Data Mining Techniques for Rental Information Management using WEKA

[1]Tansen Patel,[2]Anurag Chandra,[3]Nageshwar Verma
[1]Asst. Professor, [2]Research Scholar,[3]Research Scholar
[1]Computer Science and Engineering,
[1]Shri Shankaracharya Institute of Professional Management and Technology, Raipur, India

Abstract : Data mining is the process in which discover patterns in large datasets at the intersection of statistics, database systems and machine learning. There are various techniques used in data mining for mining the datasets as per their application areas. This research paper use Canopy technique in data mining for mining our dataset which comes under clustering techniques of data mining.WEKA software is used to implement for the result part. The paper suggests the implementation of Canopy Clustering technique of data mining and it checks the result which will be useful for managing and keeping account of every details of lessee in our Rental Information System for the society.

IndexTerms – Data Mining, WEKA, Clustering, Canopy, Association rule.

I. INTRODUCTION

Data Mining is the process where we discover patterns from large sets of data involving methods at intersection of statistics, machine learning and database systems.Data mining is a series of processes which takes input in form of data and outputs as knowledge. One of the earliest and most prominent definitions of the data mining process, which features some of its unique and distinctive characteristics, is provided by Fayyad, Piatetsky -Shapiro and Smyth (1996), who were its developer too defined it as "the non-trivial process of identifying novel, valid, potentially useful, and ultimately understandable patterns in data."The term Data Mining originally denote the algorithmic step in the data mining process, which initially was known as the KDD process which is abbreviated as Knowledge Discovery in Databases. However, with passing of time this name has been changed and data mining depending on the type of data it is used may refer to the entire process or just the algorithmic step [1].

Data mining has various uses. It uses pattern matching and statistical techniques.The data we contain with us is often vast, and noisy, meaning that it is not precise and the structure of data is complex. This is where a purely statistical technique can never be successful, so data mining is its solution. The main areas where it is used are missing value, heterogeneity, size of data, noisy data, static data, relevance, interestingness, algorithm efficiency, dynamic data, sparse data and complexity of data. Data mining has become an efficient tool for analyzing large datasets. The efficient database management systems present have been very important thing for managing a large amount of data and especially for effective retrieval of particular information from a large data wheneverrequired [2].

In Rental Information System, whenever a Lessor rents his house to a lessee, the lessor has to submit all the details provided by the lessee which will be then getting verified by nearby police station.It is necessary that the detail which is provided by the lessor should be real and valid. No type of error should be there in the data provided by the lessor which is given to him by lessee. Also the amount of data is huge and requires analysis so that proper account of that information can be kept by police officials. So, here in this case, the concept of data mining comes to existence where the data is to be verified and stored by using the data mining so that no error should be there in the data.There are various data mining techniques available with us through which we can mine data. Some of them are Classification analysis, Association rule learning, Anomaly or Outlier detection, Clustering analysis, Regression analysis, Decision trees, Combinations, Sequential patterns, Long-term(memory) processing etc.

This paper is organized as follows: Chapter 2 discusses different data mining techniques, Chapter 3 describes the methodology, and in Chapter 4, result and discussion is described. Chapter 5 concludes the paper.

II. DATA MINING TECHNIQUES

There are various types in which data mining is classified. Some of them are being explained as follows:-

2.1 Association Rule Mining

Association rule mining is one of the most active data mining methods. This rule was originally proposed for market basket analysis issues. Its purpose is to find a connection between commodities trading rules in different databases. These rules characterize the customer buying behavior patterns that can be used to guide businesses which are scientifically arranged stock inventory and shelf design. After many researching on association rule mining and doing a lot of research, their work involves exploring the theory of association rule mining with the original algorithm improvements and new design algorithm. Association rule mining (Parallel Association Rule Mining) and the number of association rules mining (Quantities Association Rule Mining) and other issues were introduced to improve the efficiency of mining rule algorithm, adaptability, availability and application promotion [3].

2.2 Decision Tree

Decision tree algorithm is one of the most important classification measures in data mining. Decision tree classifier as one type of classifier is a flow- chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node represents a class. The method that a decision tree model is used to classify a record is to find a path that from root to leaf by measuring the attributes test, and the attribute on the leaf is classification result.Decision tree is the main technology used for classification and prediction. Decision tree learning is a typical inductive algorithm based on instance, which focus on classification rules displaying as decision trees inferred from a group of disorder and irregular instance. In top-down recursive way, it compares attributes between internal nodes of decision tree, judges the downward branches according to different attribute of the node, and draws a conclusion

from leaf nodes in the decision tree. So from a root to a leaf node corresponds to a conjunctive rule, and the entire tree corresponds to a group of disjunctive expression rules [4].

