Experiment - 3

Aim:- Perform data preprocessing tasks and Demonstrate performing association rule mining on data sets

→ Explore various options available in Weka for preprocessing data and apply Unsupervised filters like Discretization, Resample Filter etc on each dataset

→ Load weather, nominal, Iris, Glass datasets into Weka and run Apriori. Algorithm with different support and confidence values

→ study the rules generated. Apply different discretization filters on numerical attributes and run the Apriori association rule algorithm. study the rules generated.

→ Derive interesting insights and observe the effect of discretization in the rule generation process.

Objectives:-

Data preprocessing is essential before its actual use The dataset is preprocessed in order to check missing values, noisy data and other inconsistencies before executing it to the algorithm

Unsupervised Filter:-

A filter that adds a new nominal attribute representing the cluster assigned to each instance by the specified clustering algorithm.

Apriori Algorithm:-

Apriori algorithm refers to an algorithm that is used in mining frequent product sets and relevant association rules. Generally the apriori algorithm operates on a database containing a huge number of transactions.

WEKA:-

WEKA (Waikota Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikota, New Zealand. Weka is free software available under the GNU General public License.

Preprocess Tab:-

It is first step in machine learning is to preprocess the data. It is used to select the data files, preprocess it and make it fit for applying the various machine learning algorithm.

Loading Data:- The first four buttons at the top of the preprocess section enable you to load data into WEKA.

→ Open file... Brings up a dialog box allowing you to browse for the data file on the local file system.

→ Open URL... Asks for a Uniform Resource Locator address for where the data is stored

→ Open DB... Reads data from a database.

→ Generate... Enables you to generate artificial data from a variety of Data Generators. Using the open file... button you can read files in a variety of formats, WEKA's ARFF format, CSV format, C4.5 format.

Current Relation: Once some data has been loaded, the Preprocess panel shows a variety of information.

→ Relation:- The name of the relation, as given in the file it was loaded from. Filters modify the name of a relation.

→ Instances:- The no. of instances in the data

→ Attributes:- the no. of attributes in the data.

Association Rule:-

An association rule has two parts, an antecedent and a consequent. An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent.

Association rule are created by analyzing data for frequent if then patterns and using the criteria support and confidence to identify the most important

relationships. support is an indication of how frequently the items appear in the database. confidence indicates the number of times the statements have been found to be true.

Support and Confidence Values:-

- Support count:-

The support count of an itemset X, denoted by X.count, in a data set T is the number of transactions in T that contain X. Assume T has n transactions.

Then

$$\text{support} = \frac{(X \cup Y).count}{n}$$

$$\text{confidence} = \frac{(X \cup Y).count}{X.count}$$

support = support(dAUCy)

confidence = support(dAUCy) / support(dAy)

Steps Required:-

1. Open WEKA Tool.
2. click on WEKA Explorer
3. click on Preprocessing tab button.
4. click on open file button.
5. choose WEKA folder in C drive
6. select and click on data option button.
7. choose labor data set and open file.
8. choose filter button and select the Unsupervised - Discretize option and apply
9. click on Associate tab and choose Aprior algorithm
10. click on start button.

Output:

Preprocess	classify	cluster	Associate	select Attributes	Visualize
open file	open URL	open DB	Generate	Undo	Edit	save

Filter

Choose | None | Apply | Stop

Preprocess	classify	cluster	Associate	Select Attributes	Visualize
open file	open URL	open DB	Generate	undo	edit	save

Filter

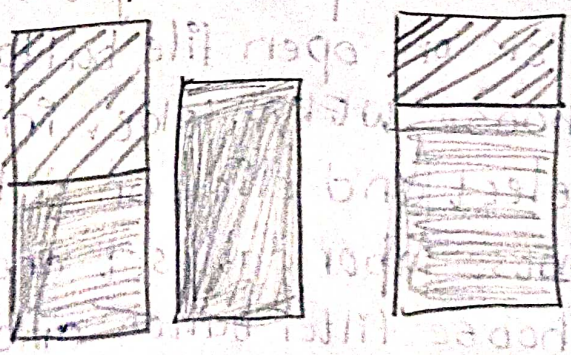
Choose | Discretize -B 10 -M -1.0 -R first-last -precision 6 | apply | Stop

Current relation

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

Select attribute

No.	label	count	weight
1	Sunny	5	
2	overcast		
3	rainy		



Experiment-5

Aim:- Demonstrate performing clustering of data sets
→ Load each dataset into Weka and run simple k-means clustering algorithm with different values of k
→ study the clusters formed. observe the sum of squared errors and centroids, and derive insights.
→ Explore other clustering techniques available in Weka.
→ Explore visualization features of weka to visualize the clusters. Derive interesting insights and explain.

Objectives:-

The goal of clustering is to find distinct groups or "clusters" within a data set. Using a machine language algorithm, the tool creates groups where items in a similar group will, in general, have similar characteristics to each other.

K-Means clustering Algorithm:-

K-Means clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning, which groups the unlabeled dataset into different clusters. Here k defines the number of predefined clusters that need to be created in the process.

