

DMWA Lab

Assignment -1

Teghdeep Kapoor

18104050

B12

Q1.

KNIME (Konstanz Information Miner) is easy to use, secluded and provides an open-source information coordination, preparation, examination and investigation stage. KNIME contains devices for information pre-handling, changing, grouping, affiliation leads etc. The benefit of the tool is that WEKA can be coordinated and broaden for conceivable outcomes with KNIME by different administrators.

Language Used - Java

Method Purpose -

- Enables user to visually create data flows easily
- Interactive data models

Advantages -

- Easily visualization of molecular data

Disadvantages -

- No methods for data wrapper
- Not automatic facility for parameter Optimization

WEKA (Waikato Environment for Knowledge Analysis) is a ML tool. It consists of all ML algorithms which are used to solve the real-life application problems.

Language Used - Java

Method Purpose -

- Easy analytics of data and predictive modelling

Advantages -

-Free Extensible

-ARFF, CSV, C4.5, binary are formats used to load files

Disadvantages -

-Weak in statistical analysis

-For parameter optimization of machine learning (No automatic facility)

RapidMiner (RM - some time ago YALE) is a free, adaptable and open-source tool executed in Java. It is a tool for ML, DM, image processing and business analytics.

Language Used - Java

Method Purpose -

-Text Mining, results in visualization, Model validation and optimization

Advantages -

-Full Faculty Model Evaluation Offers more procedures Over 1500 methods for data integration, analysis, visualization Compatible for large users.

Disadvantages -

-Only capable of SQL statements

-Working with only database files.

ORANGE uses the Python language which helps in the visualization of data in DM. It helps in predictive modelling, analysis, selection of subset and empirical analysis. This tool performs tasks like data manipulation and data transformations.

Language Used - C++ and python

Method Purpose -

Data Pre-processing, filtering, and modelling Techniques

Advantages -

-Debugging is better

-Categorization problems like scripting DM are simple.

Disadvantages -

-Weak in statistical analysis

-Limited capabilities of visual representations of data mode.

Q2.

Difference between Data Mining and Data Harvesting

1. Data mining is the process of executing data into an analysis pattern for better client. Data harvesting is the process of extracting data from websites to retrieve quality information
2. Data mining stresses more on creating an analysis chart so that brands can conduct necessary actions according to the behavior patterns of clients. 2 . Data harvesting stresses more on finding data that will help brands to execute, improvise, learn and apply solutions that will cater to assisting their needs.
3. The main agenda of data mining is to create a solution which will matter or will alter in the next few years. 3. The main agenda of data harvesting is to collect information about clients whose behavior pattern will help you better understand their needs
4. Data mining gives a predictive analysis. 4. Data harvesting gives solutions that are coming directly from the mouth of what clients are expecting.
5. Data mining provides a long term solution to assist clients fluctuating preferences. 5. Data harvesting provides solutions which are needed on the spot to assist clients
6. Data mining is an automated process 6. Data harvesting can be done automated or manually.
7. Data mining collects tons of data you have in hand and creates a clear report of what the next few years will be like with reference to clients. 7. Data harvesting extracts any data which you require so that you can easily have it in your system to keep a closer check on.
8. Another word for data mining is knowledge discovery in database. 8. Another word for data harvesting is data scraping.
9. With data mining, algorithms are used so that valuable data can be easily structured. 9. With data harvesting, the process is simple. You just need to click on the website which you want to scrape data from and the process begins henceforth.
10. A team of experts is required to conduct efficient data mining processes. 10. Data harvesting doesn't require expert's attention, even a beginner can conduct this process without any hassle.
11. Data mining tools: Rapidminer, Orange, Weka, KNIME and Sisense (top 5) 11. Data harvesting tools: Import.io, OutWithHub, Octaparse, Visual Web Ripper and Web scraper (top 5)

Q3.

Data Mining Techniques

Data mining is highly effective, so long as it draws upon one or more of these techniques:

1. Tracking patterns. One of the most basic techniques in data mining is learning to recognize patterns in your data sets. This is usually a recognition of some aberration in your data happening at regular intervals, or an ebb and flow of a certain variable over time. For example, you might see that your sales of a certain product seem to spike just before the holidays, or notice that warmer weather drives more people to your website.
2. Classification. Classification is a more complex data mining technique that forces you to collect various attributes together into discernable categories, which you can then use to draw further conclusions, or serve some function. For example, if you're evaluating data on individual customers' financial backgrounds and purchase histories, you might be able to classify them as "low," "medium," or "high" credit risks. You could then use these classifications to learn even more about those customers.
3. Association. Association is related to tracking patterns, but is more specific to dependently linked variables. In this case, you'll look for specific events or attributes that are highly correlated with another event or attribute; for example, you might notice that when your customers buy a specific item, they also often buy a second, related item. This is usually what's used to populate "people also bought" sections of online stores.
4. Outlier detection. In many cases, simply recognizing the overarching pattern can't give you a clear understanding of your data set. You also need to be able to identify anomalies, or outliers in your data. For example, if your purchasers are almost exclusively male, but during one strange week in July, there's a huge spike in female purchasers, you'll want to investigate the spike and see what drove it, so you can either replicate it or better understand your audience in the process.
5. Clustering. Clustering is very similar to classification, but involves grouping chunks of data together based on their similarities. For example, you might choose to cluster different demographics of your audience into different packets based on how much disposable income they have, or how often they tend to shop at your store.
6. Regression. Regression, used primarily as a form of planning and modeling, is used to identify the likelihood of a certain variable, given the presence of other variables. For example, you could use it to project a certain price, based on other factors like availability, consumer demand, and competition. More specifically, regression's main focus is to help you uncover the exact relationship between two (or more) variables in a given data set.
7. Prediction. Prediction is one of the most valuable data mining techniques, since it's used to project the types of data you'll see in the future. In many cases, just recognizing and understanding historical trends is enough to chart a somewhat accurate prediction of what will happen in the future. For example, you might review consumers' credit histories and past purchases to predict whether they'll be a credit risk in the future.

