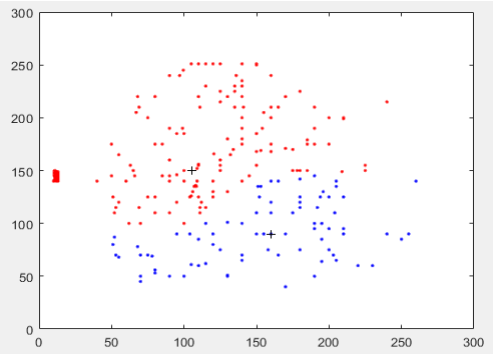


Fig 4.2 city block

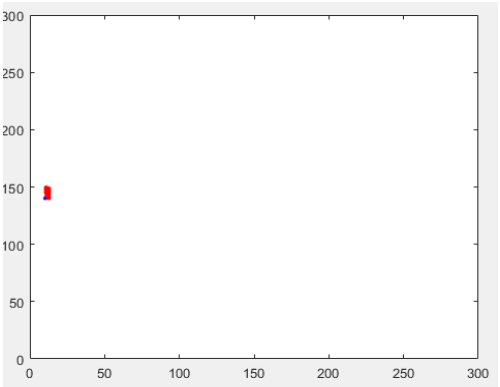


Fig 4.3 correlation

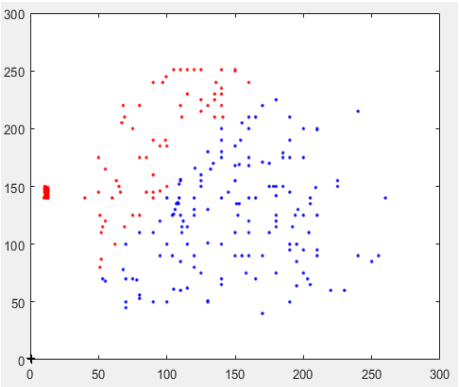


Fig 4.4 cosine

In the second dataset $r_B = 113\ and\ r_{density} = 0.11$

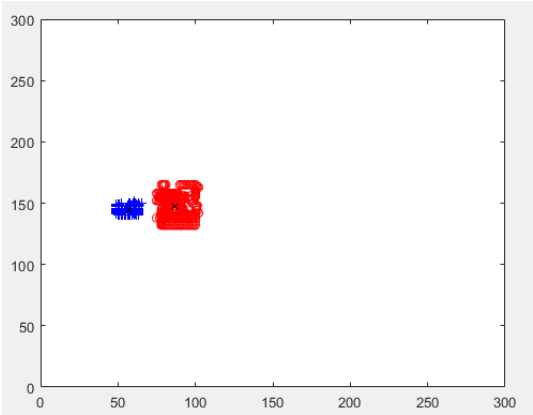


Fig 5.1 Euclidean

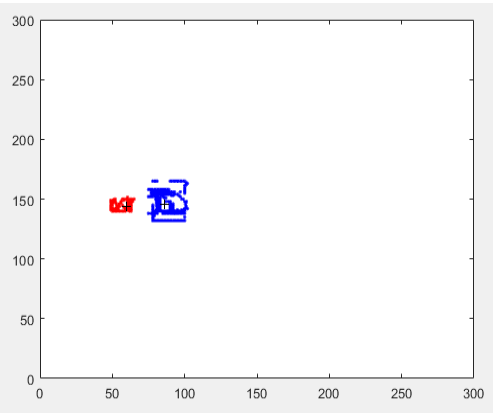


Fig 5.2 city block

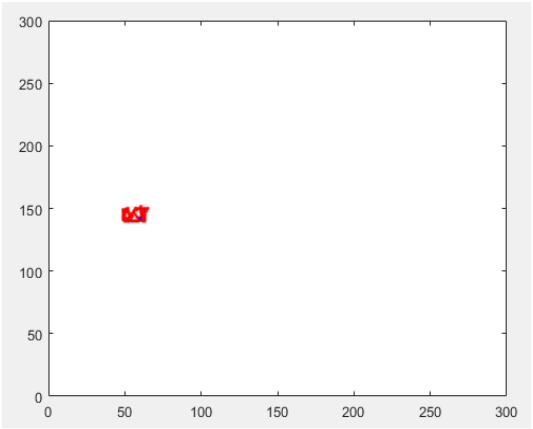


Fig 5.3 correlation

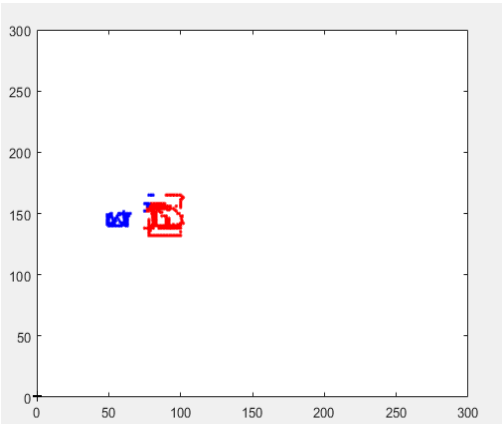


Fig 5.4 cosine

3.1.3 Cosine distance

The cosine distance is the square of the Euclidean distance and the data is normalized to the unit of length. The distance between two points is no minus the cosine of the angle between the points (as a vector). Consider a data matrix X of size mxn as a row vector x1, x2 ... ... xm of size m (1 xn) The cosine distance between vectors xs and xt is defined as follows