2.3 Clustering

Clustering can be defined specifically as the identification of similar classes of objects. This technique can be used for effective means of finding distinct groups or classes of object but it becomes expensive so clustering can be used as preprocessing way for attribute subset selection and classification.By using different clustering techniques we can identify sparse and dense regions in object space and can discover total distribution pattern and relation among the data attributes. For example, to make group of customers based on purchasing pattern, to distinctly distribute genes with similar functionality. There are different types of clustering methods. Some of them are Partitioning Methods, Hierarchical Agglomerative (divisive) methods, Density based methods, Grid-based methods, and Model-based methods [5].

2.4 Classification

Data mining tools have to take a model from the database, and also requires the user to define one or more classes. The database contains one or more attributes that define the class of a row and these are known as predicted attributes whereas the remaining attributes are called predicting attributes. A combination of values for attributes defines a class. Basically the system give a case or tuple with certain known attribute values which will be able to predict to what class this case belongs. When looking at classification rules the system has to find the terms that predict the class from the predicting attributes so firstly the person has to define conditions for each class, the data mine system then make descriptions for the classes. Once classes are referred, the system should define rules that heads the classification therefore the system should be able to find the details of each class [6].

2.5 Regression

Regression is another data mining technique which is based on learning on the basis of supervision and is used to predict a numerical target. It predicts sales, profit, numbers, square footage, temperature rates. All these can be found by using techniques of regression. Regression starts with value of data set known. It is based on training process. It estimates the value by comparing already known and predicted values. Error is the value which is the difference between expected and predicted value. Its motive is to reduce the error so that we can get accurate result [7].

III. METHODOLOGY

3.1 Canopy Clustering Technique

We will verify the tests by using clustering technique. There are various techniques and algorithms in data mining through which we can check the validity of data. Some of them are:

- Hierarchical Clustering
- K means Clustering
- Canopy clustering
- Cobweb clustering

We are going to use Canopy cluster technique to check whether the data inserted by the lessor is valid or not. Canopy clustering is very fast, easy and simple and it can make clusters accurately. This algorithm is most accurately used as pre-processing approach to clustering techniques like K-means. This can decrease computational expenses by starting with initial clustering as it can ignore points which are not in the use of canopies. In canopy clustering, objects are represented as points which is multidimensional. For making the clusters it makes use of two distance thresholds with D1>D2. Algorithm starts with a set of initial points. At each point distance is calculated and grouping decision is applied. When a point is D2 delete it from the cluster. By the end of this technique process, the algorithm specifies a set of canopies.Faster and near to accurate distance measure is used in this type of clustering. Each canopy is a group of objects which are alike. An object may be there in more than one canopy [8].

IV. RESULTS AND DISCUSSION

In this paper, we have used WEKA software which is implemented in java language. It is a open source software. It offers data mining algorithms for data preprocessing, clustering, association rule, classification and Machine learning.

The details of the persons who are present as lessee were found randomly from Chhattisgarh. It includes contact Information {10 digit no.}, Area {City name and area he is residing}, Criminal record {yes, no}, House no.{ any numerical number}, Living Status{ stay, leave}.

In WEKA the whole dataset is in ARFF file format, which consist of different attributes to indicate various things in file. In it, first of all the excel sheet is converted into .csv format and then this file is converted into .arff format. Figure 4.1 shows the sample view of lessee dataset which consists of major attributes and different values.