Q4.

Lists conferences as well as publications, and has these top 10 publications -

1. ACM SIGKDD International Conference on Knowledge discovery and data mining
2. IEEE Transactions on Knowledge and Data Engineering
3. ACM International Conference on Web Search and Data Mining
4. IEEE International Conference on Data Mining (ICDM)
5. ACM Conference on Recommender Systems
6. SIAM International Conference on Data Mining
7. arXiv Databases (cs.DB)
8. European Conference on Machine Learning and Knowledge Discovery in Databases
9. Knowledge and Information Systems
10. ACM Transactions on Knowledge Discovery from Data (TKDD)

Q5. I)

Weka Explorer

The explorer is where you play around with your data and think about what transforms to apply to your data, what algorithms you want to run in experiments. The Explorer interface is divided into 5 different tabs:

- Preprocess: Load a dataset and manipulate the data into a form that you want to work with.
- Classify: Select and run classification and regression algorithms to operate on your data.
- Cluster: Select and run clustering algorithms on your dataset.
- Associate: Run association algorithms to extract insights from your data.
- Select Attributes: Run attribute selection algorithms on your data to select those attributes that are relevant to the feature you want to predict.
- Visualize: Visualize the relationship between attributes.

Weka Experimenter

This interface is for designing experiments with your selection of algorithms and datasets, running experiments and analyzing the results. The tools for analyzing results are very powerful, allowing you to consider and compare results that are statistically significant over multiple runs.

Knowledge Flow

Applied machine learning is a process and the Knowledge Flow interface allows you to graphically design that process and run the designs that you create. This includes the loading and transforming of input data, running of algorithms and the presentation of results. It's a powerful interface and metaphor for solving complex problems graphically.

II)

Dataset File

@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,yes

sunny,80,90,TRUE,no

overcast,83,86,FALSE,yes

overcast,70,96,FALSE,yes

rainy,68,80,TRUE,yes

rainy,65,70,FALSE,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,TRUE,yes

rainy,71,91,TRUE,no

Analysis of Dataset using WEKA

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose **None** Apply Stop

Current relation: Relation: weather Instances: 14 Attributes: 5 Sum of weights: 14

Attributes: All None Invert Pattern

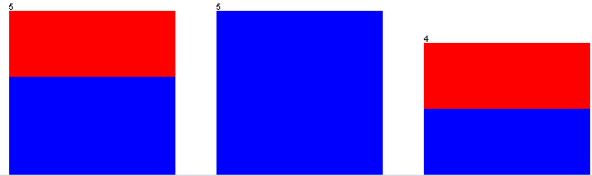
No.	Name
1	<input checked="" type="checkbox"/> outlook
2	<input type="checkbox"/> temperature
3	<input type="checkbox"/> humidity
4	<input type="checkbox"/> windy
5	<input type="checkbox"/> play

Remove

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

No.	Label	Count	Weight
1	sunny	5	5.0
2	overcast	5	5.0
3	rainy	4	4.0

Class: play (Nom) Visualize All



Status: OK Log x 0

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose **None** Apply Stop

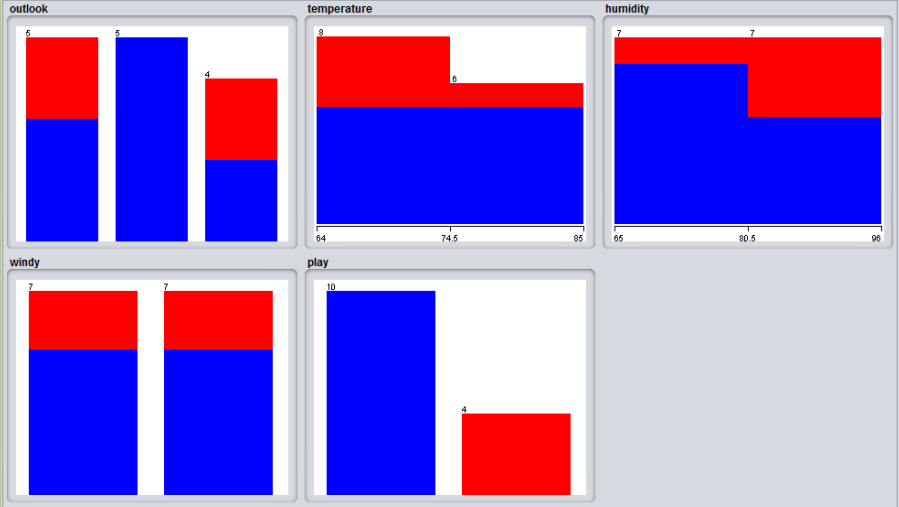
Current relation: Relation: weather Instances: 14 Attributes: 5 Sum of weights: 14

Attributes: All None Invert Pattern

No.	Name
1	<input type="checkbox"/> outlook
2	<input type="checkbox"/> temperature
3	<input type="checkbox"/> humidity
4	<input type="checkbox"/> windy
5	<input checked="" type="checkbox"/> play

Remove

All attributes



Status: OK Log x 0