$$d_{st} = 1 - \frac{x_s x_t}{\sqrt{(x_s x_s')(x_t x_t')}}$$

This figure shows the results of grouping the three data sets using a more detailed dual-core grouping method.

The figure below shows this data set. The data set consists of two hearts in a circle. We also evaluated many inseparable linear data sets of different shapes and densities.

Experience.

First dataset Clustering results of linearly non-separable data sets from

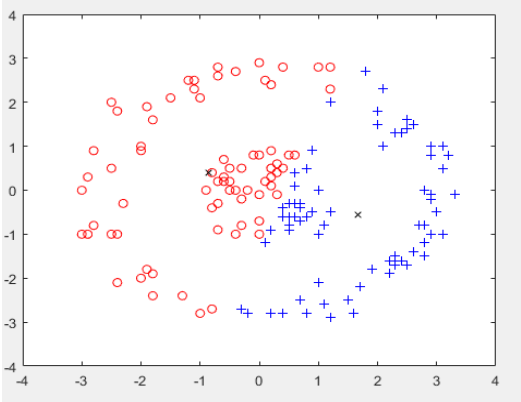


Fig 6.1 Euclidean

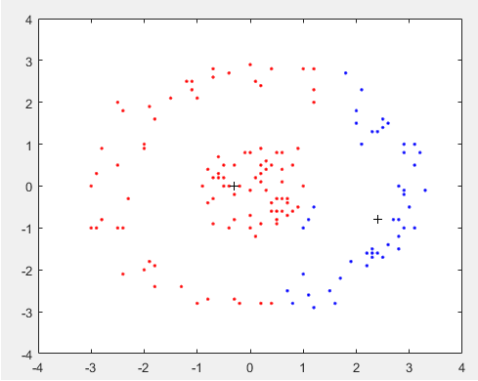


Fig 6.2 city block

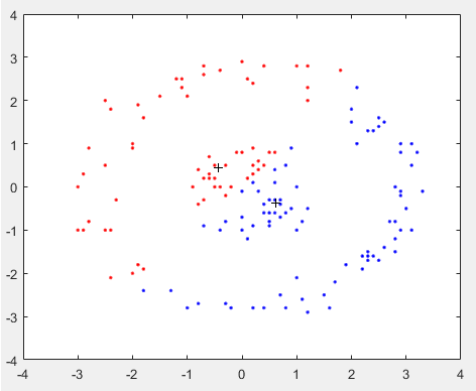


Fig 6.3 cosine

In the second dataset Clustering results of linearly non-separable data sets from

