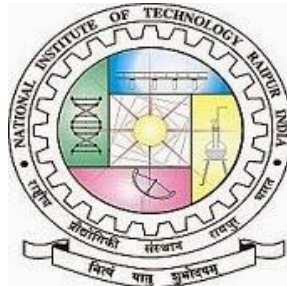


EXCEL IMPLEMENTATION OF REGRESSION CLUSTERING

B.Tech. Major Project Report

BY

**ABHISHEK MAHESHWARI (11115004)
TORAN SAHU (11115086)**



**DEPARTMENT OF COMPUTER SC. & ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY
RAIPUR - 492010, CHHATTISGARH, INDIA**

MAY, 2015

EXCEL IMPLEMENTATION OF REGRESSION CLUSTERING

A Major Project Report

Submitted in Partial Fulfillment of the Requirements for the Award of

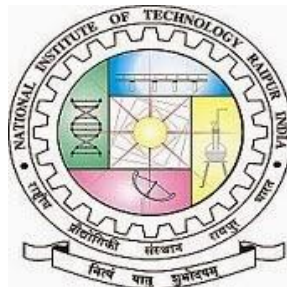
Bachelor of Technology
In
Computer Science & Engineering
(April-May 2015)

BY

ABHISHEK MAHESHWARI (11115004)
TORAN SAHU (11115086)

Under the Guidance of

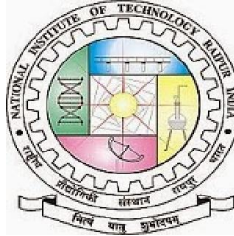
Dr. N. K. NAGWANI
(Assistant Professor)



DEPARTMENT OF COMPUTER SC. & ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY
RAIPUR - 492010, CHHATTISGARH, INDIA

MAY, 2015

**DEPARTMENT OF COMPUTER SC. & ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY
RAIPUR**



CERTIFICATE

I hereby certify that the work which is being presented in the B.Tech. Major Project Report entitled “**Excel Implementation of Regression Clustering**”, in partial fulfillment of the requirements for the award of the **Bachelor of Technology in Computer Sc. & Engineering** and submitted to the Department of Computer Sc. & Engineering of National Institute of Technology Raipur is an authentic record of my own work carried out during a period from January 2015 to May 2015 under the supervision of **Dr. N. K. NAGWANI** (Assistant Professor), CSE **Department**.

The matter presented in this thesis has not been submitted by me for the award of any other degree elsewhere.

Signature of Candidate

ABHISHEK MAHESHWARI

R.No. 11115004

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Head of the Department

Dr. N. K. NAGWANI

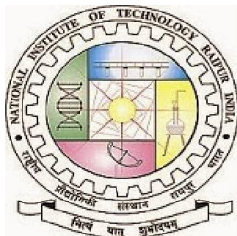
Assistant Professor
Computer Sc. & Engineering Department
NIT Raipur

Guide

Dr. N. K. NAGWANI

Assistant Professor
Computer Sc. & Engineering Department
NIT Raipur

**DEPARTMENT OF COMPUTER SC. & ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY
RAIPUR**



CERTIFICATE

I hereby certify that the work which is being presented in the B.Tech. Major Project Report entitled “**Excel Implementation of Regression Clustering**”, in partial fulfillment of the requirements for the award of the **Bachelor of Technology in Computer Sc. & Engineering** and submitted to the Department of Computer Sc. & Engineering of National Institute of Technology Raipur is an authentic record of my own work carried out during a period from January 2015 to May 2015 under the supervision of **Dr. N. K. NAGWANI** (Assistant Professor), CSE **Department**.

The matter presented in this thesis has not been submitted by me for the award of any other degree elsewhere.

Signature of Candidate

TORAN SAHU

R.No. 11115086

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Head of the Department

Dr. N. K. NAGWANI

Assistant Professor
Computer Sc. & Engineering Department
NIT Raipur

Guide

Dr. N. K. NAGWANI

Assistant Professor
Computer Sc. & Engineering Department
NIT Raipur

ACKNOWLEDGEMENT

We the student of VIII semester, Computer Sc. & Engineering Branch, NIT Raipur, extend our heartfelt thanks to our project guide **Dr. N. K. NAGWANI**, Assistant Professor, Computer Sc. & Engineering Branch, NIT Raipur for providing us an interesting topic for our project work and guiding us at every juncture to complete it successfully.

We would also thankful to all the staff members and colleagues for their encouraging support for the accomplishment of this project.

Abhishek Maheshwari

Toran Sahu

8th Semester

Comp. Sc. & Engg.

NIT Raipur

ABSTRACT

Nowadays Excel files carry most of the datasets created for various usages. Extracting information from these Excel file is an important task. So to fulfill these purposes clustering regression is applied to the Excel File. Prediction of attributes in these Excel files is important for some activities. Regression techniques are most widely used for prediction task where relationship between the independent variable and dependent variable is identified. The accuracy of the regression techniques for prediction can be improved if clustering can be used along with regression. Clustering along with regression will ensure the more accurate curve fitting between the dependent and independent variables. The objective of this proposed work is to find an optimum number of clusters in which original dataset should be clustered to ensure less prediction errors for estimating the value of dependent variable. The proposed project consists of four major stages, first of all data preparation is carried by extracting data from Excel file; in the second stage, clustering is used to group the similar type of data, in third stage regression techniques are applied over these groups (clusters) to predict the dependent variable value from individual clusters, and the last stage task concludes with finding the cluster count for which minimum error is estimated. The output (clusters) are generated in Excel file format for further uses.

LIST OF FIGURES

2.1. Different ways of representing Clusters.....	2
2.2. Fragmentation Design of a sample Excel file.....	3
5.1. Methodology of the Project.....	9
5.2. Flow of the Project.....	10
9.1. Excel file data imported to MySQL database.....	18
9.2. Cluster label column inserted corresponding to nearest center.....	19
10.1. Prompt for Excel file path.....	20
10.2. Excel file path input.....	21
10.3. Cluster count (K) input.....	21
10.4. Selection of Mode of clustering.....	22
10.5. K-Means is selected for clustering.....	22
10.6. Manual input for Cluster Centroids.....	23
10.7. Prompt for Cluster # 1 Centre Values.....	23
10.8. Input for Cluster # 1 Centre Values.....	24
10.9. Input K Cluster Centre Values.....	24
10.10. Directory tree generated for K cluster count.....	25
10.11. i cluster files generated in i th directory.....	25
10.12. Completeness & consistency of clustering shown in Count.txt.....	25
10.13. Snap of generated .arff file from Example.Xls.....	26
10.14. Regression Equations generated for K=1.....	27
10.15. Weighted MAE and Weighted RMSE Calculated for K=1.....	27
10.16. Weighted MAE and RMSE are compared for all clusters.....	28

LIST OF TABLES

3.1. Equation for different clusters for estimating dependent variable values using Linear Regression	6
3.2. Equation for different clusters for estimating dependent variable values using LMS Regression.....	7
8.1. MAE and RMSE using Linear Regression for different no. of clusters.....	16
8.2. MAE and RMSE using LMS Regression for different no. of clusters.....	17

CONTENTS

Acknowledgment.....	i
Abstract.....	ii
List of Figures.....	iii
List of Tables.....	iv
1. General Introduction	
1.1. General Introduction.....	1
2. Clustering	
2.1. Introduction.....	2
2.2. Cluster Analysis.....	2
2.3. Clustering Terminology.....	3
2.4. Clustering Techniques.....	3
2.5. Clustering Algorithm	
2.5.1. K-Means.....	3
3. Regression	
3.1. Introduction.....	5
3.2. Objective of Regression.....	5
3.3. Applications.....	5
3.4. Models of Regression	
3.4.1. Linear Regression Model.....	5
3.4.2. Least Median of Squares Regression Model.....	6
4. Regression with Clustering	
4.1. Introduction.....	7
4.2. Advantages.....	7

5. Objective of the Project	
5.1. Introduction to the Project.....	8
5.2. Methodology of the Project.....	8
5.3. Flow of the Project.....	9
5.4. Modules Included.....	10
5.5. Feature Set.....	10
6. Clustering & Regression Using Java	
6.1. Preliminary Java.....	12
6.2. Java to MS Excel Connectivity.....	12
6.3. Java to MySQL Connectivity.....	13
6.4. Excel to Arff.....	13
6.5. APIs Used	
6.5.1. Apache POI.....	14
6.5.2. Weka.....	14
7. Error Calculation	
7.1. Introduction.....	15
7.2. Error Estimation Methods	
7.2.1. Mean Absolute Error.....	15
7.2.2. Root Mean Squared Error.....	15
8. Performance Measurement	
8.1. Introduction.....	16
8.2. Performance Measurement Methods.....	16
9. Role of Database Management	
9.1. Introduction.....	18
9.2. Requirement of SQL.....	18

10. Execution flow of the Project	
10.1. Introduction.....	20
10.2. Raw Input to the Project.....	20
10.3. Mode of Clustering	
10.3.1. K-Means.....	22
10.3.2. Manual (Cluster Center Input).....	23
10.4. Generating Clustered Excel File & Count File.....	25
10.5. Clustered Excel File to Arff File.....	26
10.6. Regression	
10.6.1. Linear Regression.....	26
10.6.2. LMS Regression.....	27
10.7. Error estimation	
10.7.1. MAE.....	27
10.7.2. RMSE.....	27
10.8. Error Comparison.....	28
11. Conclusion.....	29
12. Future Work.....	30
13. Reference and Bibliography.....	31
APPENDIX.....	32

Chapter 1

General Introduction

1.1 General Introduction

Data mining is defined as the process of analysing data from different viewpoints and summarizing it into useful information by extracting the information from the huge set of data. If simply stated, data mining refers to extracting or “mining” knowledge from large amounts of data [1]. In other words we can say that data mining is mining the knowledge from data. This project work targets similar applications of data mining techniques on database stored in MS Excel file format. MS Excel files carry most of the useful datasets in huge amount. So extracting information from these datasets is an important task.

In this proposed work the given MS Excel file is clustered and multiple MS Excel files are generated, each representing a separate cluster. These generated MS Excel files holds completeness and consistency with respect to the original MS Excel file. The generation of separate cluster files in MS Excel format favours better usage of data mining algorithms for knowledge extraction. Advantages of clustering a given huge MS Excel file includes the better prediction of a missing attribute value. Regression techniques are used to predict the value of an attribute based on multiple independent attributes. The error in prediction is minimized by selecting the optimum number in which the original MS Excel file is clustered. The selection of the optimum cluster count is done by comparing the weighted error estimated over a cluster count value.

