BINUS UNIVERSITY
GLOBAL CLASS

Mid Exam Answer

Student name: *Christopher Angelo - 2440041503*

Course: *Data Mining (COMP6746001)* – Professor: *Ms. Dewi Suryani*
Due date: *Nov 25th, 2022*

Essay

Knowledge discovery in databases (KDD) is the process of discovering interesting patterns or knowledge from a collection of extensive data. There are several steps in the KDD process. Please describe them in detail according to your understanding. Then, provide an example case and elaborate on how each step of the KDD process is employed in that example.

Answer 1. Knowledge Discovery in Database is essentially the process of extracting information (knowledge) from a database using pattern analysis, data mining, and machine learning. The process of KDD is divided into 6 steps, namely:

1. **Data Cleaning** — Detecting improper data and correcting them (or deleting them from the dataset) to ensure the proper functioning of the data mining process. This step is important because it is the core data that we are modifying; it can affect the accuracy of the results of the data mining process.

2. **Data Integration** — Combining data from multiple sources into a single dataset, usually by merging them (SQL JOIN & MERGE) to each other.

3. **Data Selection** — Selecting the data that will be used for the data mining process. This step is usually done by filtering and selecting a specific data based on the task and metrics that is want to be extracted or mined.

4. **Data Transformation** — Transforming the data into a format that is suitable for the data mining process. For example, it is often you do this by normalizing the data (e.g. using Min-Max Normalization) or by converting the data into a different format (e.g. converting a categorical data into a numerical data).

5. **Data Mining** — The actual process of extracting knowledge from the data using a data mining algorithm (e.g. K-Means Clustering, Apriori Algorithm, etc.).

6. **Knowledge Representation** — Representing the knowledge that has been extracted from the data in a format that is easy to understand by humans (e.g. using a decision tree). This step is often done by using a visualization tool (e.g. Tableau, Power BI, etc.) and is usually presented to the stakeholders of a company.

Suppose you have a dataset of crimes happenning in a city. You want to know the most common crime that happens in the city. You can use the KDD process to answer this question. The steps are as follows:

1. **Data Cleaning** — You need to clean the data by removing data that is not complete (e.g. the data of the crime that does not have a complete information).

2. **Data Integration** — You need to integrate the data from multiple sources (e.g. the data of the crime that is reported by the police and the data of the crime that is reported by the citizens).
3. **Data Selection** — You need to select the data that is relevant to the question (e.g. the data of the crime that is in the city). You can also help constraining the model to be up-to-date by filtering the data and selecting the data that is in the last 5 years.
4. **Data Transformation** — You need to transform the data into a format that is suitable for the data mining process (e.g. converting the crime category into a numerical data / label).
5. **Data Mining** — Use a data mining software / algorithm to help you with this process.
6. **Knowledge Representation** — Represent the knowledge that has been extracted from the data in a format that is easy to understand by humans (e.g. using a visualization tool).

Illustrate how the data warehouse is related to data mining based on your perspective, and attach any facts to support your statement.

Answer 2. Data warehouse and data mining is related to one another as it is both a part of the tools when doing Business Intelligence.

Data warehouse touches on how the data of something is stored, organized and 'linked' to each other. A good data warehouse is one that is able to store data in a way that is easy to access and understand. It also will help on the datamining process where it reduces the amount of data cleaning and preprocessing that needs to be done.

Data mining is the process of extracting information (knowledge) from a dataset. It is usually done by using a data mining algorithm (e.g. K-Means Clustering, Apriori Algorithm, etc.). In this context, the data mining process will be done using the data that is stored in the Data warehouse. Improving the data warehouse will also improve the data mining process.

Case Study

Calculate the mean, median, mode, and quartiles based on the total quantities per product sold on Oct 15th, 2022.