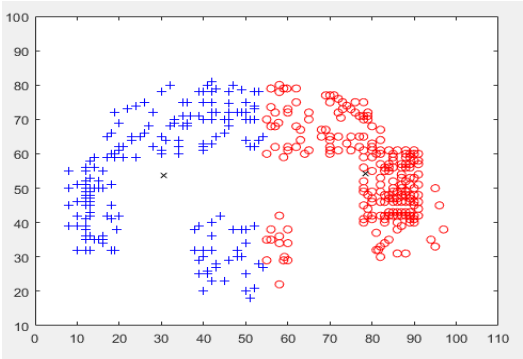


Fig 7.1 Euclidean

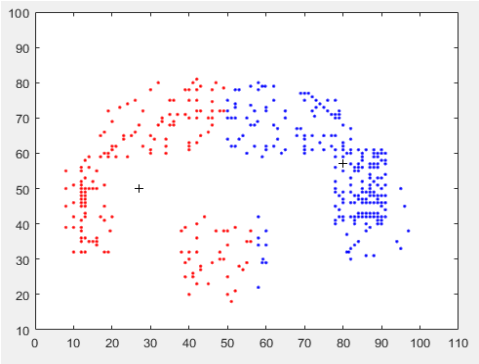


Fig 7.2 City block

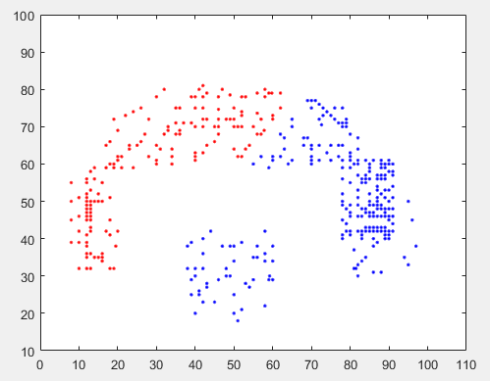


Fig 7.3 cosine

In third data set Clustering results of linearly non-separable data sets from

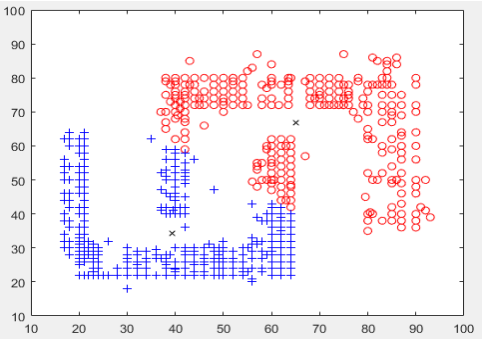


Fig 8.1 Euclidean

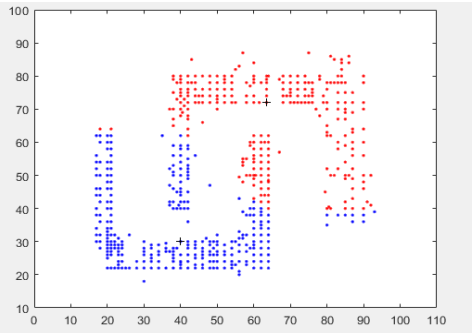


Fig 8.2 City block

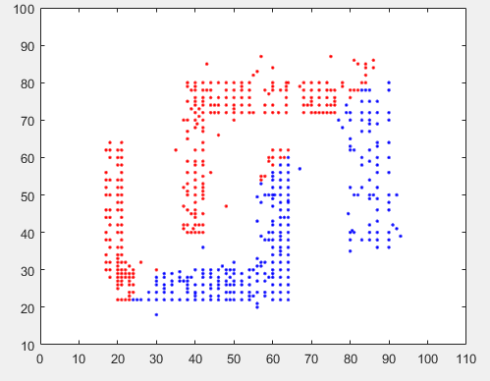


Fig 8.3 Cosine

City blocks	Euclidean	Cosine	correlation
$\sum_{j=1}^{k}\lvert a_j - b_j\rvert$	$\sqrt{\sum_{j=1}^{k}(a_j - b_j)^2}$	$d_{st} = 1 - \frac{x_s x_t}{\sqrt{(x_s x_s')(x_t x_t')}}$	$d_{st} = 1 - \frac{(x_s - x_s')(x_t - x_t')'}{\sqrt{(x_s - x_s')(x_s - x_s')'}\sqrt{(x_t - x_t')(x_t - x_t')'}}$

Table 3.1 distance measure formula

CHAPTER-4

IMPLEMENTATION AND RESULTS

4.1 Datasets Used

The data is generated in a 2D plan to observe and validate the group results for each data point and each group. First, assess the variation in the distance effect between the two clusters.

This report uses four public database datasets to assess the performance of two clustering algorithms: Astroparticle, Splice, Log Stat (Australian Credit Approved), and Iris. The astroparticle and splice datasets come from the Australian Statlog / LIBSVM (Reference Vector Machine Library) and the datasets come from the VCI (Irvine VC Machine Learning Repository).

Cosmic Particle Dataset is an application for classifying cosmic particle physics. It contains 2 classes and 4 functions. Australian records / files apply to credit card applications. Credit cards are approved or rejected based on 14 features of the application. The purpose of a splice data set is to identify two classes of splice sections in a DNA sequence. Each property is represented by 60 large properties. The iris data set iris (setosa, versicolor, virginica) contains three categories. Describe four iris plants. Table 1 shows the detailed characteristics of the data for the four data sets.

Table 1
Data characteristics of real data sets.

