Comparative analysis of clustering algorithms for the study of home loan applicants using WEKA tool.

Mrs.Yogita Bhapkar
Department of Computer Science,
BVDU Yashwantrao Mohite College,Pune, India
yvbhapkar@gmail.com

Dr.Ajit More
MCA Director
BVDU IMED Pune, India
ajit.more@bharatividyapeeth.edu

Abstract— *Data Mining is the process to pull out patterns from large dataset. It purely based on statistics and artificial intelligence. Data mining has performed great contribution in knowledge discovery in the industries. It is upcoming research area. Data mining has variety of applications in every field including science, information technology, biology, medical science, Banking and many more. This research paper is restricted to only bank application. Predicting loan defaulters is at the core application in every bank which offers loan to the customers. Against this background classification and clustering these data mining techniques can used to check loan defaulters. various classifiers has used to classify loan applicants into classes like those who are safe and hence get loan from bank and those who are risky so can't get loan. Clustering is also data mining technique of partitioning given set of objects into disjoint clusters. Cluster analysis is one of the major data analysis methods and can use in bank for analyzing loan defaulters. Also it supports different categories and several methods. This study mainly focuses on study of loan applicants by using clustering approach. Here performance of K-means algorithm, hierarchical clustering algorithm and EM i.e. expectation maximization algorithm is measured. Data mining tool used here is WEKA performances of algorithms are compared on the basis of accuracy and running time.*

Keywords— Data mining-means clustering, Hierarchical clustering, EM, WEKA.

I. Introduction

Day by day there is expensive growth of data. Data plays important role for every business organizations, including private, public, and government organizations. These organizations are investing in building data warehouses that contain millions of records but they are not getting good returns. Basically data warehouse is a repository of huge amount of data. This is stored under unified schema. But hence organizations have rich data but poor knowledge. Due to lack of knowledge they can not produce sufficient output. Hence they need database analysis technique to extract useful information from data and thus help for analysis and decision making. Data mining is the process to extract knowledge and pull out patterns from large datasets. It is also data analysis technique which helps to discover previously unknown information from data. Data mining and knowledge discovery in database have attracting a significant amount of research in all the fields.

Today data mining is becoming popular research area. Data mining supports set of techniques that can be used to extract relevant and interesting knowledge from data. Data mining techniques such ass association, rule mining, classification, clustering and prediction.

As we know that banking industries have important role in the economy .data mining help the banking industries to deal with upcoming challenges in the economical conditions. To provide loan to the customers is core application.

Clustering is a data mining technique use for making group of abstract objects into classes of similar objects are grouped into one cluster and dissimilar objects are grouped in another cluster. Clustering is a data mining technique that have been attracting a significant amount of research in science, information technology field, medical science, image processing, document classification ,clustering analysis is also used in banking industries also. In this research paper we worked on fundamental approach of clustering is that of grouping similar data together.

II. Requirements of Clustering

1. Finds natural grouping of instances.
2. It deals with different types of attributes.
3. Minimal requirements for domain knowledge
4. It deals with outliers also.
5. High interpretability and usability.
6. Clusters are generated by arbitrary shapes.

III. Clustering Methods

1. Partitioning method: constructs various partitions of database objects.
2. Hierarchical method: it creates hierarchical decomposition of the set of data.
3. Density based method: it is based on connecting and density functions.
4. Grid based method: based on multiple level granularity structure.
5. Model based method: a model is hypothesized for each of the clusters and idea is to find best fit of that model to each other.

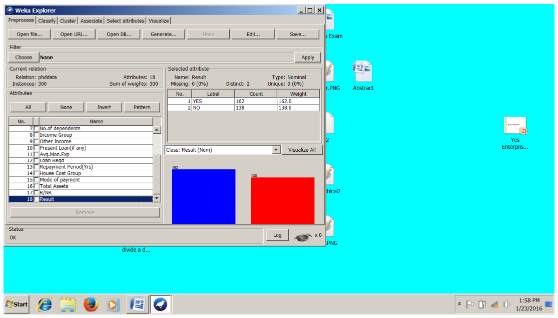
Many algorithms exist for clustering. This study mainly focus on K-means algorithm which comes under partitioning clustering, Hierarchical clustering algorithm which is of type agglomerative and divisive and EM algorithm which is of model based clustering.

IV. WEKA TOOL

In this research paper we used WEKA data mining tool. WEKA aims to provide a comprehensive collection of machine learning algorithms and data preprocessing tools to researchers. It provides variety of algorithms for data mining and machine learning. It is open source and freely available. It is platform-independent. It is easily useable and easily understandable tool.

