Imperial Journal of Interdisciplinary Research (IJIR)
Vol-3, Issue-4, 2017
ISSN: 2454-1362, http://www.onlinejournal.in

A Study of Classification and Clustering Algorithm in Data Mining

Sri Sanjana N[1] & Jaya Prathosh R[2] & Dr. N. Chitra Devi[3]
[2]Student, Information Technology, [3]HOD (CSE&IT)
Adithya Institute of Technology, Coimbatore

***Abstract**: Data mining ssis the process of analyzing data from different perspectives and summarizing it into useful information. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Weka is a data mining tool that contain the many machine leaning algorithms. It is provide the facility to classify the data through various algorithms. This paper aims about the study of commonly used algorithms like K means clustering and Decision tree classification.*

***Keywords** — Data Mining, WEKA Classification, Clustering, K means, Decision tree.*

1 INTRODUCTION

Data mining (sometimes called data or knowledge discovery)is the use of automated data analysis techniques to uncover previously undetected relationships among data items. Data mining often involves the analysis of data stored in a data warehouse. There are several data mining techniques are preprocessing, association, classification, pattern recognition and clustering. In our work performs by clustering and classification techniques. In this paper we are working with the clustering and classification because it is most important process, if we have a very large database. The development of Information Technology has generated large amount of databases and huge data in various areas. The research in databases and information technology has given rise to an approach to store and manipulate this precious data for further decision making. Data mining is a Process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis.
Some of the data mining application includes,

1) Marketing: Customer profiling, retention, identification of potential customer, market segmentation.
2) Fraud detection: Identify credit card fraud and intrusion detection.
3) Scientific data analysis: Identify the research decision making data.
4) Text and web mining: used to search text or information on web or given raw data.
5) Agriculture: higher yield and higher market price for crops
Any other applications that involve large amount of data.

LEARNING TECHNIQUES

Machine learning (Mitchell, 1997) is a mature and well-recognized research area of computer science, mainly concerned with the discovery of models, patterns, and other regularities in data. Research areas related to machine learning and data mining include database technology and data warehouses, pattern recognition and soft computing, text and web mining, visualization, and statistics.

– Database technology and data warehouses are concerned with the efficient storage, access and manipulation of data.
– Pattern recognition and soft computing typically provide techniques for classify-ing data items.
– Text and web mining are used for web page analysis, text categorization, as well as filtering and structuring of text documents; natural language processing can provide useful tools for improving the quality of text mining results.
– Visualization concerns the visualization of data as well as the visualization of data mining results.
– Statistics is a classical data analysis discipline, mainly concerned with the analysis of large collections of numerical data.

2 SUPERVISED LEARNING TECHNIQUES

Supervised learning is a data mining task of inferring a function from labeled training data.The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and the desired output value (also called the supervisory signal).

SUPRVISED LEARNING ALGORITHM:

DECISION TREE:

Decision trees are combined of computational and mathematical techniques to aid the representation, generalization and categorization of a given set of data. A Decision tree is a format which contains a root node, branches, and leaf nodes. Each internal node denoted as check on associate degree attribute, every branch denoted as the end result of a check and every leaf node denoted as a category. The top most node within the tree is called as root node. The main goal is to produce a model that predicts the value of a required variable based upon many input variables the decision tree model also uses the prediction based rules classification. The known label of test data is compared along with the classified result.

ALGORITHM FOR DECISION TREE:

Step 1: The leaflet is labeled with the same class if the instances belong to the same class.
Step 2: For each parameters, the potential information will be evaluated and the gain in information will be taken from the test on the parameter.
Step 3: Finally the best parameter will be selected based on the present selection parameter.