Data set	Number of clusters	Number of features	Number of data points (samples in each cluster)	Description	Source
Astroparticle	2	4	3089 (2000/1089)	Astronomical application	LIBSVM
Statlog (Australian credit approval)	2	14	690 (307/383)	Credit card approval	UCI
Splice	2	60	1000 (517/483)	Splice junctions in DNA sequence	LIBSVM
Iris	3	4	150 (50/50/50)	Iris plant	UCI

Table 4.1 Data characteristic of real data set

4.2 Implementation

The implementation of the work is done over the MATLAB based on clustering algorithm such as k-mean algorithm used different distance matrix and FCM is given below.

4.2.1 Code:

k-means for Euclidean distance measure

```
edit kmeans
>>Data= [X; Y]:
>> k=2; [IDX, C] = kmeans (data, k);
>> plot (data (IDX==1,1), data (IDX==1,2),'ro'); hold on;
>> plot (data (IDX==2,1), data (IDX==2,2),'bo'); hold on;
>> plot (C (:1), C(:2),'rx'); hold on;
>> axis ([0 110 5 100])
```

K-means for cosine distance measure

```
%edit kmeans
data= [X; Y];
k=2; [IDX, C] =kmeans (data,2,'Distance', 'cosine');
plot (data (IDX==1,1), data (IDX==1,2),'r.'); hold on;
plot (data (IDX==2,1), data (IDX==2,2),'b.'); hold on;
plot(C(:,1),C(:,2),'k+');
axis ([0 110 5 100])
data= [X; Y];
```

k-means for city blocks distance measure

```
data= [X; Y];
k=2; [IDX, C] =kmeans (data,2,'Distance', 'city block');
plot (data (IDX==1,1), data (IDX==1,2),'r.'); hold on;
plot (data (IDX==2,1), data (IDX==2,2),'b.'); hold on;
plot (C (:,1), C(:,2),'k+');
axis ([0 110 5 100])
data= [X; Y];
```

k-means for correlation distance measure

```
data= [X; Y];
[cidx, ctrs] = kmeans(X,2,'Distance','correlation');
 plot(data(cidx==1,1), data(cidx==1,2),'r.'); hold on;
 plot(data(cidx==2,1), data(cidx==2,2),'b.'); hold on;
 plot(ctrs(:,1), ctrs(:,2),'kx'); hold on;
 axis ([0 110 5 100])
 data= [X; Y];
```

X=[38 78;36 72;39.5 71.5;37.4 69.6;37.4 65;40 65;30 66;30 62;37 61;38 61.5;38.4 60;42 59;37 58;37.5 55;39 58;39 54;40 53;45 58;48 57;48 53;50 52;54 56;56 58;44 79;43 76;49 77;50 71;53 72;43 70;45 70.5;44 68;51 68;53 69.5;58 69;56 67;58 67.5;57.5 62;57 62;53 64;53 61;49 63;32 52;33 51;30 58;31 57;33 51;37 52;33 55;32 61;35 50;36 50;38 50;39 50;40 50;41 50;42 50;43 50;42.5 50;44 50;43.5 50;44.5 50;30 68;30.4 68;31 69;32 80;30.4 69;30 71;30 72;30 73;30 75;30 76;31 68;31 72;31 75;31 69;31 70;31 74;32 68;34 75;34 80;32 80;31 80;30 80;40 61;40.5 62;44 61;42 63;45 65;47 62;48 60;41 60;43 62;46 65;40 53;41 53;45 58;42 53;45 58;43 55;40 54]	Y=[2 25;7 23;8 30;13 19;13 23;12 27;13 32;19 23;19 28;19 30;23 18;24 22;24 28;25 34;31 24;32 19;33 17;36 26;38 28;37 30;39 20;39 25;40 32;41 33;41 16;44 27;47 28;46 24;57 23;58 28;58 19;59 23;61 28;63 23;62 18;64 30;67 15;68 20;69 30;70 29;72 21;75 27;76 28;78 17;78 23;79 25;80 34;83 27;83 30;83 18;84 20;87 25;89 25;89 22;92 28;94 23;97 23;95 30;52 23;53 28;50 24;54 28;50.6 23.6;51.4 27.6;53 24;52 28;53 27;52 29;70 25;71 26;69 25;68 24;65 22;89 24;88 22;89 26;90 15;91 16;94 18;90 19;93 18;95 15;8 15;9 16;12 15;20 4;10 15;12 12;12 14;12 15;16 15;13 18;15 15;13 13;14 14;16 16;18 18;19 20;19 20;18 19;20 12;21 13;23 15;25 12;29 15;22 15;29 18;30 20;28 16;28 20;27 20;41 18;45 18;43 19;44 19;45 15;46 16;47 19;48 19;50 20;51 20;52 20;55.6 20;53.5 20]