Visualization:-

Weka's Visualize panel lets you look at a dataset and select different attributes - preferably numeric ones - for the x and y axes. Instances are shown as points, with different colors for different classes. You can sweep out a rectangle and focus the dataset on the points inside it.

WEKA:-

WEKA (Waikota Environment for knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikota, New Zealand. Weka is free software available under the GNU General public License.

Cluster:.

Cluster Analysis in Data Mining means that to find out the group of objects which are similar to each other in the group but are different from the object in other groups. A clustering algorithm finds group of similar instances in the entire dataset. WEKA supports several clustering algorithms such as EM, Filtered clusterer, Hierarchical cluster, simple k Means etc.

Selecting a clusterer:

By now you will be familiar with the process of selecting and configuring objects: clicking on the clustering scheme listed in the clusterer box at the top of the window brings up a GenericObject Editor dialog with which to choose a new clustering scheme.

Cluster Modes:-

The cluster mode box is used to choose what to cluster and how to evaluate the results. The first three options are the same as for classification: Use training set, Supplied test set and Percentage split except that now the data is assigned to clusters instead of trying to predict a specific class. The fourth mode, Classes to clusters evaluation compares how well the chosen clusters match up with a pre-assigned class in the data.

Ignore Attributes:

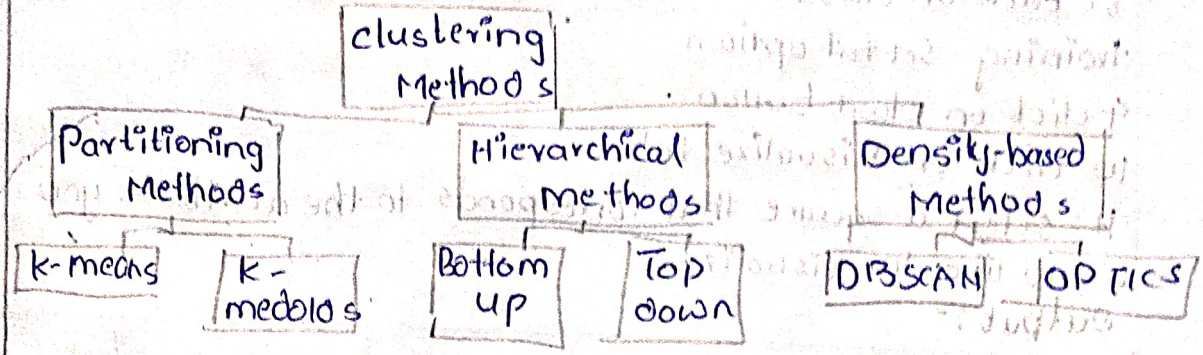
Often, some attributes in the data should be ignored when clustering. The Ignore attributes button brings up a small window that allows you to select which attributes are ignored. Clicking on an attribute in the window highlights it, holding down the SHIFT key selects a range of consecutive attributes, and holding down CTRL toggles individual attributes on and off. To cancel the selection, back out with the Cancel button. To activate it, click the select button. The next time clustering is invoked, the selected attributes are ignored.

Working with Fillers:-

The Filtered clusterer meta-clusterer offers the user the possibility to apply filters directly before the cluster is learned. This approach eliminates the manual application of a filter in the Preprocess panel, since the data gets processed on the fly. Useful if one needs to try out different filter setups.

Learning clusters:-

The cluster section, like the classify section, has start/stop buttons, a result text area and a result list. These all behave just like their classification counterparts. Right-clicking an entry in the result list brings up a similar menu, except that it shows only two visualization options: Visualize cluster assignments and Visualize tree. The latter is grayed out when it is not applicable.



selecting Instances:-

Sometimes it is helpful to select a subset of the data using visualization tool.

1. select Instance:

Click on an individual data point. It brings up a window listing attributes of the point. If more than one point will appear at the same location, more than one set of attributes will be shown.

2. Rectangle:

You can create a rectangle by dragging it around the point.

3. Polygon:

You can select several points by building a free-form polygon. Left-click on the graph to add vertices to the polygon and right-click to complete it.

4. polylines:

To distinguish the points on one side from the once on another, you can build a polyline. left click on the graph to add vertices to the polyline and right-click to finish

steps Required:

1. Open WEKA Tool
2. click on WEKA Explorer
3. click on Preprocessing tab button.
4. click on open file button.
5. choose WEKA folder in C drive.
6. select and click on data option button
7. choose iris data set and open file.
8. click on cluster tab and choose k-means and select use training set tat option
9. click on start button.
10. click on Visualize tab
11. select a square that corresponds to the attributes you would like to visualize.

output:-

Preprocess	classify	cluste	Associate	Select Attributs	Visualize

clusterd

choose SimpleKMeans -init 0 -max-candidates 100

cluste modie

- Use training set
- supplied test set
- percentage split
- classes to clusters evaluation
- Store clusters for Visualization
- Ignore attributes

start stop

cluster output

Preprocess | classify | cluster | Associate | select attribute | Visualize

plot Matrix	Outlook	temperature	humidity	windy	play
play					
Windy					
humidity					
temperature					
Outlook					

plot size [100] —

point size [1] —

Jitter :

update

select Attributes

Sub sample % : 100