INPUT:
D //Training data
OUTPUT :
T //Decision tree

```
DTBUILD (*D)
{
T=φ;
T= Create root node and label with splitting attribute;
T= Add arc to root node for each split predicate and
label;
for each arc do
D= Database created by applying splitting
Predicate to D;
if stopping point reached for this path, then
T'= create leaf node and label with appropriate class;
else
T'= DTBUILD(D);
T= add T' to arc;
}
```

3 UNSUPERVISED LEARNING TECHNIQUES

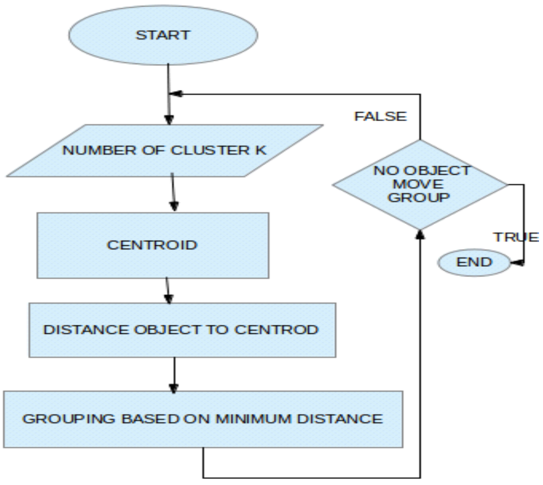
Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses. The most common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data.

UNSUPRVISED LEARNING ALGORITHM: K MEANS CLUSTERING

K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume kclusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as bar center of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more.
Finally, this algorithm aims at minimizing an objective function knows as squared error function given by,

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\|x_i - v_j\|\right)^2$$

Where,
'||xi – vj||' is the Euclidean distance between xi and vj.
'ci' is the number of data points in I'th cluster.
'c' is the number of cluster centers.



Algorithm for K Means Clustering

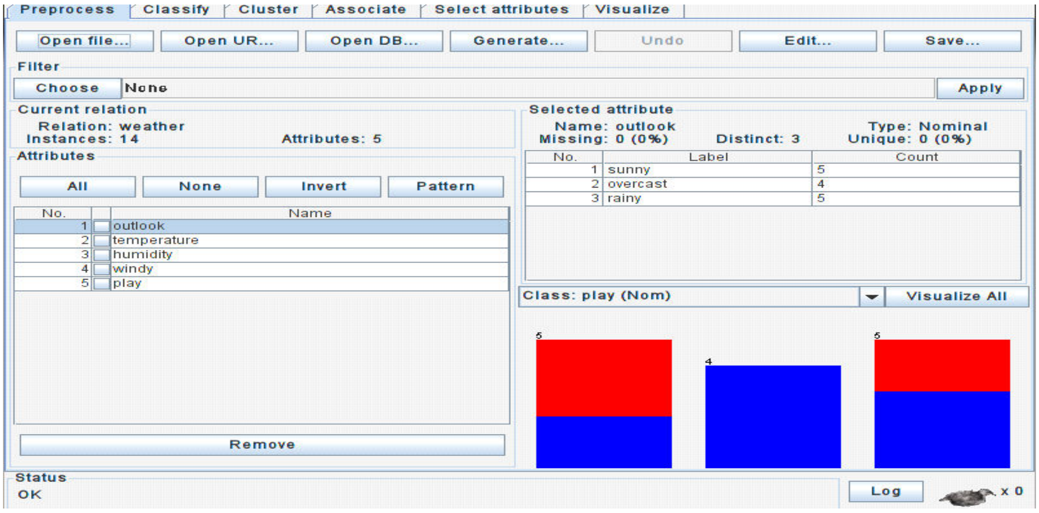
```
Input: C= {c1, c2, c3.....cn}, cluster sets,
D= {d1, d2, Dn} data sets
Output: find mean value µi
Begin
Choose any cluster from Data set D
Repeat
While (Cj € D)
Assign Z as a Cluster centric
Select similar data
Compute Mean Value µi
End
```