Kmeans correlation
data= [X; Y];
[cidx, ctrs] = kmeans(X,2,'Distance','correlation');
plot(data(cidx==1,1), data(cidx==1,2),'r.'); hold on;
plot(data(cidx==2,1), data(cidx==2,2),'b.'); hold on;
plot(ctrs(:,1), ctrs(:,2),'kx'); hold on;
axis ([0 110 5 100])
data= [X; Y];
	Kmeans Euclidean
	edit kmeans
	>> k=2; [IDX, C] = kmeans (data, k);
	>> plot (data (IDX==1,1), data (IDX==1,2),'ro'); hold on;
	>> plot (data (IDX==2,1), data (IDX==2,2),'bo'); hold on;
	>> plot(C(:,1), C(:,2),'rx'); hold on;
	>> axis ([0 110 5 100])

Figure 4.2 dataset

CHAPTER-5

CONCLUSION AND FUTURE WORK

This article introduces a new different distance metric that includes the distance of each group of centroids from the k-means algorithm as well as the variation of any point. In this report we use the different distance measure to different dataset we using to compare the performance analysis based on that we have found that Euclidean distance and city block similar result we have performed our experience on four datasets. By default, the performance of a dataset is estimated by comparing various metrics of cosine interval, correlation, city-block, and Euclidean in the k-means algorithm. Euclidean distance measurements are used for an actual dataset in a public database. we can also try to extend our study for another partition algorithm FCM (Fuzzy c-mean algorithm) on different distance measure.

CHAPTER-6
REFERENCES

[1] Application of fuzzy theory to term-based portfolio selection," *Fourth International Symposium on Uncertainty Modeling and Analysis, 2003. ISUMA 2003.*, 2003, pp. 198-202, doi: 10.1109/ISUMA.2003.1236162.

[2] Q. Zhao, G. Li and S. Xing, "FCM Algorithm Based on the Optimization Parameters of Objective Function Point," *2010 International Conference on Computing, Control and Industrial Engineering*, 2010, pp. 331-333, doi: 10.1109/CCIE.2010.200.

[3] Pusadan, Mohammad & Buliali, Joko & Ginardi, R.V.Hari. (2020). Cluster Phenomenon to Determine Anomaly Detection of Flight Route. 10.1016/j.procs.2019.11.151.

[4] B. S. Shedthi, S. Shetty and M. Siddappa, "Implementation and comparison of K-means and fuzzy C-means algorithms for agricultural data," *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 2017, pp. 105-108, doi: 10.1109/ICICCT.2017.7975168.

[5] A. Kapoor and A. Singhal, "A comparative study of K-Means, K-Means++ and Fuzzy C-Means clustering algorithms," *2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT)*, 2017, pp. 1-6, doi: 10.1109/CIACT.2017.7977272.

[6] KOÇ and T. ÖLMEZ, "Improved Fuzzy C-means and K-means Algorithms for Texture and Boundary Segmentation," *2018 6th International Conference on Control Engineering & Information Technology (CEIT)*, 2018, pp. 1-6, doi: 10.1109/CEIT.2018.8751905.

[7] Shao-Hong Yin and Min Li, "Study on a modified Fuzzy C-Means Clustering Algorithm," *2010 International Conference On Computer Design and Applications*, 2010, pp. V5-484-V5-486, doi: 10.1109/ICCDA.2010.5541025.

[8] Comparative Study of Distance Measures for the Fuzzy C-means and K-means

Non-Supervised Methods Applied to Image Segmentation Martín Vélez-Falconía,b , Josué Marína, b, Selena Jiménezа and Lorena Guachi-Guachia,b,c

[9] Comparison of K-Means and Fuzzy C-Means Algorithms on Different Cluster Structures Zeynel Cebeci1 , Figen Yildiz 2 2015

[10] M. A. Nadaf and S. S. Patil, "Performance evaluation of categorizing technical support requests using advanced K-Means algorithm," *2015 IEEE International Advance Computing Conference (IACC)*, 2015, pp. 409-414, doi: 10.1109/IADCC.2015.7154740

[11] M. A. Nadaf and S. S. Patil, "Performance evaluation of categorizing technical support requests using advanced K-Means algorithm," *2015 IEEE International Advance Computing Conference (IACC)*, 2015, pp. 409-414, doi: 10.1109/IADCC.2015.7154740.

[12] Finding Similarity in Articles using Various Clustering Techniques Deeksha1 , Shashank Sahu2

[13] Effect of Different Distance Measures on the Performance of K-Means Algorithm: An Experimental Study in Matlab Dibya Jyoti Bora, Dr. Anil Kumar Gupta Department of Computer Science & Applications, Barkatullah University, Bhopal, India

[14] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

[15]