Cluster analysis is an important research field it has its own unique position in a large number of data analysis and processing. A good clustering method will produce high quality clusters in which the intra-class similarity is high, the inter-class similarity is low, and the quality of a clustering result also depends on both the similarity measure used by the method and its implementation. The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns. Regression analysis is a statistical process for the prediction of dependent variables from the relationship among independent variables. Regression is done for a new data for which some estimation has to be done. First relationship equation is find out from the given test data and then that equation is used to predict the value of dependent variable.

Chapter 2

Clustering

2.1 Introduction

Cluster analysis is the process of making groups of similar objects in separate from a global set of objects. The objects in a group or cluster should have minimum intra cluster distance. Objects in same group are more similar to each other in comparison to the objects in other groups. Clustering is used in many fields nowadays like machine learning, pattern recognition, statistical analysis, image analysis and information retrieval.

2.2 Cluster Analysis

Cluster analysis is not a particular algorithm, but it is a process to be executed to get the more useful information[7]. Multiple algorithms can be applied to achieve this purpose. These algorithms differs in their notion, complexity and efficiency. The intended use of dataset and the intended use of the result defines the selection of algorithm and parameter setting. Cluster analysis is not an automatic task, it is an iterative process of knowledge extraction and it involves trials and failures. So until the result is achieved, the data is pre-processed and parameters are remodelled.

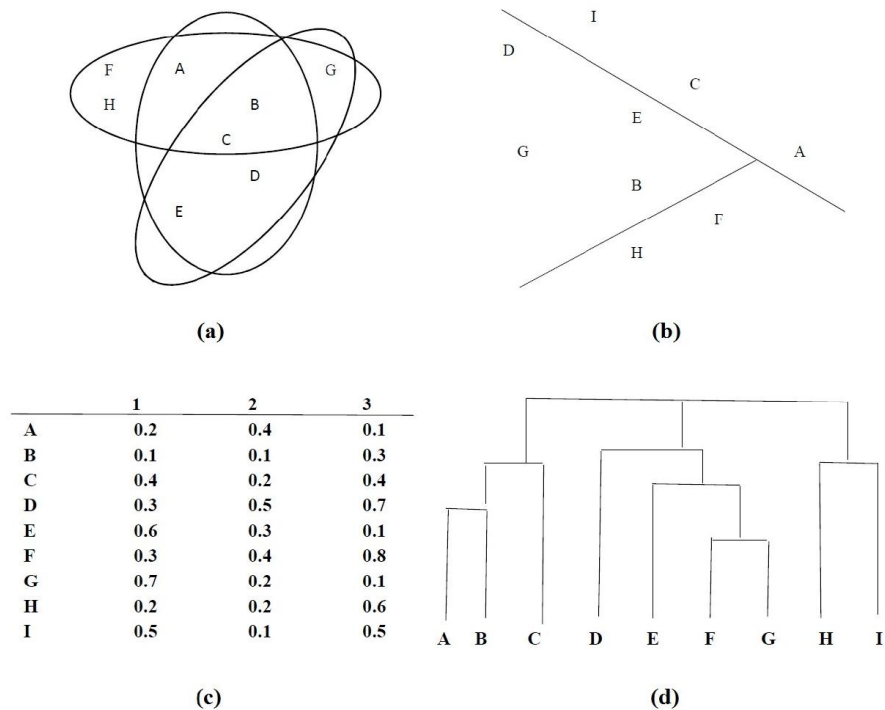


Fig 2.1: Different ways of representing Clusters

2.3 Clustering Terminologies

A suitable dataset for clustering is a collection of points belonging to some universal set of points. The dataset points are vectors of real numbers. The length of the vector is equivalent to the number of dimensions of the dataset. The components of the vector are called as coordinates of the represented points. All universal sets for which we can perform a clustering have a distance measure, giving a distance between any two points in the set. The common Euclidean distance is used for all dataset.

2.4 Clustering Techniques

Each row or data interpretation provided by user in .XLS file is compared using given Algorithm and then a cluster label column is added to the respective row. After addition of column, new files are created which depicts the clusters. The number of files generated by the program is equals to number of cluster centres provided by the user as input.

If k = number of cluster centres provided by user Example.xls is data file to be clustered, Then cluster1.xls, cluster2.xls, cluster3.xls upto clusterk.xls are generated.

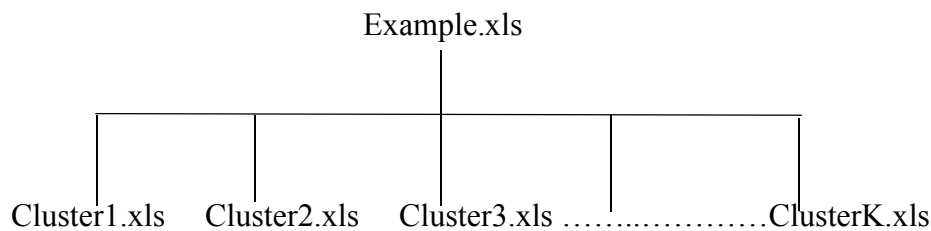


Fig 2.2: Fragmentation Design of a sample Excel file

Properties of fragmented Excel files:

1. Number of rows (example.xls) = Number of rows (Cluster1.xls)+ Number of rows (Cluster2.xls)+ Number of rows (Cluster3.xls)++ Number of rows (Clusterk.xls)
2. $\bigcap_{i=1 \text{ to } k} \text{cluster_i.xls} = \text{NULL}$

2.5. Clustering Algorithms

2.5.1. K-Means

K-Means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. K-Means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

Given a set of observations $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$, where each observation is a d -dimensional real vector, k -means clustering aims to partition the n observations into k ($k \leq n$) sets $\mathbf{S} = \{S_1, S_2, \dots, S_k\}$

so as to minimize the within-cluster sum of squares (WCSS). In other words, its objective is to find:

$$\arg \min \sum_{i=1}^k \sum_{x \in S_i} ||x - \mu||^2$$

Where μ_i is the mean of points in S_i .

The algorithm is often presented as assigning objects to the nearest cluster by distance. The standard algorithm aims at minimizing the WCSS objective, and thus assigns by "least sum of squares", which is exactly equivalent to assigning by the smallest Euclidean distance. Using a different distance function other than (squared) Euclidean distance may stop the algorithm from converging.

Euclidean distance is used to compute that a particular vector belongs to which Cluster by calculating and comparing different Euclidean distances.

Euclidean distance (d) = square root of sum of squared difference between respective dimensions of vector ($x_1, x_2, x_3, \text{upto } x_n$) and cluster mean vector ($\mu_1, \mu_2, \mu_3 \text{ upto } \mu_n$).

$$d = \sqrt{((x_1 - \mu_1)^2 + (x_2 - \mu_2)^2 + \dots + (x_n - \mu_n)^2)}$$

Algorithm of K-Means [3]

1. Clusters the data into k groups where k is predefined.
2. Select k points at random as cluster centers.
3. Assign objects to their closest cluster center according to the Euclidean distance function.
4. Calculate the centroid or mean of all objects in each cluster.
5. Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.

K-Means is relatively an efficient method. However, we need to specify the number of clusters, in advance and the final results are sensitive to initialization and often terminates at a local optimum.. A practical approach is to compare the outcomes of multiple runs with different k and choose the best one based on a predefined criterion. In general, a large k probably decreases the error but increases the risk of over fitting.

Chapter 3

Regression

3.1. Introduction

A statistical process for the estimation of the relationship between the dependent variable and one or more independent variables. There are several technique which solves this purpose. So the choice of technique chosen to solve the purpose of getting relationship among variables depends on the requirement to get the preciseness of the result. Regression analysis helps us in understanding that how the typical value of the dependent variable change while other variables are changed. Regression analysis is widely used for prediction and forecasting where it has significant overlap with the field of machine learning.

3.2. Objective of Regression

The main objective of Regression analysis is to predict or forecast the value of dependent variable. Regression analysis is to explain variability in dependent variable by means of one or more of independent or control variables.

3.3. Applications

There are three broad classes of applications of regression analysis.

- 1. Descriptive or explanatory:** interest may be on describing “What factors influence variability in dependent variable?” For example, factor contributing to higher sales among company’s sales force.
- 2. Predictive:** for example setting normal quota or baseline sales. We can also use estimated equation to determine “normal” and “abnormal” or outlier observations.
- 3. Decision purpose:** “What if” analysis can be done using regression, Estimating variable and fixed costs having calibrated cost function.

3.4. Models of Regression

3.4.1 Linear Regression Model

In statistics, linear regression is an approach for modelling the relationship between a dependent variable y and one or more independent variables denoted by X . The case of one explanatory variable (independent variable) is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.

Given data $\{y_i, x_{i1}, \dots, x_{ip}\}$ from $i = 1$ to n of n statistical units, a linear regression model assumes that the relationship between the dependent variable y_i and the p -vector of regression x_i is linear. This relationship is modeled through a disturbance term or error variable ε_i — an unobserved random variable that adds noise to the linear relationship between the dependent variable and regressions. Thus the model takes the form

$y_i = \beta_{1i}x_{i1} + \dots + \beta_{ip}x_{ip} + \varepsilon_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i$, $i = 1, \dots, n$.
where T denotes the transpose, so that $\mathbf{x}_i^T \boldsymbol{\beta}$ is the inner product between vectors \mathbf{x}_i and $\boldsymbol{\beta}$.

Often these n equations are stacked together and written in vector form as

$$Y = X\boldsymbol{\beta} + \varepsilon,$$

Where Y , X , $\boldsymbol{\beta}$ and ε are the matrices of the respective values.

