Data Mining and Web algorithm

Lab Assignment 1:

[Feb 14-19, 2022]

Patil Amit Gurusidhappa
19104004
B11

1. Explore the various data mining tools, namely Rapid Miner, Orange, Weka, KNIME etc. Compare these tools on the basis of following parameters.
a. Features

RapidMiner Features	**Orange**	**Weka**	**Knime**
1. Data Access. Access, load & analyze any type of data. 2. Data Exploration. Extract statistics & key information. 3.Data Prep. Expertly cleanse data for predictive analytics	Orange consists of a canvas interface onto which the user places widgets and creates a data analysis workflow. Widgets offer basic functionalities such as reading the data, showing a data table, selecting features, training predictors, comparing learning algorithms, visualizing data elements, etc	machine learning, data mining, preprocessing, classification, regression, clustering, association rules, attribute selection, experiments, workflow and visualization	Scalability through sophisticated data handling (intelligent automatic caching of data in the background while maximizing throughput performance)

d. Open source/licensed

RapidMiner Features	**Orange**	**Weka**	**Knime**
Yes	Yes	Yes	Yes

e. Supportive OS

RapidMiner Features	Orange	Weka	Knime
1. Windows 7, Windows 8, Windows 8.1, Windows 10 (64-bit highly recommended) 2. Linux (64-bit only) 3. MacOS X 10.10 - 10.14	Windows macOS Linux	Windows macOS Linux	Windows macOS Linux

b. Supportive Algorithms

RapidMiner Features	Orange	Weka	Knime
Classification algorithms · 5. k-nearest neighbors · 6. Tree-based algorithms · 7. Support vector machine · 8. Neural networks	Orange Data Mining Library Logistic Regression. Random Forest. Simple Random Forest. Softmax Regression. k-Nearest Neighbors. Naive Bayes. Support Vector Machines. Linear Support Vector Machines	data preparation, classification, regression, clustering, association rules mining, and visualization	clustering, rule induction, decision tree, association rules, naïve bayes, neural networks, support vector machines, etc

c. Application areas

RapidMiner Features	Orange	Weka	Knime
RapidMiner provides data mining and machine learning procedures including: data loading and transformation (ETL), data preprocessing	The tool has components for machine learning, add-ons for bioinformatics and text mining and it is packed with features	Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and	used in pharmaceutical research, it is also used in other areas such as CRM customer data analysis, business

and visualization, predictive analytics and statistical modeling, evaluation, and deployment.	for data analytics.	visualization	intelligence, text mining and financial data analysis. Recently attempts were made to use KNIME as robotic process automation (RPA) tool.

2. Discuss four most useful data mining techniques along with its applications.

1. Tracking patterns. 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.

Application : 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.

Application:, 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;

Application :, 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.

Application : 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.

Application : 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.

Application : 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.

Application : review consumers' credit histories and past purchases to predict whether they'll be a credit risk in the future

3. Explore the various Journals, Conferences, and Symposiums in the area of data Mining.

Conferences:

Legal Data Mining Conference: How can machine learning improve legal decision-making?

SIAM International Conference on Data Mining 2021: Doctoral Forum Panel

Day 02 - NCDM 2021 - 5th National E-Conference on Data Mining & ICT

Journals

	Publication	h5-index	h5-median	Type
1	ACM SIGKDD International Conference on Knowledge discovery and data mining	67	98	Conference
2	IEEE Transactions on Knowledge and Data Engineering	66	111	Journal
3	ACM International Conference on Web Search and Data Mining	58	94	Conference
4	IEEE International Conference on Data Mining (ICDM)	39	64	Conference
5	Knowledge and Information Systems (KAIS)	38	52	Journals

4. Do Practice on free and open source software Waikato Environment for Knowledge Analysis (WEKA) tool which allows you to mine your own data for trends and patterns. Familiarize yourself with the process of data format and loading of data set. Refer the WEKA manual given in help folder for the Introduction of WEKA.

After successful installation try to explore answers for the following:

1. What is the purpose of the following in Weka:
- **The Explorer**

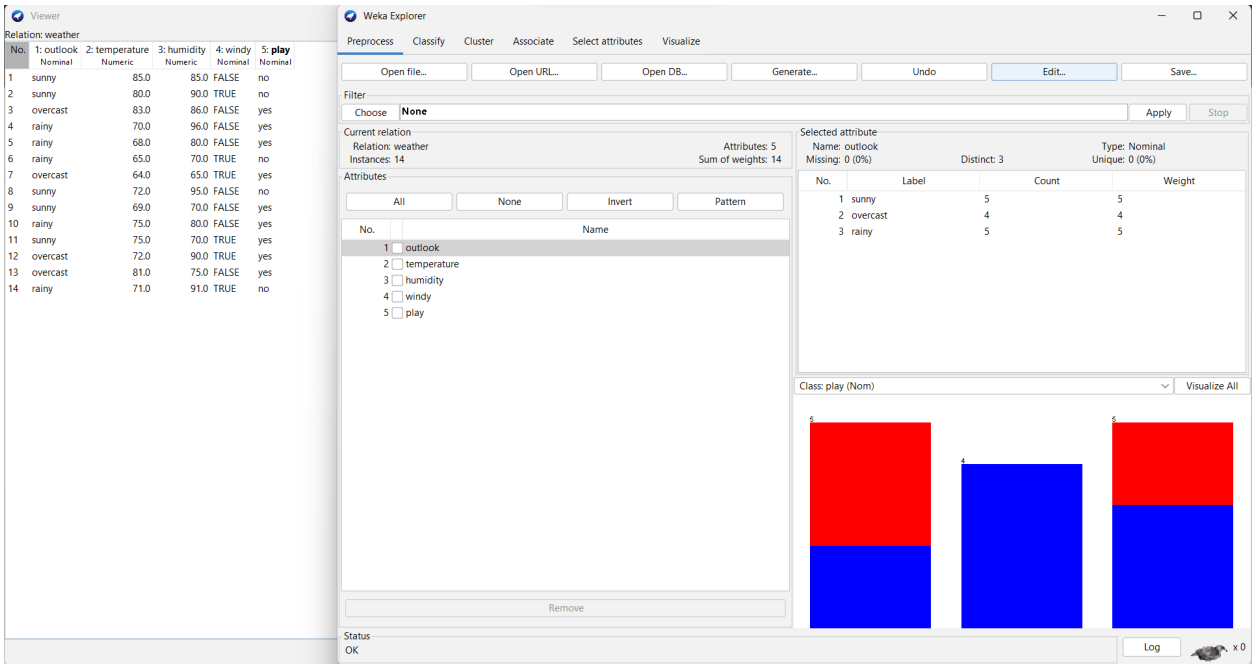
The Weka Knowledge Explorer is an easy to use graphical user interface that harnesses the power of the weka software

● **The Knowledge Flow interface**
The Knowledge Flow Interface is an alternative to the Explorer, and it lets you lay out filters, classifiers, and evaluators interactively on a 2D canvas

● **The Experimenter**
Weka Experimenter allows you to design your own experiments of running algorithms on datasets, run the experiments and analyze the results.

● **The command-line interface**
The Command Line interface is useful for running experiments automatically on a server, without using a GUI

2. Create a tiny data set for weather problem below in arff (attribute relation file format) format (introduction given on next page) and analyze on WEKA.



@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature numeric
@attribute humidity numeric
@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