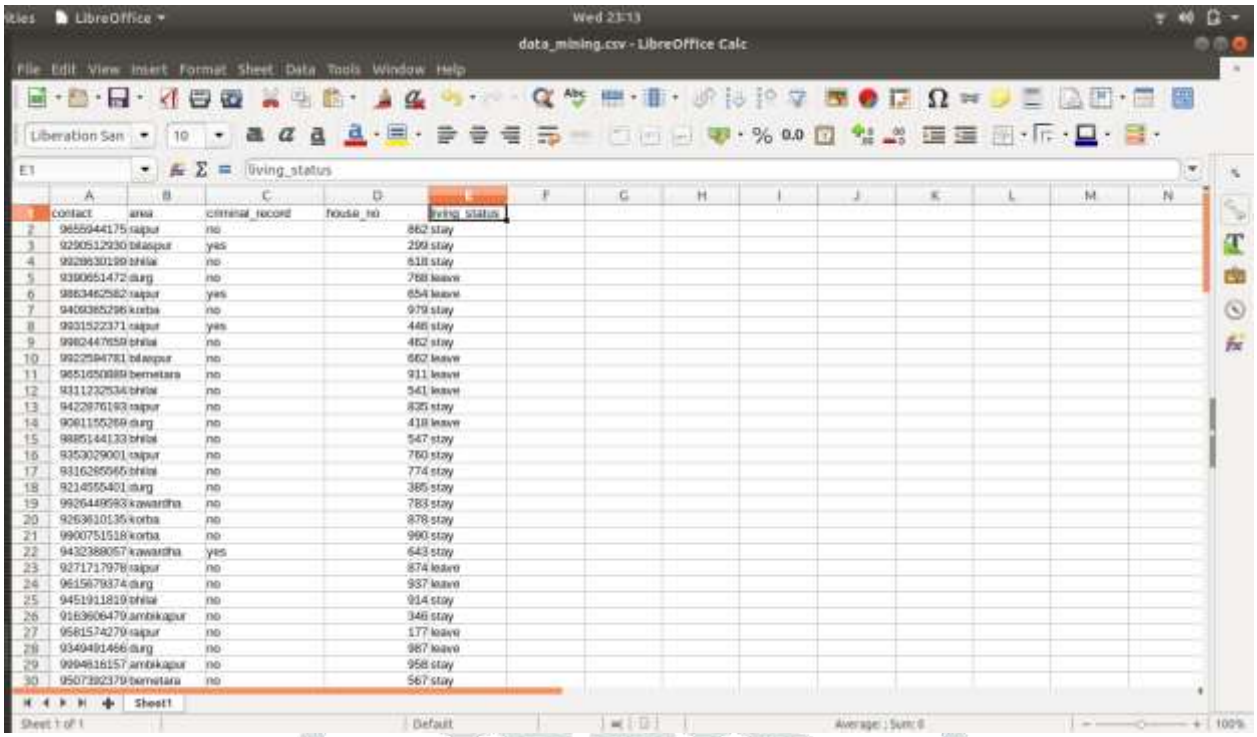

Figure 4.1

In Figure 4.2 the analysis of the lessee dataset is done by using canopy clustering. In it, at first the instances are calculated and the canopies are made. Then the number of canopies are calculated and the T2 and T1 radius are calculated. After it, time taken to build model and then at last the list of unclustered instances is being calculated and then shown.