Equation for each cluster file is generated for all K values, as follows

Table. 3.1: Equation for different clusters for estimating dependent variable values using Linear Regression

No. of Cluster	Cluster No.	Equation
1	1	SCF = -0.7815*D+-0.0324*R1+0.001*Delta1+2.6885
2	1	SCF = -0.3929*D+-0.0041*R1+-0.0001*Delta1+2.3781
	2	SCF = -1.9591*D+-0.0226*R1+0.0021*Delta1+2.89
3	1	SCF = -0.3929 * D + -0.0041 * R1 + -0.0001 * Delta1 + 2.3781
	2	SCF = -0.0213 * R1 + 0.0042 * Delta1 + 2.2511
	3	SCF = -0.0296 * R1 + -0.0004 * Delta1 + 2.5568
4	1	SCF = -0.3871 * D + -0.0042 * R1 + -0.0006 * Delta1 + 2.3834
	2	SCF = -0.0213 * R1 + 0.0042 * Delta1 + 2.2511
	3	SCF = -0.0296 * R1 + -0.0004 * Delta1 + 2.5568
	4	SCF = -0.4038 * D + -0.004 * R1 + 0.0002 * Delta1 + 2.368
5	1	SCF = -0.3871 * D + -0.0042 * R1 + -0.0006 * Delta1 + 2.3834
	2	SCF = -0.0213 * R1 + 0.0042 * Delta1 + 2.2511
	3	SCF = -0.0292 * R1 + -0.0017 * Delta1 + 2.5659
	4	SCF = -0.4038 * D + -0.004 * R1 + 0.0002 * Delta1 + 2.368
	5	SCF = -0.0205 * R1 + 0.0003 * Delta1 + 2.5239

3.4.2. Least Median of Squares

Classical least squares regression consists of minimizing the sum of the squared residuals. Many authors have produced more robust versions of this estimator by replacing the square by something else, such as the absolute value.

In this article a different approach is introduced in which the sum is replaced by the median of the squared residuals[6].

$$\min_{\boldsymbol{\theta}} \text{median}_i (y_i - y_i')^2$$

Equation for each cluster file is generated for all K values, as follows

Table. 3.2: Equation for different clusters for estimating dependent variable values using LMS Regression

No. of Cluster	Cluster No.	Equation
1	1	$SCF = -0.7753 * D + -0.0335 * R1 + 0.0005 * Delta1 + 2.7005$
2	1	$SCF = -0.4189 * D + -0.0028 * R1 + 0 * Delta1 + 2.3849$
	2	$SCF = -1.4192 * D + -0.0272 * R1 + 0 * Delta1 + 2.8271$
3	1	$SCF = -0.4189 * D + -0.0028 * R1 + 0 * Delta1 + 2.3849$
	2	$SCF = -0.0529 * R1 + 0.0021 * Delta1 + 2.624$
	3	$SCF = -0.0228 * R1 + 0 * Delta1 + 2.5358$
4	1	$SCF = -0.428 * D + -0.0021 * R1 + -0.0004 * Delta1 + 2.3948$
	2	$SCF = -0.0529 * R1 + 0.0021 * Delta1 + 2.624$
	3	$SCF = -0.0228 * R1 + 0 * Delta1 + 2.5358$
	4	$SCF = -0.445 * D + -0.0022 * R1 + 0.0001 * Delta1 + 2.3936$
5	1	$SCF = -0.428 * D + -0.0021 * R1 + -0.0004 * Delta1 + 2.3948$
	2	$SCF = -0.0529 * R1 + 0.0021 * Delta1 + 2.624$
	3	$SCF = -0.0334 * R1 + -0.0015 * Delta1 + 2.581$
	4	$SCF = -0.445 * D + -0.0022 * R1 + 0.0001 * Delta1 + 2.3936$
	5	$SCF = -0.0154 * R1 + 0.0001 * Delta1 + 2.5198$

Chapter 4

Regression with Clustering

4.1. Introduction

In the proposed work K-Means clustering technique along with the two regression technique, namely linear regression (LR), least median of squares (LMS), are used to predict the dependent variable value. After creating the group of similar records from the MS-Excel data, regression techniques are applied to identify the relationship between dependent variable with other independent components of the data.

4.2. Advantages

If the clustering can be combined along with the regression technique then estimation errors for data can be minimized. It is demonstrated from the experiments that if the optimum number of clusters can be created on raw data before applying regression then prediction errors can be minimized in efficient manner [4].

The accuracy of the regression techniques for prediction can be improved if clustering can be used along with regression. Clustering along regression will ensure the more accurate curve fitting between the dependent and independent variables.

Chapter 5

Objective of the Project

5.1 Introduction to the Project

The main objective of the project is to Cluster the MS-Excel file into separate clustered excel files and find the optimum number of cluster count (K) in which file should be clustered to get minimum error when predicting the value for dependent variable. The process starts by fragmenting the given data using cluster centres provided by the user as input or using K-means directly. The process of fragmentation incorporates the task to import Excel file to SQL database, processing data from SQL database by comparing respective distances from each cluster centre using Euclidean method in manual mode or using K-means and then applying regression on each cluster file separately to calculate errors and finally concludes by generating output as a K value against which minimum error is calculated. Input and output is handled using MS-Excel pack 2013.

5.2 Methodology of the Project

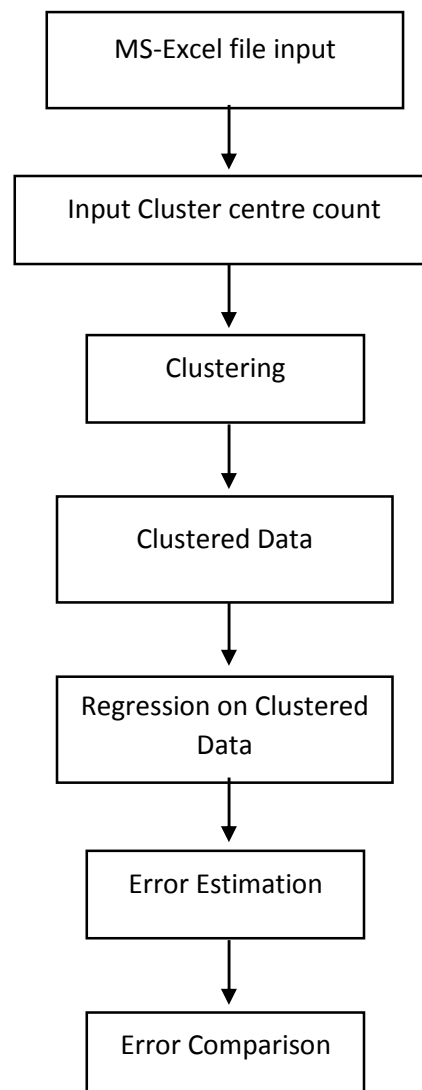


Fig. 5.1: Methodology of the Project

5.3 Flow of the Project

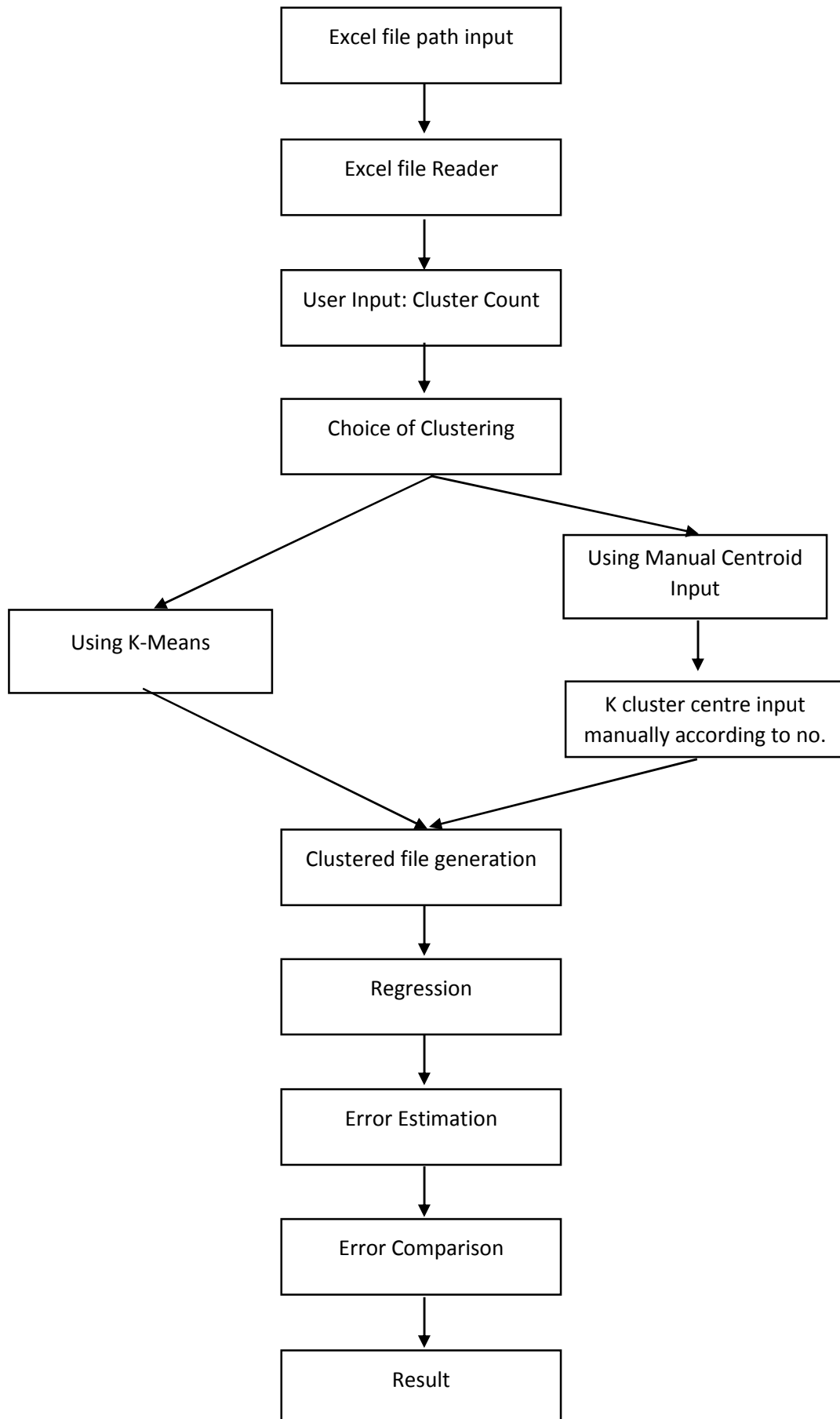


Fig. 5.2: Flow of the Project

5.4 Modules Included

1. **Excel File Reader:** This module takes file path in standard format and converts it in the specified format if required and extracts the filename.
2. **SQL Import:** This module imports the Excel file data in SQL format for better and efficient processing.
3. **Cluster Count (K):** Cluster Centre count is asked from user according to the requirement of user.
4. **Choice Of Clustering:** There are two choices of clustering
 - i. K- Means
 - ii. Manual
5. **Clustered Excel files:** In both modes the excel file is clustered and separate clustered excel files are generated in separate directories.
6. **Regression on Clustered Excel files:** After Separating Clusters for each K value both regression methods are applied on each cluster file and regression equation for data using each method is generated.
7. **Error Estimation:** MAE and RMSE errors are calculated for each cluster file using regression equation and then Weighted MAE and Weighted RMSE are calculated for each K value.
8. **Error comparison:** The stored errors are compared and the K value is found against which minimum error is calculated.

