

DATA WAREHOUSE AND DATA MINING

PATEL CHITRAL ANILBHAI

201903103510035

❖ Practical-1

- Case study of any three data mining applications and make a detailed note on them.

Data mining is a process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.

Data mining is widely used in diverse areas. There are a number of commercial data mining system available today and yet there are many challenges in this field. In this tutorial, we will discuss the applications and the trend of data mining.

Data Mining Applications

Here is the list of areas where data mining is widely used –

- Future health care
- Market basket Anylysis
- Education
- Manufacturing engineering
- Customer Relationship Management
- Fraud Detection
- Intrusion Dection
- Customer segmantion

- Financial Banking
- Criminal investigation
- Retail Industry
- Telecommunication Industry
- Bioinformatic

Future Healthcare

Data mining holds great potential to improve health systems. It uses data and analytics to identify best practices that improve care and reduce costs. Researchers use data mining approaches like multi-dimensional databases, machine learning, soft computing, data visualization and statistics. Mining can be used to predict the volume of patients in every category. Processes are developed that make sure that the patients receive appropriate care at the right place and at the right time. Data mining can also help healthcare insurers to detect fraud and abuse.

Market Basket Analysis

Market basket analysis is a modelling technique based upon a theory that if you buy a certain group of items you are more likely to buy another group of items. This technique may allow the retailer to understand the purchase behaviour of a buyer. This information may help the retailer to know the buyer's needs and change the store's layout accordingly. Using differential analysis comparison of results between different stores, between customers in different demographic groups can be done.

Education

There is a new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational Environments. The goals of EDM are identified as predicting students' future learning behaviour, studying the effects of educational support, and advancing scientific knowledge about learning. Data mining can be used by an institution to take accurate decisions and also

to predict the results of the student. With the results the institution can focus on what to teach and how to teach. Learning pattern of the students can be captured and used to develop techniques to teach them.

Manufacturing Engineering

Knowledge is the best asset a manufacturing enterprise would possess. Data mining tools can be very useful to discover patterns in complex manufacturing process. Data mining can be used in system-level designing to extract the relationships between product architecture, product portfolio, and customer needs data. It can also be used to predict the product development span time, cost, and dependencies among other tasks.

Research Analysis

History shows that we have witnessed revolutionary changes in research. Data mining is helpful in data cleaning, data pre-processing and integration of databases. The researchers can find any similar data from the database that might bring any change in the research. Identification of any co-occurring sequences and the correlation between any activities can be known. Data visualisation and visual data mining provide us with a clear view of the data.

❖ Practical-2

- **Write summarized description of any 10 data mining tools including Weka and Rapid Miner.**

There are many tools in the market both open source and proprietary with varying levels of sophistication. At the root, each tool helps with implementing a data mining strategy, but the difference lies in the level of sophistication you the customer of these software needs. There are tools that do well in a specific domain such as the Financial domain or the Scientific domain.

Let's look at the more popular ones in the market.

- Rapid Miner
- Oracle Data Mining
- IBM SPSS Modeler
- Knime
- Python
- Orange
- Kaggle
- Rattle
- Weka
- Teradata

1. RAPID MINER

A data science software platform providing an integrated environment for various stages of data modelling including data preparation, data cleansing, exploratory data analysis, visualization and more. The techniques that the software helps with are machine learning, deep learning, text mining and predictive analytics. Easy to use GUI tools that take you through the modelling process. This tool written entirely in Java is an open-source framework and is wildly popular in the data mining world.

2. ORACLE DATA MINING

Oracle, the world leader in database software, combines its prowess in database technologies with Analytical tools and brings you Oracle Advanced Analytics Database part of the Oracle Enterprise Edition. It features several data mining algorithms for classification, regressing, prediction, anomaly detection and more. This is proprietary software and is supported by Oracle technical staff in helping your business build a robust data mining infrastructure at the enterprise scale.

The algorithms integrate directly with Oracle database kernel and operate natively on data stored in its own database, eliminating the need for extraction of data into standalone analytics servers. The Oracle Data Miner provides GUI tools taking the user through the process of creating, testing and applying data models

3. IBM SPSS MODELER

IBM is again a big name in the data space when it comes to large enterprises. It combines well with leading technologies to implement a robust enterprise-wide solution. IBM SPSS Modeller is a visual data science and machine learning solution, helping in shortening the time to value by speeding up operational tasks for data scientists. IBM SPSS Modeler will have you covered from drag and drop data exploration to machine learning.

The software is used in leading enterprises for data preparation, discovery, predictive analytics, model management and deployment. The tool helps organizations to tap into their data assets and applications easily. One of the advantages of proprietary software is its ability to meet robust governance and security requirements of an organization at the enterprise level, and this reflects in every tool that IBM offers on the data mining front.