Data file used for this research is bank database containing customer’s data which are applying for home loan. Data file “bank.arff” and includes 300 instances. As an illustration of performing clustering in WEKA, we used implementation of clustering algorithms to cluster the customers in this bank data set, and to characterize the resulting customer segments.

Following Figure shows the main WEKA Explorer interface with the bank data file when loaded.



I) K-means clustering :

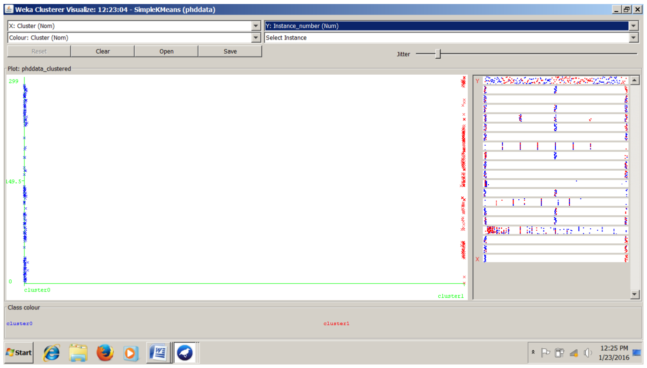
K-means is the most common partitioned clustering algorithm. Which is used to partition n observations into K clusters in which each observation belong to cluster with nearest mean. It is simple, non supervised iterative learning method. The idea behind classifying set of data objects into K number of clusters where K is fixed initially.

This algorithm is easily implemented and have less execution time. It first fix initial group centroids . Then assign each object to the group that has closest centroid. Once all the objects are assigned it recalculate positions of centroids. Again repeat the same process until centroids not change.

Following are the results of K-means algorithm when applied on bank database:

```
Simple K-Means Algorithm
Number of Clusters: 2
Number of iterations: 5
Clustered Instances

0      111 ( 37%)
1      189 ( 63%)
```



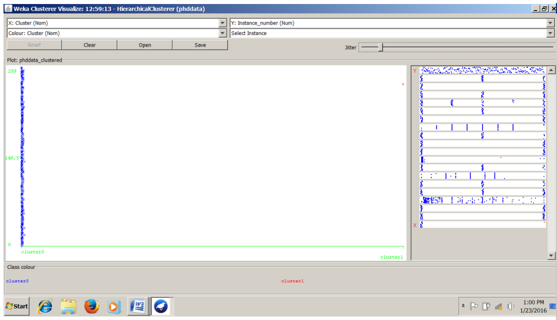
II) Hierarchical clustering:

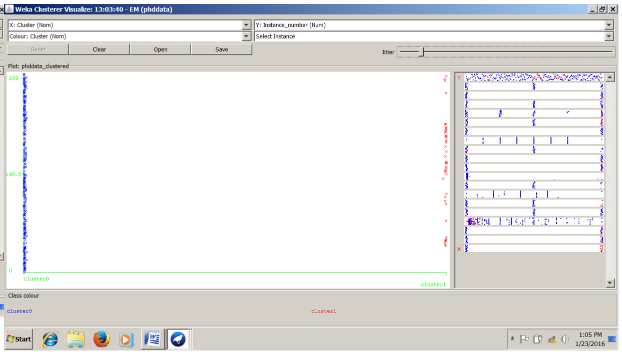
Here data is not partitioned into clusters in a single step, but it takes series of partitions which seeks to build a hierarchy of clusters. In this paper we use agglomerative approach for hierarchical clustering which start with assigning data items into own clusters it means if we have N items then we have N clusters, each of the cluster contains one item only. Then it finds the most similar pair of clusters and merges them into a single cluster. Then it computes similarities between new cluster and each of the old Clusters. Repeat this until we will get a single cluster with all items.

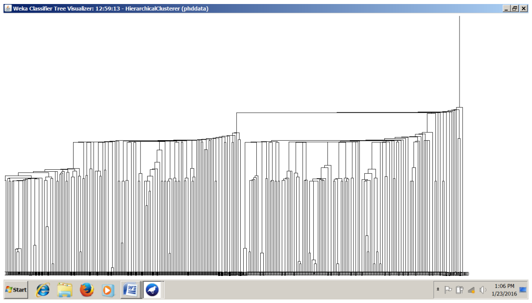
Following is the result of Hierarchical clustering:

```
Number of Clusters: 2
Clusters Instances
   0      299 (100%)

    1          1 (  0%)
```






Weka Classifier Tree Visualizer: 12:59:13 - HierarchicalClusterer (phddata)
Start
1:06 PM
1/23/2016

V. EXPECTATION MAXIMIZATION ALGORITHM:

This is extension to K-means algorithm, which is based on maximum likely hood concept. Here data set is usually model with a fixed number of Gaussian distributions that are initialized randomly and whose parameters are iteratively optimized to fit better to the data set. EM algorithm maximizes the likelihood function.EM algorithm assign data to cluster with some probability. It gives probability model for distributions. It can calculate cluster membership probability for any instance.