5.5 Feature Set

1. Software tools / IDE (integrated Development Environment)
 - a. Eclipse IDE
 - b. Java runtime Environment 6.0
 - c. MySQL Workbench 6.1CE
 - d. Microsoft Excel pack 2013

Chapter 6

Clustering & Regression Using Java

6.1. Introduction

Java is a computer programming language that is object-oriented, class-based, concurrent and specifically designed to make it platform independent. It also facilitates the developers "write once, run anywhere" (WORA) feature, means the code that runs on one platform can be run on another without recompiling it once again. Java applications are compiled to byte code that can run on any Java virtual machine (JVM) independent of computer architecture. Java is nowadays one of the most popular programming languages in use.

6.2 Java to MS Excel Connectivity (JDBC-ODBC Driver: DSN-Less)

A JDBC-ODBC bridge consists of a JDBC driver and an ODBC driver. JDBC driver uses an ODBC driver to connect to a target database. JDBC-ODBC driver makes translation of JDBC method calls into ODBC function calls. This bridge is used by programmers when a particular database doesn't supports a JDBC driver directly. By establishing this connection between Ms Excel file and Java program, data from excel file are imported to the MySQL database for further process.

Sample Code:

```
public void connToXls(String fp){
    try{
        Class.forName( "sun.jdbc.odbc.JdbcOdbcDriver" );
        //using DSN-less connection
        public static Connection cxl = DriverManager.getConnection( "jdbc:odbc:Driver={Microsoft
        Excel Driver (*.xls)};DBQ="+fp);
        public static Statement stmntxl = cxl.createStatement();
    }
    catch( Exception e ){
        System.err.println("connToXls method Exception in 'Excelreader.java file' : " +e);
    }
}
```

6.3. Java to MySQL Connectivity

Java Database Connectivity (JDBC) is used as an interface for accessing relational databases from Java. By using JDBC a connection to the database is established, database queries are issued and updated as well as results are received. JDBC provides an interface which allows you to perform SQL operations independent of the instance of the used database.

To connect to MySQL from Java, MySQL JDBC driver is required, that is called as *MySQL Connector/J*.

Sample Code:

```
public class ConnToSql {  
    public static Connection main()throws Exception{  
        final String JDBC_DRIVER = "com.mysql.jdbc.Driver";  
        final String DB_URL = "jdbc:mysql://localhost:3306/";  
        final String USER = "root";  
        final String PASS = "mysqlsamplepassword";  
        Connection csql = null;  
        try{  
            Class.forName(JDBC_DRIVER);  
            csql= DriverManager.getConnection(DB_URL, USER, PASS);  
        }  
        catch( SQLException e ){  
            System.err.println("Exception in connToSQL.java "+e);  
        }  
        return csql;  
    }  
}
```

6.4. Excel to ARFF

WEKA APIs take input from .arff (Attribute-relation File Format) file. Excel file is converted to .arff files before applying regression on that. The conversion of Excel to .arff file is done directly by getting connection from Excel and writing a new file with .arff extension and following the specified format in .arff online documentation.

ARFF format is:

@relation filename_name

@attribute attribute_name attribute_type

@attribute attribute_name attribute_type

@data

____ , ____

____ , ____

____ , ____

____ , ____

____ , ____

The above specified format is used as input to the WEKA APIs for finding regression equations.

6.5. APIs Used

6.5.1. Apache POI

It is a project run by the Apache Software Foundation. Apache POI (Poor Obfuscation Implementation) provides pure Java libraries for reading and writing, Microsoft Office formatted files, like Word, PowerPoint and Excel.

Writing an excel file: Writing a file using POI is very simple and involve following steps:

1. Create a workbook
2. Create a sheet in workbook
3. Create a row in sheet
4. Add cells in sheet
5. Repeat step 3 and 4 to write more data

6.5.2. WEKA

WEKA is a workbench (Software) for Waikato Environment for Knowledge Analysis. It is the incorporation of several standard ML (Machine Learning) techniques[8]. With it, a specialist in a particular field is able to use ML to derive useful knowledge from databases that are far too large to be analysed by hand. Weka API has most of the Data Mining functions which can be used directly to get information out of the raw database.

`weka.classifiers.functions.LeastMedSq.class` and `weka.classifiers.functions.Linear.class` is used to implement Least Median Regression and Linear Regression respectively. `weka.clusterers.SimpleKMeans` and `weka.core.Instances` are used to implement K-means algorithm for clustering the data. The only requirement for implementing Weka API inbuilt functionalities is that the database should be in .arff format. So for this to happen first the database in Excel format is converted into .arff format using simple Excel connectivity and direct file creating with .arff extension and following the .arff file format to write the created file.

Chapter 7

Error Calculation

7.1. Introduction

The standard error of the estimate is a measure of the accuracy of predictions. The regression equation is used to predict the value of dependent variable. So the measure of the difference between predicted and original value is called Error.

7.2. Error Estimation Methods

Two such common errors in regression based prediction are Mean Absolute Error (MAE) and Root Squared Error (RMSE). The Predicted value is termed as Y' and Original value is Y .

The difference of predicted value (Y') and original value (Y) is measure and used in both types of error estimation methods.

$$\text{Error in prediction is } = (Y - Y')$$

7.2.1. Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) is the Average of the absolute value of the residuals (error). The MAE is very similar to the RMSE but is less sensitive to large errors. The MAE is calculated using the following:

$$\begin{aligned} \text{MAE} &= \text{Average of all the errors in prediction over a cluster file.} \\ \text{MAE} &= \frac{1}{n} \left(\sum |y_i - \hat{y}_i| \right) \end{aligned}$$

7.2.2. Root Mean Squared Error (RMSE)

The Root Mean Squared error (RMSE) is the square root of the average squared distance of data point from the fitted line. The RMSE is calculated using the following:

$$\begin{aligned} \text{RMSE} &= \text{Square root of the average of squared errors in prediction.} \\ \text{RMSE} &= \sqrt{\frac{(\sum y_i - \hat{y}_i)^2}{n}} \end{aligned}$$

Chapter 8

Performance Measurement

8.1. Introduction

The performance of regression based prediction technique is carried in terms of errors in regression. Separate errors are calculated for each cluster file and then for every K value weighted errors are calculated over all 1, 2, 3, upto K cluster files.

As there are two error calculation methods MAE and RMSE. So accordingly there are two performance measures over each K value.

8.2 Performance Measurement Methods

After clustering is done, regression technique is applied on each individual clusters for predicting the value of dependent variable of the database. Two different Regression techniques SLR and LMS are applied on each individual cluster and prediction error MAE and RMSE are recorded [4].

The MAE and RMSE errors are tabulated in table no. 8.1 for Linear Regression

Table. 8.1: MAE and RMSE using Linear Regression for different no. of clusters

No. of Cluster	Cluster No.	No. of Records	Weighted MAE	Weighted RMSE
1	1	233	0.0415	0.0543
2	1	63	0.0136	0.0176
	2	170		
3	1	63	0.0012	0.0015
	2	91		
	3	79		
4	1	38	7.1938E-5	8.2584E-5
	2	91		
	3	79		
	4	25		
5	1	38	8.4169E-5	1.0028E-4
	2	91		
	3	54		
	4	25		
	5	25		

The MAE and RMSE errors are tabulated in table no. 8.2 for Least Median of Squares regression.

Table. 8.2: MAE and RMSE using LMS Regression for different no. of clusters

No. of Cluster	Cluster No.	No. of Records	Weighted MAE	Weighted RMSE
1	1	233	0.0406	0.0549
2	1	63	0.0129	0.0194
	2	170		
3	1	63	0.0578	0.05788
	2	91		
	3	79		
4	1	38	1.3001E-4	1.9226E-4
	2	91		
	3	79		
	4	25		
5	1	38	0.0108	0.0108
	2	91		
	3	54		
	4	25		
	5	25		

The overall weighted MAE and Overall weighted RMSE values are also recorded for each cluster. The overall weighted MAE and weighted RMSE are calculated as follows [4]:

$$\text{Weighted MAE} = \frac{\sum_{i=1}^k MAE_i \times N_i}{\sum_{i=1}^k N_i}$$

$$\text{Weighted RMSE} = \frac{\sum_{i=1}^k RMSE_i \times N_i}{\sum_{i=1}^k N_i}$$

Chapter 9

Role of Database Management

9.1. Introduction

A database management system (DBMS) is a collection of programs that enables you to store, modify, and extract information from a database. There are many different types of DBMSs, ranging from small systems that run on personal computers to huge systems that run on mainframes.

MySQL Workbench 6.1 CE:

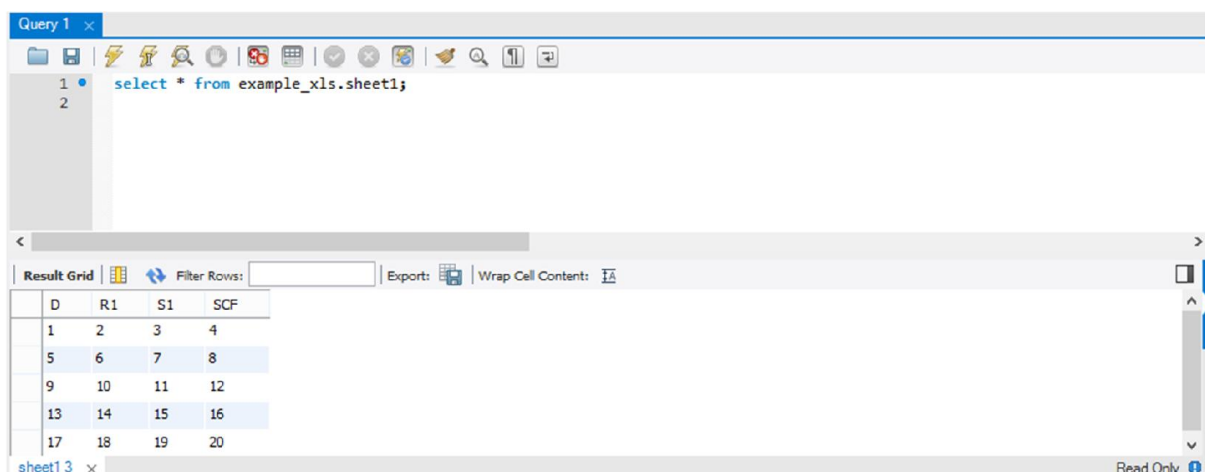
MySQL Workbench is a visual database design tool that integrates SQL development, administration, database design, creation and maintenance into a single integrated development environment for the MySQL database system.