4. KNIME

Konstanz Information Miner is an open-source data analysis platform, helping you with build, deployment and scale in no time. The tool aims to help make predictive intelligence accessible to inexperienced users. It aims to make the process easy by it is a step by step guide based GUI tools. The product markets itself as an End to End Data Science product, that helps create and production data science using its single easy and intuitive environment.

5. PYTHON

Python is a freely available and open-source language that is known to have a quick learning curve. Combined with is the ability as a general-purpose language and it is a large library of packages that help build a system for creating data models from the scratch, Python makes for a great tool for organizations who want the software they use to be custom built to their specifications.

With Python, you won't get the fancy stuff that proprietary software offers, but the functionality is there for anybody to pick up and creates their own environment with graphical interfaces of their liking. What also supports python is the large online community of package developers who ensure the packages on

offer are robust and secure. One of the features Python is known for in this field is powerful on the fly visualization features it offers.

6. ORANGE

Orange is a machine learning and data science suite, using python scripting and visual programming featuring interactive data analysis and component-based assembly of data mining systems. Orange offers a broader range of features than most other Python-based data mining and machine learning tools. It is a software that has over 15 years of active development and use. Orange also offers a visual programming platform with GUI for interactive data visualization.

7. KAGGLE

The largest community of data scientists and machine learning professionals. Kaggle although started as a platform for machine learning competitions, is now extending its footprint into the public cloud-based data science platform arena. Kaggle now offers code and data that you need for your data science implementations. There are over 50k public datasets and 400k public notebooks that you can use to ramp up your data mining efforts. The huge online community that Kaggle enjoys is your safety net for implementation-specific challenges.

8. RATTLE

The rattle is an R language based GUI tool for data mining requirements. The tool is free and open-source and can be used to get statistical and visual summaries of data, the transformation of data for data models, build supervised and unsupervised machine learning models and compare model performance graphically.

9. WEKA

Waikato Environment for Knowledge Analysis (Weka) is a suite of machine learning tools written in Java. A collection of visualization tools for predictive modelling in a GUI presentation, helping you build your data models and test them, observing the model performances graphically.

10. TERADATA

A cloud data analytics platform marketing its no code required tools in a comprehensive package offering enterprise-scale solutions. With Vantage Analyst, you don't need to be a programmer to code complex machine learning algorithms. A simple GUI based system for quick enterprise-wide adoption.

❖ **Practical-3**

Perform following tasks as per given instruction:

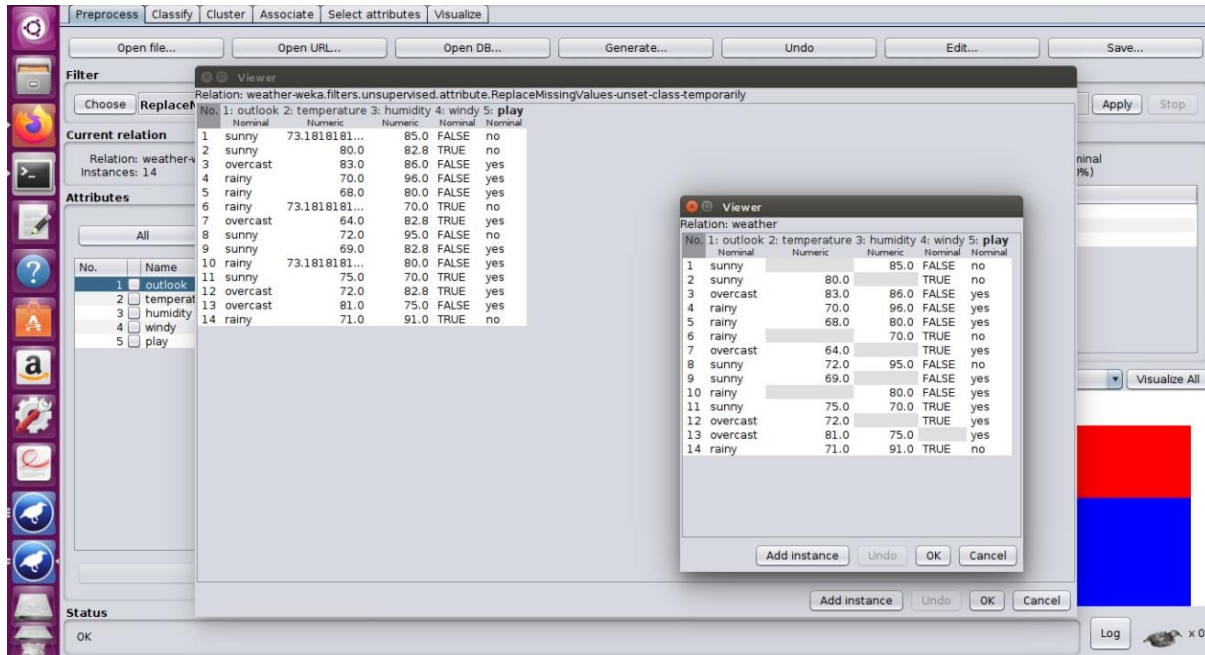
- a. Apply Data Cleaning methods on given data set using Weka.**
- b. To perform data cleaning by handling missing values.**