4 RESULT AND OUTCOME

DATASET:

```
@relation weather
@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}
@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

CLUSTERING RESULT:



Current relation
Relation: weather
Instances: 14 Attributes: 5

Selected attribute
Name: outlook Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

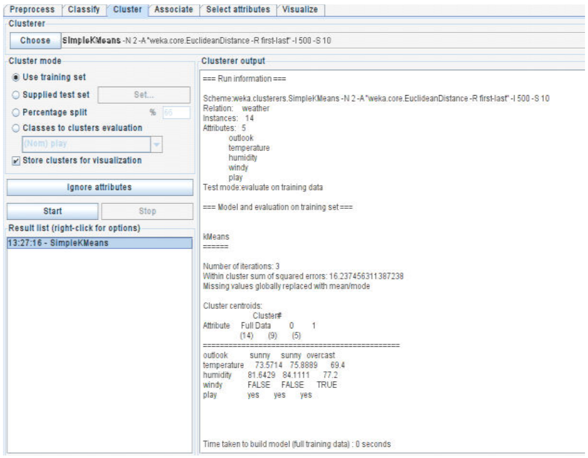
No.	Label	Count
1	sunny	5
2	overcast	4
3	rainy	5

Attributes

No.	Name
1	outlook
2	temperature
3	humidity
4	windy
5	play

Class: play (Nom)

Status
OK



Clusterer
Choose SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

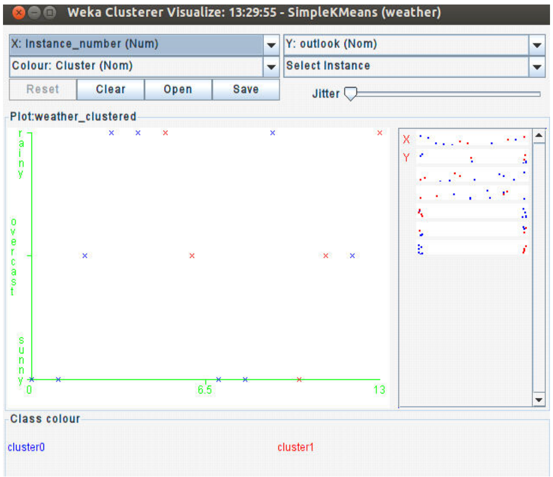
Cluster mode
● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation (Nom) play
☑ Store clusters for visualization

Result list (right-click for options)
13:27:16 - SimpleKMeans

Clusterer output

```
=== Run information ===

Scheme:weka.clusterers.SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation:     weather
Instances:    14
Attributes:   5
              outlook
              temperature
              humidity
              windy
              play
Test mode:evaluate on training data

=== Model and evaluation on training set ===


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 16.237456311387238
Missing values globally replaced with mean/mode

Cluster centroids:
                         Cluster#
Attribute      Full Data        0        1
                    (14)      (9)      (5)
==============================================
outlook            sunny    sunny  overcast
temperature      73.5714  75.8889      69.4
humidity         81.6429  84.1111      77.2
windy              FALSE    FALSE      TRUE
play                 yes      yes       yes