9.2. Requirement of SQL

SQL provides a better support and easy implementation methods from Java language. On the other hand Excel Files are hard to handle and manipulate directly using Java. So in order to ease the processing of Euclidean method comparison first Excel file is imported into SQL database and then processed there by creating other temporary files in SQL. After all the calculations and generation of temporary files, Final clustered Excel files are created using Apache POI. So SQL is used as the middle agent to manipulate and generate Excel files.

Case Example with screenshots:

1. **Input file structure:** Excel file imported to MySQL database in the original Structure format as in provided Excel file. The Structure of input file is shown below:



The screenshot shows the MySQL Workbench interface. At the top, there's a 'Query 1' tab with a SQL query: `select * from example_xls.sheet1;`. Below the query editor, the 'Result Grid' is displayed, showing a table with 4 columns: D, R1, S1, and SCF. The table contains 6 rows of data. The interface also includes a toolbar with various icons for file operations, a 'Filter Rows' field, and an 'Export' button. The status bar at the bottom indicates 'sheet13' and 'Read Only'.

	D	R1	S1	SCF
1	2	3	4	
5	6	7	8	
9	10	11	12	
13	14	15	16	
17	18	19	20	

Fig. 9.1: Excel file data imported to MySQL database

2. **File structure after addition of Cluster label:** Cluster centre is computed with the nearest distance among all others to the respective row of data file and then the label of that cluster

centre is added to the corresponding row. The structure of the file after adding Cluster label column is shown below:

The screenshot shows a query editor window titled 'Query 1'. The SQL query is `select * from example_xls.sheet2;`. Below the query editor is a 'Result Grid' showing data from 'example_xls.sheet2'. The grid has five columns: 'D', 'R1', 'S1', 'SCF', and 'ClusterLabel'. The data is as follows:

D	R1	S1	SCF	ClusterLabel
1	2	3	4	1
5	6	7	8	1
9	10	11	12	3
13	14	15	16	3
17	18	19	20	3

The 'ClusterLabel' column contains values 1, 1, 3, 3, and 3, which correspond to the nearest cluster centre label for each row. The interface also includes a 'Filter Rows' section and an 'Export' button.

Fig. 9.2: Cluster label column is inserted corresponding to nearest cluster centre label

Chapter 10

Execution Flow of the Project

10.1. Introduction

The process of proposed work incorporates the task to import Excel file to SQL database, processing data from SQL database by comparing respective distances from each cluster centre using Euclidean method in manual mode or using K-means and then applying regression on each cluster file separately to calculate errors and finally concludes by generating output as a K value against which minimum error is calculated.

10.2. MS-Excel File Input to the Project

To execute program, run the main.java file in compatible IDE. “Enter the path of Excel File” will prompt on screen. Then provide the input Excel file path. Then after “Enter the count of cluster centres (K)” will prompt.

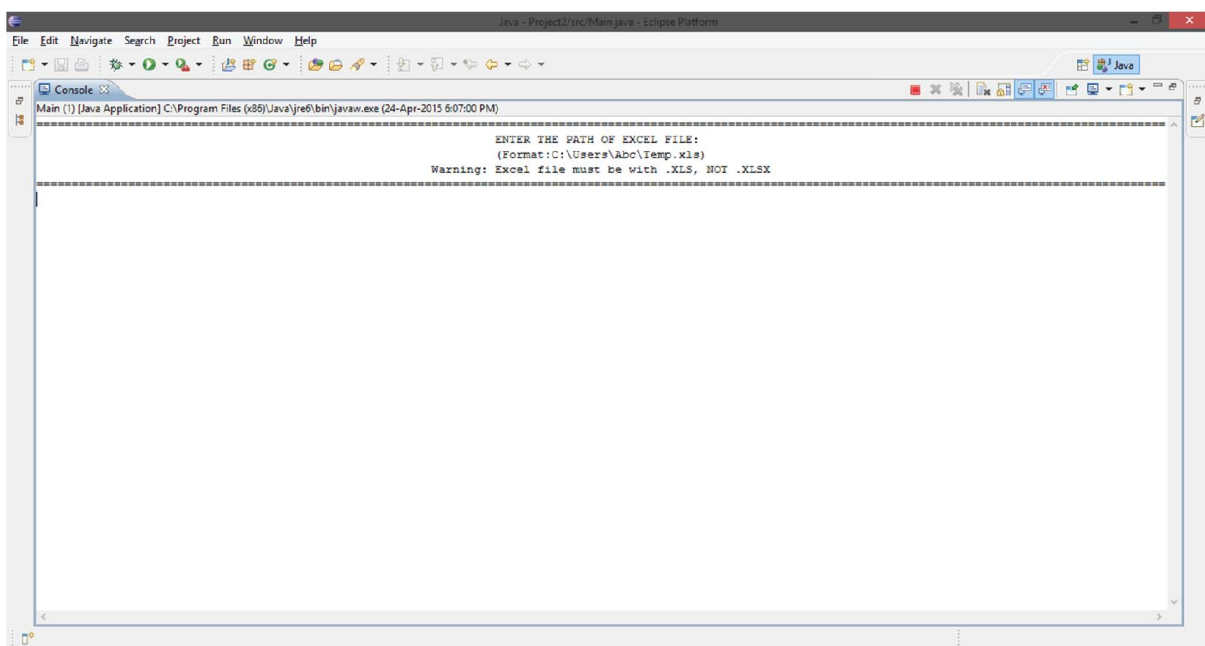


Fig No. 10.1: Prompt for Excel file path

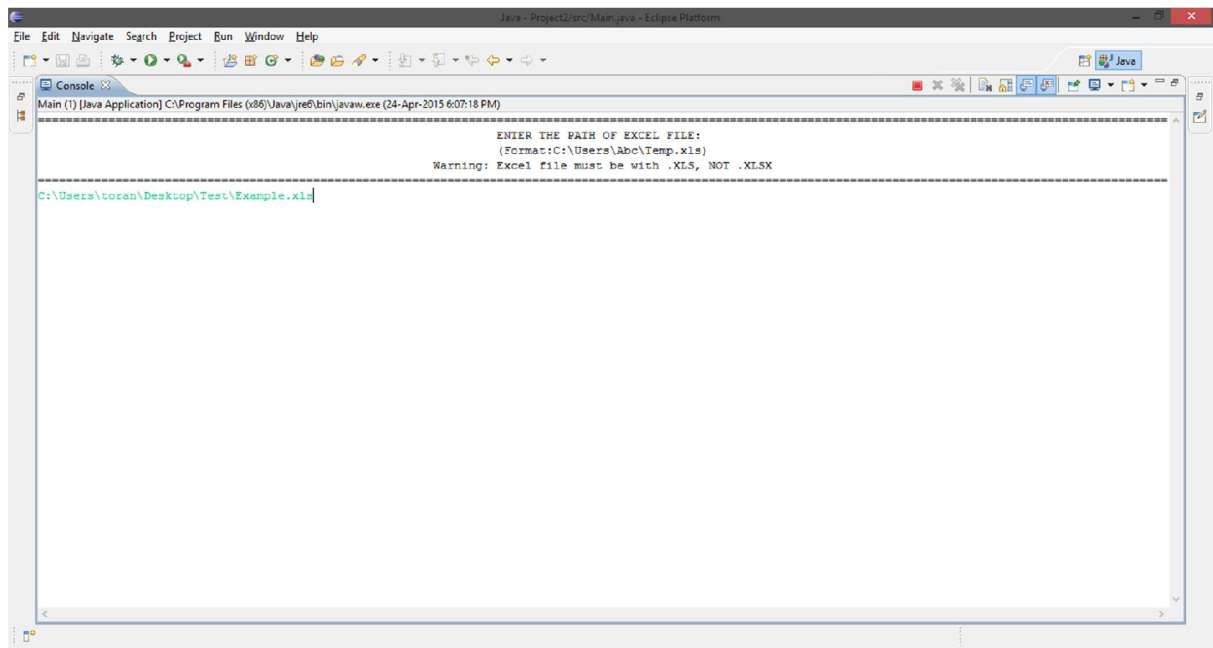


Fig No. 10.2: Excel file path input

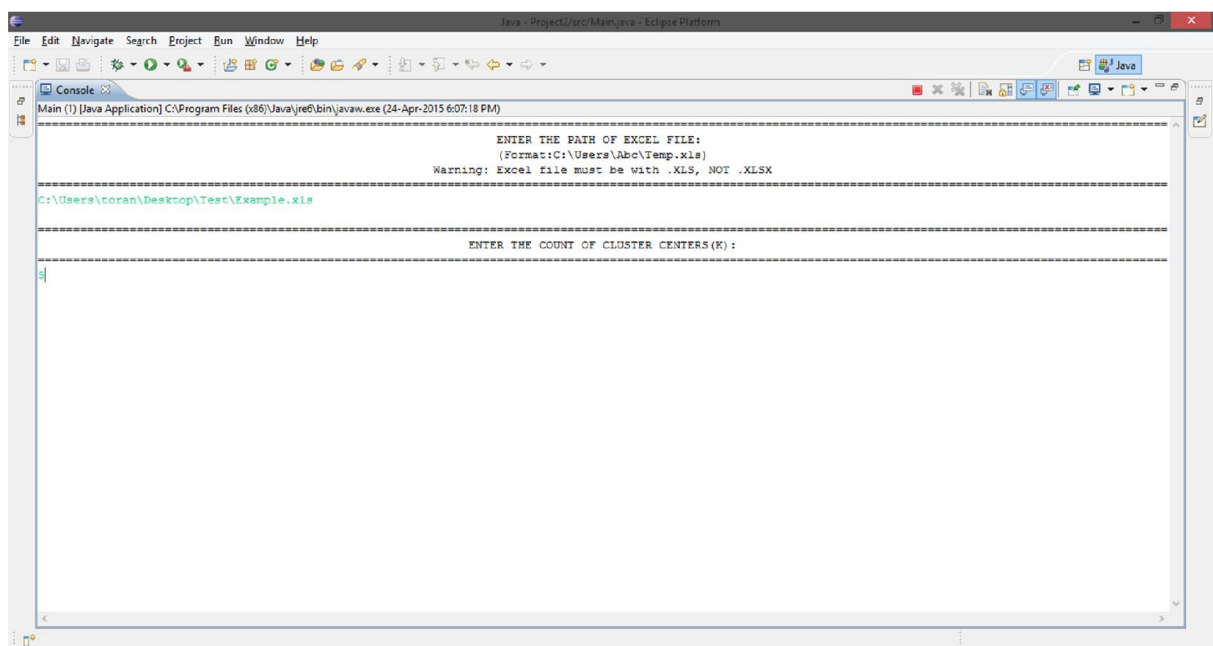


Fig No. 10.3: Cluster count (K) input

10.3. Choice of Clustering

Two choices are made available for the user to cluster the dataset. First choice is K-Means, which implements Weka API's inbuilt functionality to implement the algorithm. The second choice is manual, under which user has to avail the cluster centres around which clustering will be done using Euclidean distance comparison.