**Preprocessing for the missing value, by replacing them
with all the following for the given dataset:**

- 1) With the global constant like Unknown or–infinity.**
- 2) Use the central tendency of attribute Mean
(numerical attributes).**
- 3) Use the class wise attribute Mean or median.**
- 4) Apply all three normalization methods to any one
numeric attribute.**

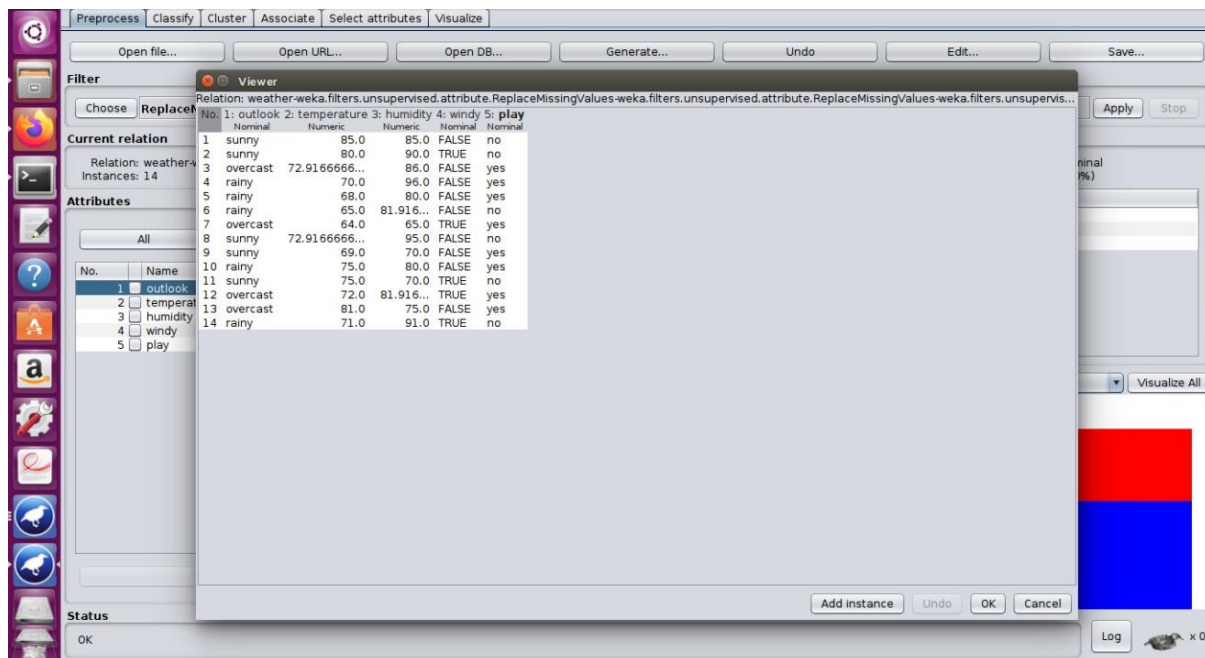
Pre-processing for the missing value, by replacing them with all the following: -

Replacing missing value:- In The Missing value is replace by counting average of a column.