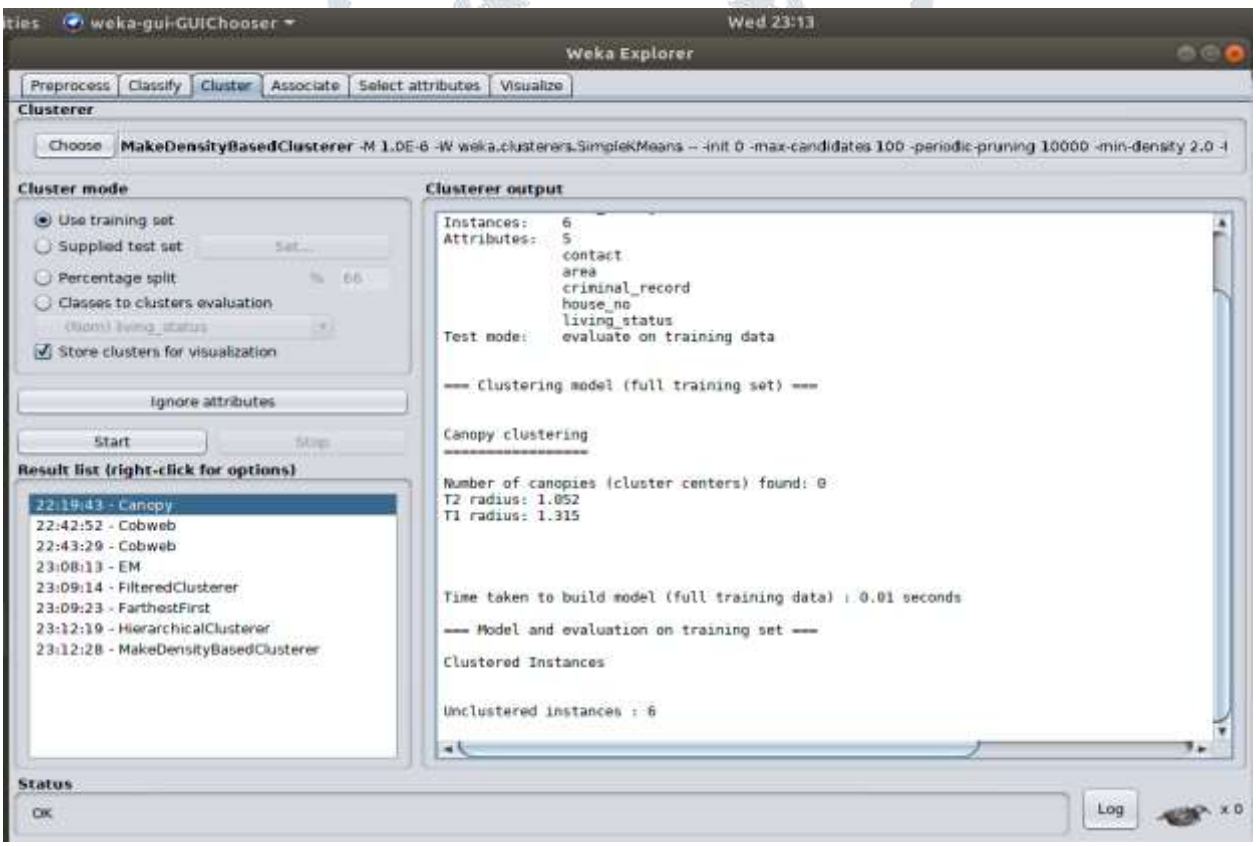

Figure 4.2

V. CONCLUSION

Using the clustering technique in data mining, we have used canopy technique and we found the result as displayed above in figures. We now conclude that the data inserted is valid and the result is shown in figure.This type of clustering is very fast, easy and simple and it can make clusters with precision. This algorithm is mostly used as pre-processing approach to clustering techniques like K-means. We have used this approach in Rental information system and applied it in the details of the lessee present which helped us in obtaining group of information of objects which are alike.

REFERENCES

[1]Gary M. Weiss, Ph.D.,Brian D. Davison, Ph.D.," DATA MINING", To appear in the Handbook of Technology Management, H. Bidgoli (Ed.), John Wiley and Sons, 2010.

[2] Ms. Aruna J. Chamatkar, Dr. P.K. Butey," Importance of Data Mining with Different Types of Data Applications and Challenging Areas", IJERA ISSN : 2248-9622, Vol. 4, Issue 5, May 2014.

[3]Liang Zhao, Deng-Feng Chen, Sheng-Jun Xu and Jun Lu,"The Research of Data Mining Classification Algorithm that Based on SJEP", International Journal of Database Theory and Application Vol.8,, pp. 223-234, 2015.

[4]*Qing-yunDai,Chun-ping Zhang and Hao Wu," Research of Decision Tree Classification Algorithm in Data Mining"* ,International Journal of Database Theory and Application Vol.9,, pp.1-8, No.5, 2016.

[5] M. Ramageri,Mrs. Bharati, " DATA MINING TECHNIQUES AND APPLICATIONS",Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305,ISSN : 0976-5166. 2010.

[6] IS. Sasikala, IIS. NathiraBanu,"Privacy Preserving Data Mining Using Piecewise Vector Quantization (PVQ)", IJARCST 2014, 302 Vol. 2, Issue 3, ISSN : 2347 – 8446, ISSN : 2347 – 9817, 2014

[7]Mansi Gera, ShivaniGoel,"Data Mining - Techniques, Methods and Algorithms: A Review on Tools and their Validity ",International Journal of Computer Applications (0975 – 8887) Volume 113 – No. 18, March 2015.

[8]SrinivasSivarathri, A.Govardhan," Analysis of Clustering Approaches for Data Mining In Large Data Sources", International Journal on Recent and Innovation Trends in Computing and Communication Volume: 2 Issue: 9, ISSN: 2321-8169, 2590 – 259, 2014.