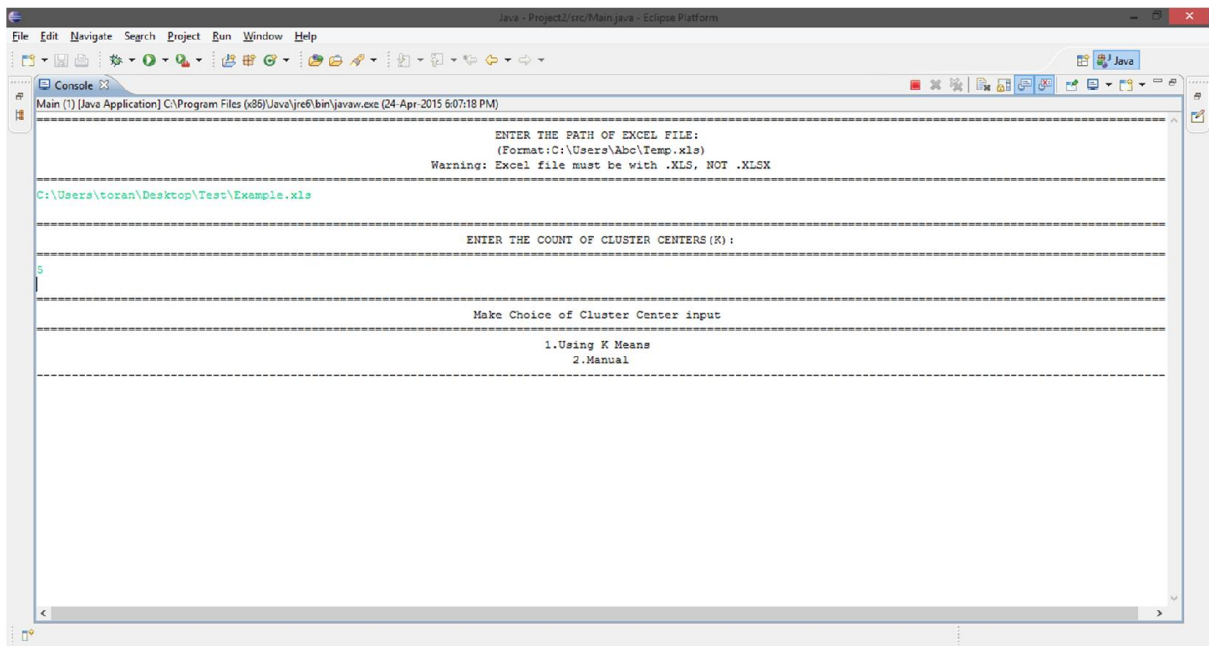


Fig No. 10.4: Selection of Mode of clustering

10.3.1. K-Means

On selecting the first choice i.e. K-Means, the program executes Weka API's method to cluster the dataset and then regression methods are applied to generate the equations which is followed by the error estimation and error comparison.

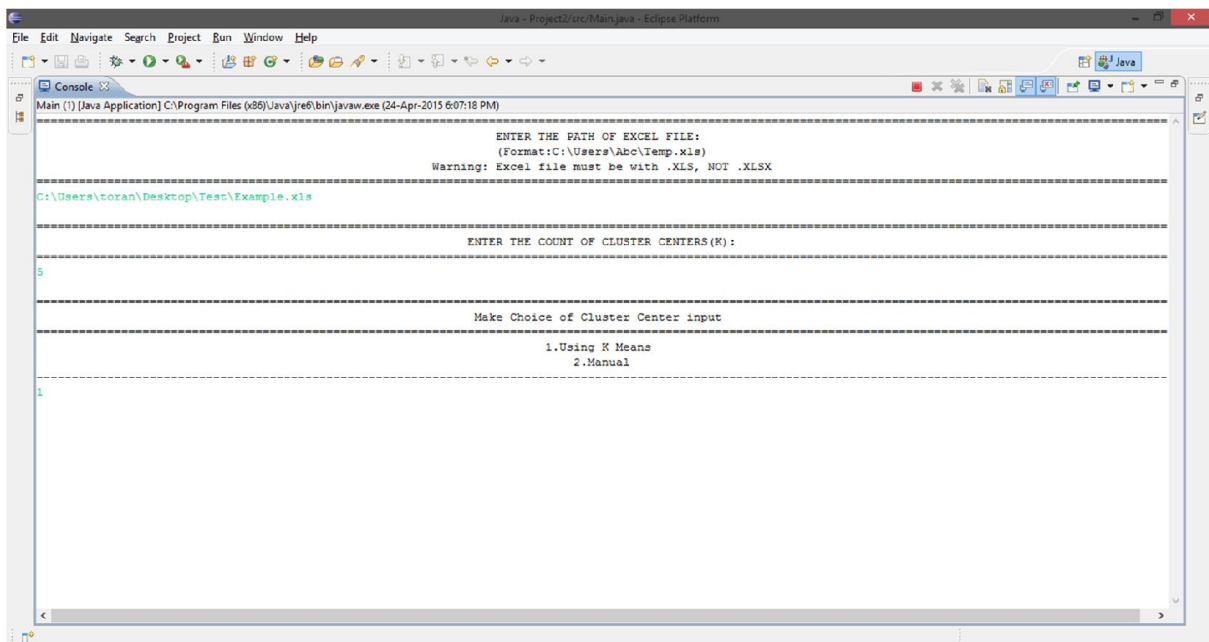


Fig No. 10.5: K-Means is selected for clustering

10.3.2. Manual Cluster Centre Input

1. After providing path and value of K, “Make choice of cluster centre input” will prompt on terminal. Manual Cluster Centre input method is selected