Time taken to build model (full training data) : 0 seconds
```

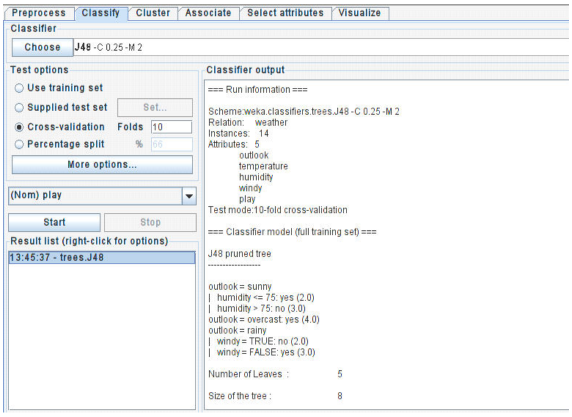
VISUALIZE CLUSTER:



Weka Clusterer Visualize: 13:29:55 - SimpleKMeans (weather)

X: Instance_number (Num)
Y: outlook (Nom)
Colour: Cluster (Nom)
Select Instance

DECISION TREE RESULT:



J48 -C 0.25 -M 2

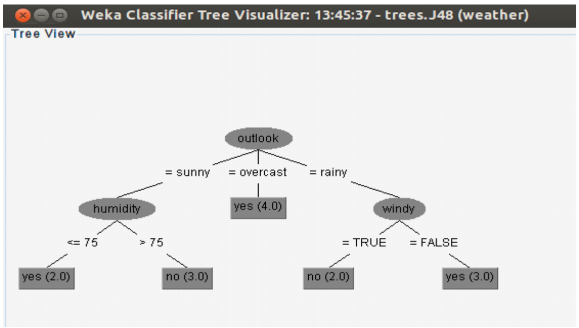
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) play

Result list (right-click for options)
13:45:37 - trees.J48

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     weather
Instances:    14
Attributes:   5
              outlook
              temperature
              humidity
              windy
              play
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

outlook = sunny
|   humidity <= 75: yes (2.0)
|   humidity > 75: no (3.0)
outlook = overcast: yes (4.0)
outlook = rainy
|   windy = TRUE: no (2.0)
|   windy = FALSE: yes (3.0)

Number of Leaves  :          5

Size of the tree :           8
```

VISUALIZE TREE:



5 CONCLUSION

In this paper, the analysis of two algorithms like k means and decision tree are implemented. These two algorithm allow more flexibility with output and can be more powerful weapons in data mining. The classification tree literally creates a tree with branches, nodes, and leaves that lets us take an unknown data point and move down the tree, applying the attributes of the data point to the tree until a leaf is reached and the unknown output of the data point can be determined. The advantage of decision tree is that it provides a theoretical framework for taking into account not only the experimental data to design an optimal classifier, but also a structural behavior for allowing better generalization capability.

The clustering algorithm takes a data set and sorts them into groups, so that conclusions can be made within these groups. An important factor in choosing an appropriate clustering algorithm is the shape of clusters in datasets to be analyzed. Clustering differs from classification by not producing a single output variable, which leads to easy conclusions, but instead requires that you observe the output and attempt to draw own conclusions.

REFERENCES:

1)Aggregated K Means Clustering and Decision Tree Algorithm for Spirometry Data(Indian Journal of Science and Technology, Vol 9(44), DOI: 10.17485/ijst/2016/v9i44/103107, November 2016) K. Rohini and G. Suseendran

2)Analysis of K-Means Algorithm Using Classification Techniques in Mammographic Dataset (International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online) Vol. 3, Issue 4, pp: (237-242), Month: October - December 2015, Available at: **www.researchpublish.com**)

3)Anshl Goyal and Rajnimehta "Performance Comparison Naïve Bayes and J48 Classification Algorithms using Bank Dataset"ISSN 0973-4562, Vol-7, No.11, 2012.

4)Gangnjot Kawrand Amit Chabra "Improved J48 Classification Algorithm for the Prediction of Diabetes", Vol-98, No-22, July 2014.

5)Purusothaman G, Krishnakumari P. A Survey of Data Mining Techniques on Risk Prediction: Heart Disease.

Indian Journal of Science and Technology. 2015 June; 8(12):1–5.

6)Machine learning repository, https://archive.ics.uci.edu/ml/datasets.html

7)Michel Steinbach,George Karypis and Vipin Kumar,"A Comparison of Document Clustering Techniques",

Computer Science Engineering,Technical Report #00-034.

8)https://moodle.umons.ac.be/pluginfile.php/43703 /mod_resource/content/2/WekaTutorial.pdf

9)A Decision Tree Scoring Model Based on Genetic Algorithm and K-Means Algorithm IEEE