Following is the result of EM algorithm:

Number of Clusters: 2

Clusters Instances

```
0      253 ( 84%)
1       47 ( 16%)
Log likelihood: -41.79124
```

VI. CONCLUSION

In this paper we introduce the concept of data mining. This study is limited to data mining technique clustering. Weka is the data mining tool used for performing experiments on bank database. The main aim of this paper is to provide detailed introduction about clustering algorithms. From all available clustering algorithms we selected K-means algorithm, Hierarchical algorithm and EM algorithm. In this study we found that K-means is simplest algorithm as compared to others hence it is widely used for clustering. This method minimizes cluster squared error. It is fast and easier to understand. It is relatively efficient. But it is not able to handle noisy data and outliers.

Hierarchical clustering results are usually presented in the form of dendrogram. Representation of dendrogram is easy to understand for small datasets but it really hard to understand in case of large datasets.

EM algorithm work we well in case of overlapping data. Time taken to build model using K-means clustering is less than clustering method. This work can be concluding by stating that K-means algorithm gives good performance for the database used for this study.

VII. ACKNOWLEDGEMENT

Many thanks to all researchers who contributed their work in this area . I wish to express my sincere thanks to Dr.Ajeet More for his constant encouragement and valuable guidance. I am also thankful to principal Dr.K.D.Jadhav and all faculty members of computer science department. Finally thanks to my family members for their support.

VIII. References

1) TAHAIRE Nu Phyu, “Survey of classification techniques in data mining” IMECS,2009

2) CBIT-IIITB Working paper ,”Default prediction in banking loans through data mining,2006-11

3) Andre Carlos Ponce de Leon Ferreira de Carvalho,”Credit risk assessment and data mining”,2009

4) Choosing the right data mining technique:classification of methods and intelligent recommendation,iEMSs 2010 by Karina Gibert

5) Credit risk analysis using a hybrid data mining model,2007by S. Kotsaiantis

6) Improving the accuracy and efficiency of the k-means clustering algorithm ,WCE 2009 by K.A.Abdul Nazeer.

7) Performance evaluation of k-means and heirarichal clustering in terms of accuracy and running time,IJCSIT 2012 by Nidhi Singh.

8) Assessing loan riskys:A data mining case study,1999 IEEE

9) A Proposed classification of data mining techniques in credit scoring, Jan 2011

A survey of credit & behavioral scoring International journal of forecastiva ,2000

10) Kazi Imran Moin,Dr.Qazi Baseer Ahmed,”Use of data mining in banking”IJERA ISSN:2248-9622,Mar-Apr 2012

11) Data Mining Techniques, http://www.dataminingtechniques.net

12) Vivek Bhambri,”Application of Data mining in banking sector”,IJCST Vol.2,Issue 2,June2011

13) Abbas Keramati,Niloofar Yousefi ,”Proposed classification of data mining techniques in credit scoring,ICIEOM,Jan 2011

14) Decision tree analysis on J48 algorithm for data mining,IJARCSSE,2013

15) Data Mining Techniques for Marketing, Sales, and Support Michael J.A. Berry and Gordon Linoff, Wiley

16) Data Mining: Concep ts and Techniques Jiawei Han and Micheline Kamber, Morgan Kaufmann

17) Principles of Data Mining Paperback – 2004

18) by Hand David (Author), Mannila Heikki (Author), Smyth Padhraic

19) K. Chitra, B.Subashini, Customer Retention in Banking Sector using Predictive Data Mining Technique, International Conference on Information Technology, Alzaytoonah University, Amman, Jordan, www.zuj.edu.jo/conferences/icit11/paperlist/Papers/

20) [2] K. Chitra, B.Subashini, Automatic Credit Approval using Classification Method, International Journal of Scientific & Engineering Research (IJSER), Volume 4, Issue 7, July-2013 2027 ISSN 2229-5518.

21) [3] K. Chitra, B.Subashini, Fraud Detection in the Banking Sector, Proceedings of National Level Seminar on Globalization and its Emerging Trends, December 2012.

22) [4] K. Chitra, B.Subashini, An Efficient Algorithm for Detecting Credit Card Frauds, Proceedings of State Level Seminar on Emerging Trends in Banking Industry, March 2013.

23) [5] Petra Hunziker, Andreas Maier, Alex Nippe, Markus Tresch, Douglas Weers, and Peter Zemp, Data Mining at a major bank: Lessons from a large marketing application http://homepage.sunrise.ch/homepage/pzemp/info/pkdd98.pdf