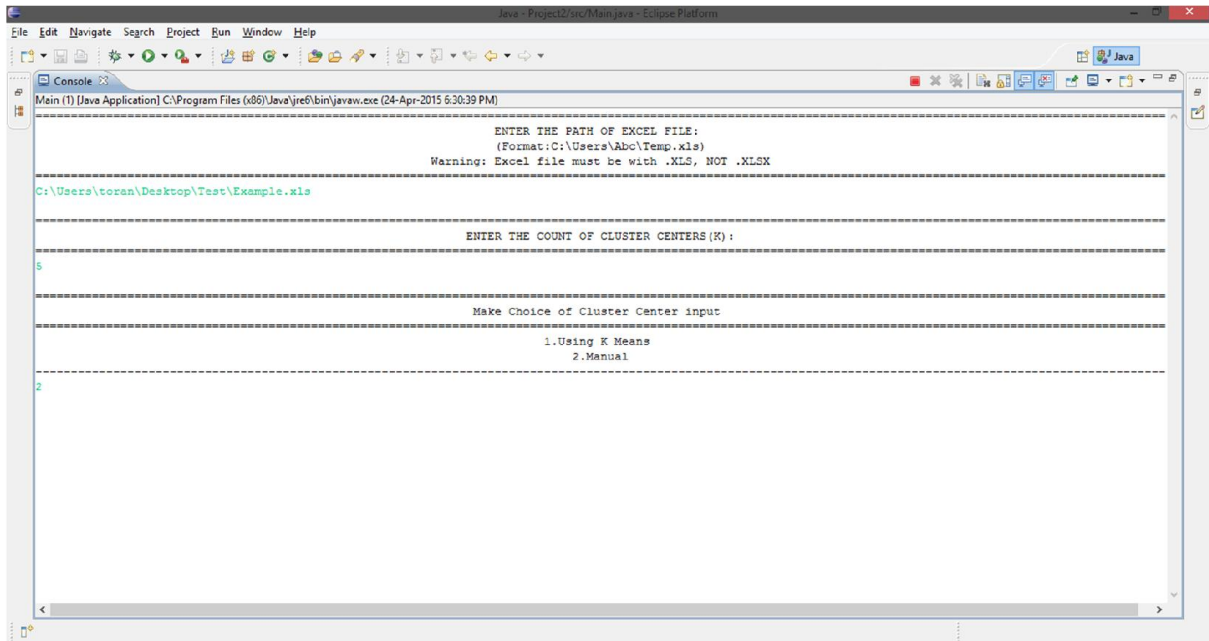


Fig. 10.6: Manual input for Cluster Centroids

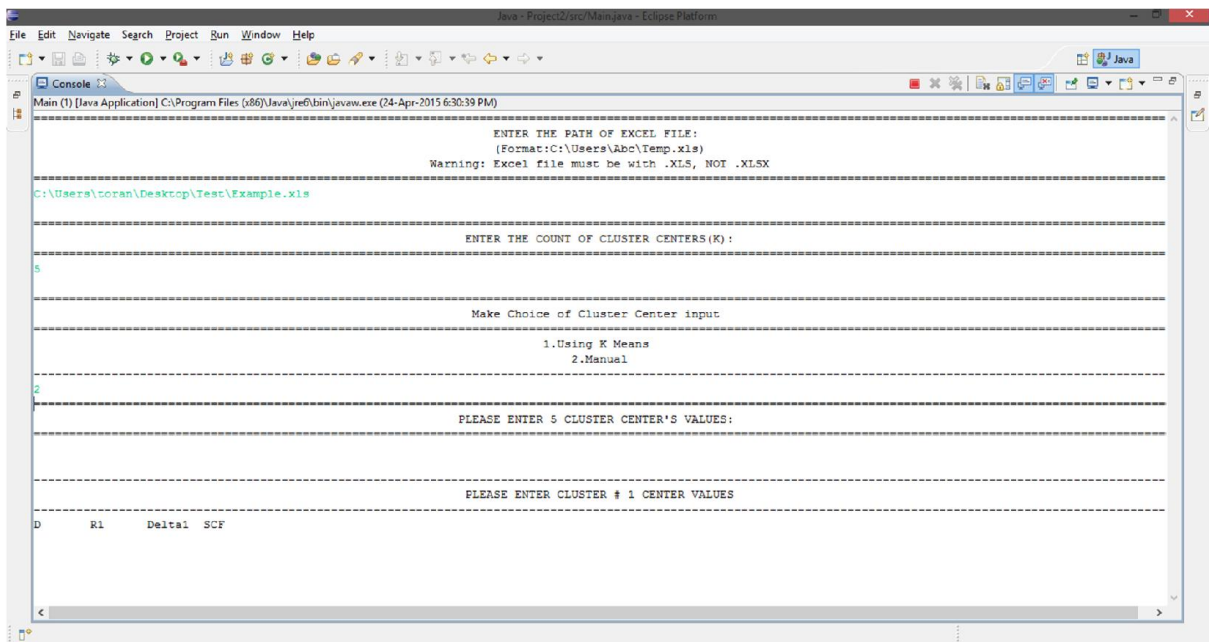


Fig. 10.7: Prompt for Cluster # 1 Centre Values

2. Enter centre #1 values.

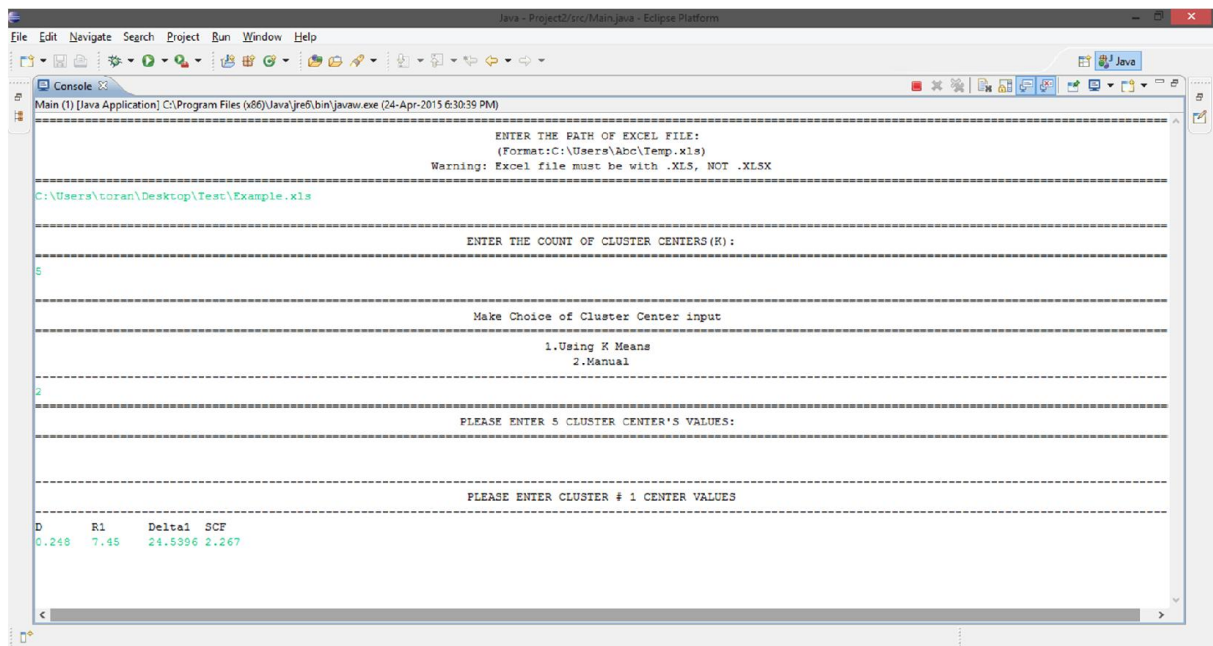


Fig. 10.8: Input for Cluster # 1 Centre Values

3. Similarly entering other cluster centre values for until cluster count

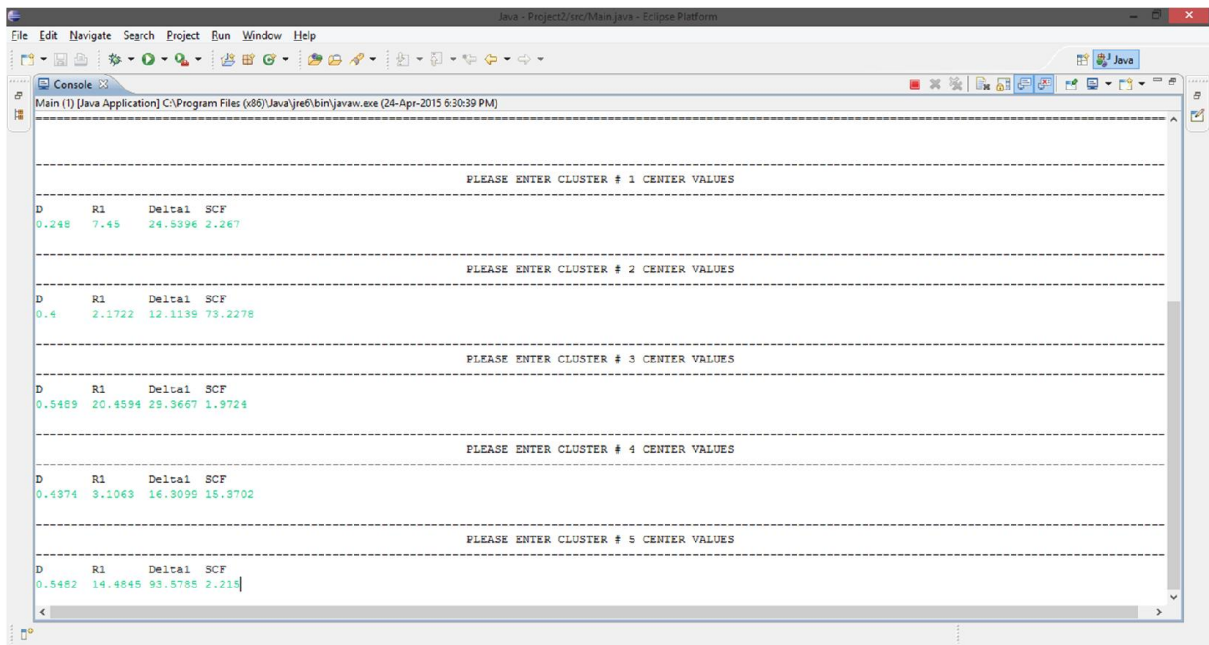


Fig. 10.9: Input K Cluster Centre Values

10.4. Generating Clustered Excel File and Count File

Program will generate Excel files and the “count.txt” file to show the completeness and consistency of the generated files in the original directory from which the dataset was selected if the steps above this are executed without any error:

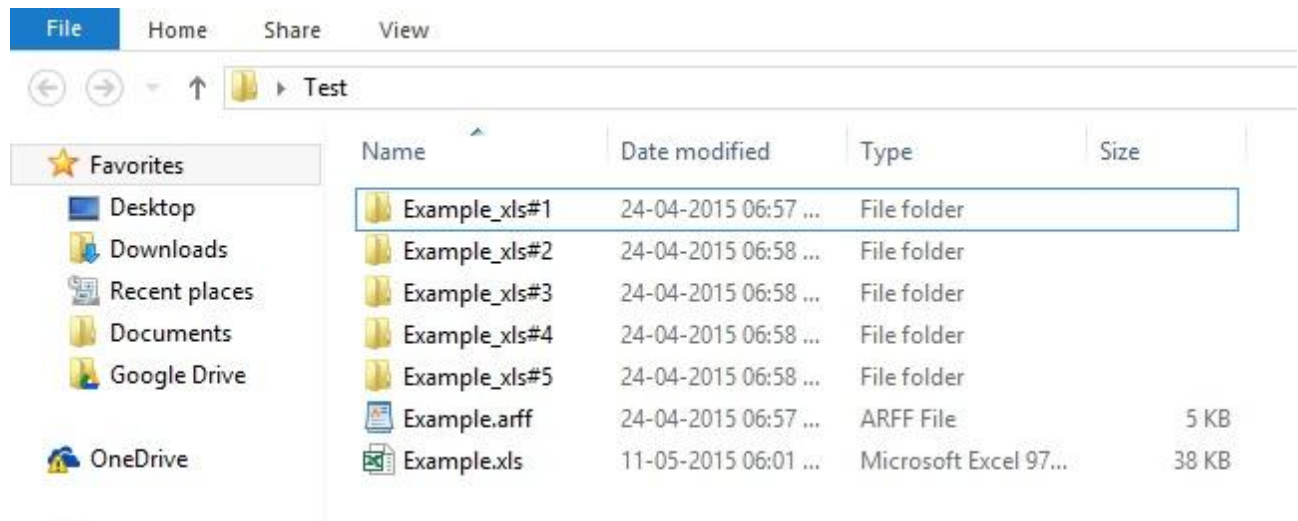


Fig. 10.10: .Directory tree generated for K cluster count

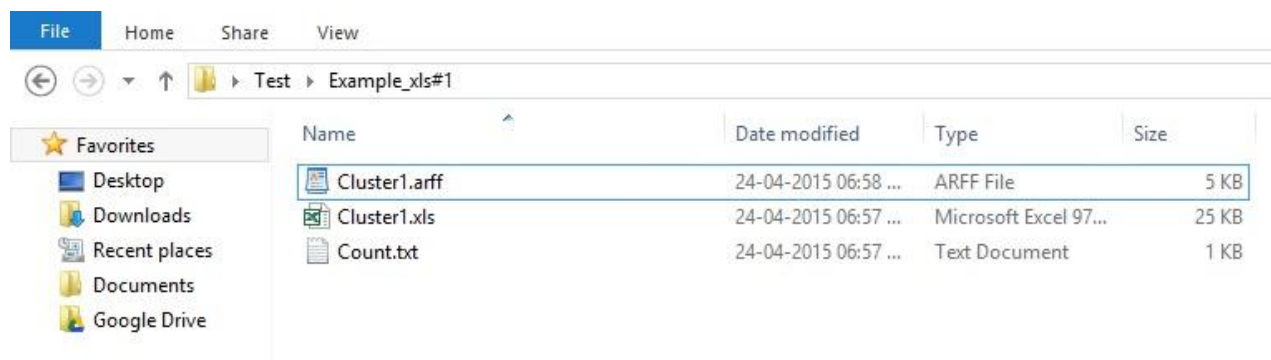


Fig. 10.11. i cluster files generated in i^{th} directory

```
Cluster1 Count #--->38      Count %--->16.30901287553648%
Cluster2 Count #--->91      Count %--->39.05579399141631%
Cluster3 Count #--->54      Count %--->23.175965665236053%
Cluster4 Count #--->25      Count %--->10.729613733905579%
Cluster5 Count #--->25      Count %--->10.729613733905579%
```