With the global constant like Unknown or -infinity:-

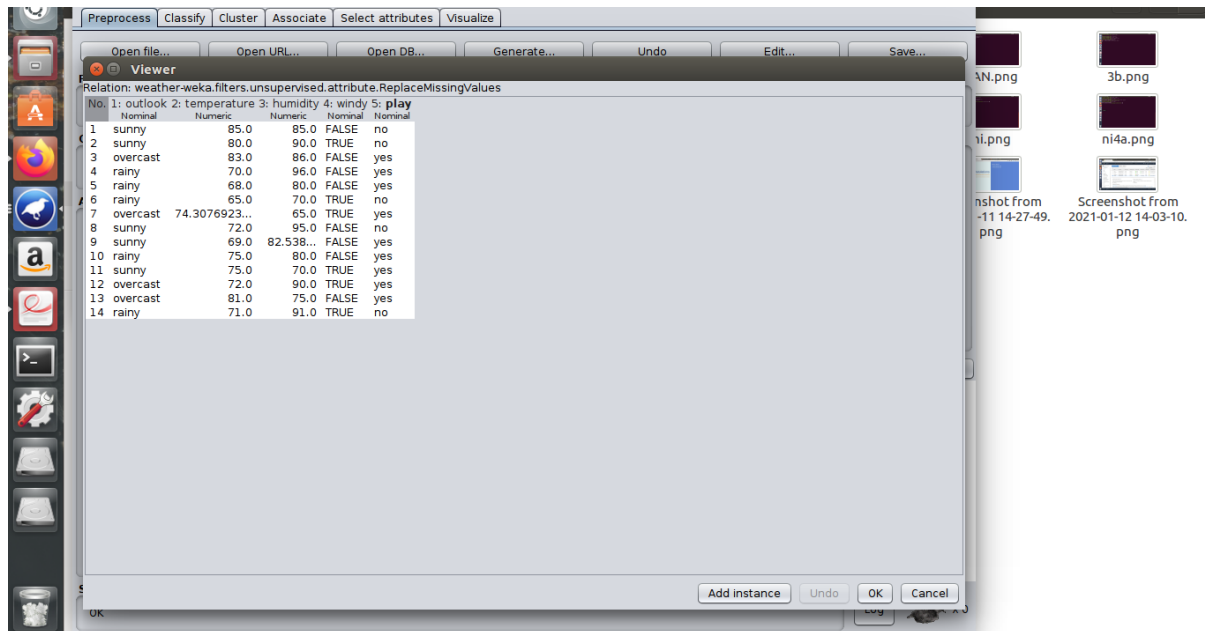
In This Null value is replaced by writing null or infinity value like 99999.....



Use the central tendency of attribute Mean (numerical attributes).

- Use the class wise attribute Mean or median.

- Apply all three normalization methods to any one numeric attribute.

A. Normalise:

Relation: weather-weka.filters.unsupervised.attribute.ReplaceMissingValues

No.	1: outlook	2: temperature	3: humidity	4: windy	5: play
	Nominal	Numeric	Numeric	Nominal	Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	74.3076923...	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	82.538...	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

B. Standardise:

Relation: weather.symbolic-weka.filters.unsupervised.attribute.ReplaceMissingWithUserConstant-Afirst-last-R0-Fyyyy-MM-ddTHH:mm:ss-weka.filters.unsupervis...

No.	1: outlook	2: temperature	3: humidity	4: windy	5: play
	Nominal	Nominal	Nominal	Nominal	Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	ABCD	FALSE	yes
4	rainy	mild	ABCD	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	ABCD	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

C. Math-expression:

Relation: weather-weka.filters.unsupervised.attribute.MathExpression-E(A-MIN)/(MAX-MIN)

No.	1: outlook	2: temperature	3: humidity	4: windy	5: play
	Nominal	Nominal	Nominal	Nominal	Nominal
1	sunny	hot	high	FALSE	no
2	sunny	0.76190476...	0.6451...	TRUE	no
3	overcast	0.90476190...	0.8064...	FALSE	yes
4	rainy	0.28571428...	1.0	FALSE	yes
5	rainy	0.19047619...	0.4838...	FALSE	yes
6	rainy	0.04761904...	0.1612...	TRUE	no
7	overcast	0.0	0.0	TRUE	yes
8	sunny		0.9677...	FALSE	no
9	sunny	0.23809523...	0.1612...	FALSE	yes
10	rainy	0.52380952...	0.4838...	FALSE	yes
11	sunny	0.52380952...	0.1612...	TRUE	yes
12	overcast	0.38095238...	0.8064...	TRUE	yes
13	overcast	0.80952380...	0.3225...	FALSE	yes
14	rainy	0.33333333...	0.8387...	TRUE	no

D.

Relation: weather-weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0

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

	Nominal	Numeric	Numeric	Nominal	Nominal
1	sunny	1.0	0.6451...	FALSE	no
2	sunny	0.75	0.8064...	TRUE	no
3	overcast	0.9	0.6774...	FALSE	yes
4	rainy	0.25	1.0	FALSE	yes
5	rainy	0.15	0.4838...	FALSE	yes
6	rainy	0.0	0.1612...	TRUE	no
7	overcast		0.0	TRUE	yes
8	sunny	0.35	0.9677...	FALSE	no
9	sunny	0.2	0.1612...	FALSE	yes
10	rainy	0.5	0.4838...	FALSE	yes
11	sunny	0.5	0.1612...	TRUE	yes
12	overcast	0.35		TRUE	yes
13	overcast	0.8	0.3225...	FALSE	yes
14	rainy	0.3	0.8387...	TRUE	no

Add instance Undo OK Cancel