Fig. 10.12. Completeness and consistency of clustering shown in Count.txt (for K=5)

10.5. Clustered Excel file to ARFF file

To implement the Weka API excel dataset is converted into .arff file format using connection to Excel file and writing new file with the same name but .arff extension with specific format.

```
@relation Example.arff
@ATTRIBUTE D numeric
@ATTRIBUTE R1 numeric
@ATTRIBUTE Delta1 numeric
@ATTRIBUTE SCF numeric
@data
0.3,9,3.75,2.2113
0.3,9,5.25,2.1966
0.3,9,7.5,2.1812
0.3,9,9,2.1742
0.3,9,10.95,2.1679
0.3,9,12,2.1658
0.3,9,15,2.163
0.3,9,18.75,2.1644
0.3,9,22.5,2.1651
0.3,9,26.25,2.1665
0.3,9,27.75,2.1679
0.3,9,30,2.184
0.3,9,33.75,2.2134
0.3,10.5,3.75,2.1217
0.3,10.5,5.25,2.1077
0.3,10.5,7.5,2.0951
0.3,10.5,9,2.0902
0.3,10.5,10.95,2.086
0.3,10.5,12,2.0853
0.3,10.5,15,2.0867
0.3,10.5,18.75,2.0937
0.3,10.5,22.5,2.107
0.3,10.5,26.25,2.1266
0.3,10.5,27.75,2.1287
```

Fig. 10.13. Snap of generated .arff file from Example.Xls

10.6. Regression

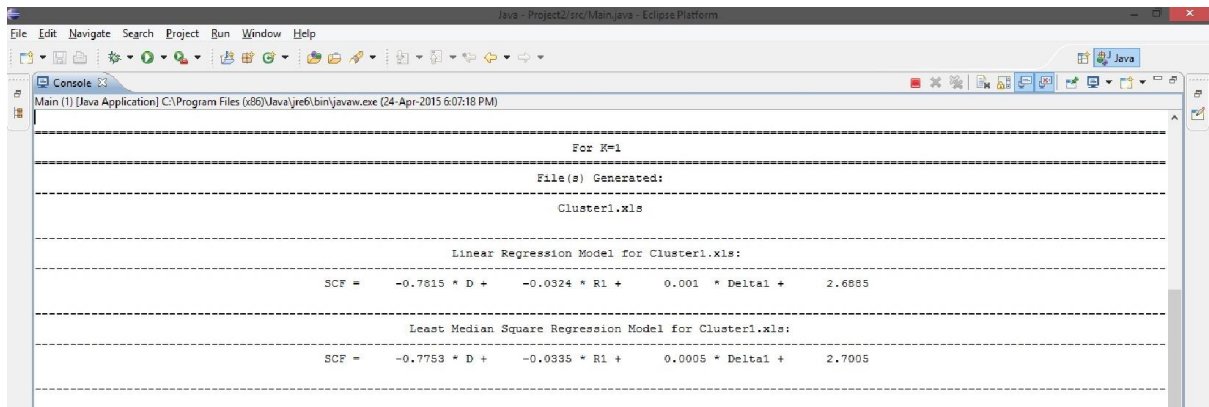
Regression methods are applied to the .arff file generated.

10.6.1. Linear Regression

Linear Regression is applied to the generated .arff file to get the regression equation. And then this equation is used to calculate the errors.

10.6.2. LMS Regression

Least Median of Squares Regression is applied to the generated .arff file to get the regression equation. And then this equation is used to calculate the errors.



```
For K=1
File(s) Generated:
Cluster1.xls

Linear Regression Model for Cluster1.xls:
SCF = -0.7815 * D + -0.0324 * R1 + 0.001 * Delta + 2.6885

Least Median Square Regression Model for Cluster1.xls:
SCF = -0.7753 * D + -0.0335 * R1 + 0.0005 * Delta + 2.7005
```

Fig. 10.14. Regression Equations generated for K=1

Similarly, equations are generated for each cluster files as shown above.

10.7. Error Estimation

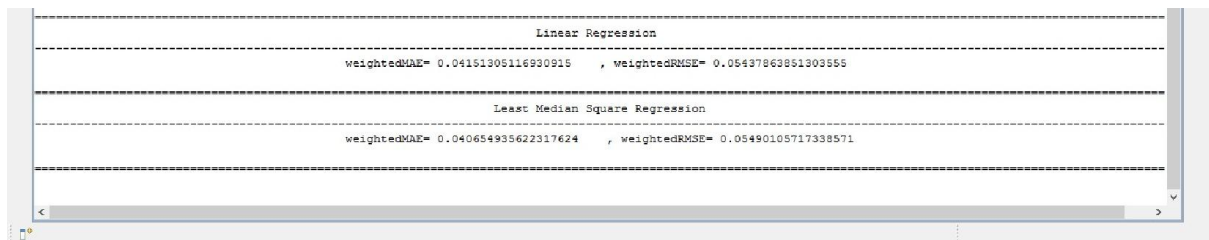
To measure the performance of the particular cluster count, the average of the residual is calculated in following two forms.

10.7.1. MAE

Mean absolute error is calculated to estimate the performance of the regression clustering for each cluster file.

10.7.2. RMSE

Mean absolute error is calculated to estimate the performance of the regression clustering for each cluster file.



```
Linear Regression
weightedMAE= 0.04161305116930915 , weightedRMSE= 0.05437863851303555

Least Median Square Regression
weightedMAE= 0.040654935622317624 , weightedRMSE= 0.05490105717338571
```

Fig. 10.15. Weighted MAE and Weighted RMSE Calculated for K=1

Similarly, Error is calculated for each value of K (cluster count).

10.8. Error Comparison

Error comparison is done on the basis of calculated Weighted MAE and Weighted RMSE over each Cluster Count value and then the cluster count against which minimum error is found is printed as the output of the project.

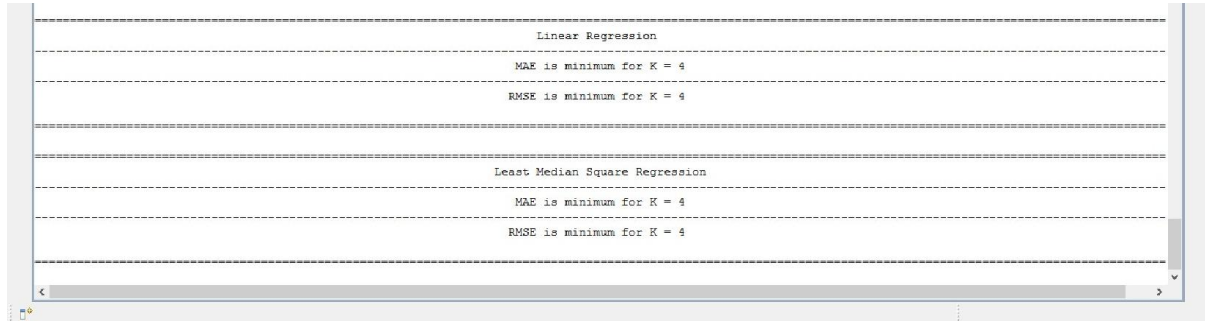


Fig. 10.16. Weighted MAE and Weighted RMSE are compared for all clusters

Chapter 11

Conclusion

In the proposed work three major steps, clustering, applying regressions, and performance measurement are carried out. K-Means clustering technique along with Linear Regression (LR) and Least Median of Squares (LMS) regression technique are applied on MS Excel file for predicting dependent variable value. The separate clustered MS Excel files are generated which hold consistency and completeness with respect to original MS Excel File. The performance is evaluated by estimating corresponding errors using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), error estimation techniques.

The Java powered program employed here uses Ms Excel file for data input and MySQL workbench for data processing, which helps in fragmentation of the data into different clusters for given cluster centres, and followed by regression on clustered excel file. Errors are estimated per cluster file and compared to find out the cluster value for which error is minimum.

Chapter 12

Future Work

In future the whole project can be implemented using GUI for perspective of easiness. Also excel file writing can be implemented using some techniques other than the Apache POI. The clustered data can be used to plot graph from them to analyse. Partitioning based clustering algorithm K-Means is used in the proposed work, density based clustering algorithms can be explored for creating the groups of the similar records from MS-Excel data as future scope of the work. Many more Regression techniques can be implemented for further optimization in the result.

Reference

- [1]. J. Han and M. Kamber, “Data Mining Tools and Techniques”, Morgan Kaufmann Publishers.
 - [2]. Database system concept, Korth & Sudarshan, MH.
 - [3]. J. A. Hartigan and M. A. Wong, “Algorithm AS 136: a K-Means clustering algorithm,” Journal of the Royal Statistical Society C: Applied Statistics, vol. 28, no. 1, pp. 100–108, 1979.
 - [4]. N. K. Nagwani and S. V. Deo, “Estimating the Concrete Compressive Strength Using Hard Clustering and Fuzzy Clustering Based Regression Techniques”, pp.4-9, 2014.
 - [5]. B. Zhang, “Regression clustering,” in Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM '03), pp. 451–458, Melbourne, Fla, USA, November 2003.
 - [6]. P. J. Rousseeuw, “Least median of squares regression,” Journal of the American Statistical Association, vol. 79, no. 388, pp. 871–880, 1984.
 - [7]. L. Kaufman, P. J. Rousseeuw, "Finding Groups in Data-An Introduction to Cluster Analysis," Wiley Series in Probability and Mathematical Statistics, 1990.
 - [8]. G. Holmes, A. Donkin, and I. H. Witten, “Weka: a machine learning workbench,” in Proceedings of the 2nd Australian and New Zealand Conference on Intelligent Information Systems, pp. 357–361, December 1994.
-