Answer 1. From the transaction data provided, the following are the transactions made in October 15:

ID	Items	Total
TR123450	{[AP:2], [OB:5], [CR:3], [BM:6]}	Rp 200 250
TR123451	{[OB:10], [PC:3], [WB:6], [CR:3]}	Rp 355 500
TR123452	{[AP:3], [WB:2], [CR:5], [GC:3]}	Rp 273 000
TR123453	{[AP:1], [OB:2], [PC:1], [WB:1], [CR:3], [GC:1], [BM:3]}	Rp 192 750
TR123454	{[AP:5], [GC:5]}	Rp 242 500
TR123455	{[OB:12], [WB:3], [CR:6], [BM:6]}	Rp 355 500

Converting these transactions to bought item counts and sorting it, the data will look as follows:

Item	Calculation	Count
PC	0 + 3 + 0 + 1 + 0 + 0	4
GC	0 + 0 + 0 + 1 + 5 + 0	6
AP	2 + 0 + 3 + 1 + 5 + 0	11
WB	0 + 6 + 2 + 1 + 0 + 3	12
BM	6 + 0 + 0 + 3 + 0 + 6	15
CR	3 + 3 + 5 + 3 + 0 + 6	20
OB	5 + 10 + 0 + 2 + 0 + 12	29

The mean, median, mode, and quartiles are calculated as follows:

$$\bar{x} = \frac{11 + 29 + 4 + 12 + 20 + 15 + 6}{7} = 13.8\bar{5}$$

$$\text{Median} = x_4 = 12$$

$$\text{Mode} = 29$$

Normalize the total amount per transaction using min-max normalization, Zscore normalization, and normalization by decimal scaling

Answer 2. Variables with different scales than each other do not contribute equally to the model. Feeding an un-normalized data might create a bias in the model. This is why we normalize data before feeding it to a model.

The following table is the transaction data, highlighting the total amount:

Total
Rp. 200 250
Rp. 355 500
Rp. 273 000
Rp. 192 750
Rp. 242 500
Rp. 333 750
Rp. 336 000
Rp. 192 250
Rp. 189 000
Rp. 268 500

Normalizing data using min-max normalization. Min-max normalization is a normalization technique that scales the data to a fixed range of 0 to 1. The formula for min-max normalization is as follows:

$$x_{n_{scaled}} = \frac{x_n - \min_x}{\max_x - \min_x}$$

From the data above, the minimum value of the price $\min_x = 189000$. The maximum value from the data $\max_x = 355500$. Normalizing the data using min-max normalization, the following will be the calculation and the normalized value (rounded to 2 decimal places):

Total	Equation	Normalized
Rp. 200 250	$\frac{200250-189000}{355500-189000}$	0.06
Rp. 355 500	$\frac{355500-189000}{355500-189000}$	1.00
Rp. 273 000	$\frac{273000-189000}{355500-189000}$	0.50
Rp. 192 750	$\frac{192750-189000}{355500-189000}$	0.02
Rp. 242 500	$\frac{242500-189000}{355500-189000}$	0.32
Rp. 333 750	$\frac{333750-189000}{355500-189000}$	0.86
Rp. 336 000	$\frac{336000-189000}{355500-189000}$	0.88
Rp. 192 250	$\frac{192250-189000}{355500-189000}$	0.01
Rp. 189 000	$\frac{189000-189000}{355500-189000}$	0.00
Rp. 268 500	$\frac{268500-189000}{355500-189000}$	0.47

Normalizing data using Z-score normalization. When using Z-score normalization, we would need to first calculate the average of the data (mean) and the standard deviation of the data. The formula for Z-score normalization is as follows:

$$x_{i_{scaled}} = \frac{x_i - \bar{x}}{\sigma}$$

Calculating the mean of the data, the calculation is as follows:

$$\bar{x} = \frac{\text{Sum of all data}}{\text{Data count}} = \frac{2583500}{10} = 258350$$

Calculating the standard deviation of the data, the calculation is as follows:

$$\begin{aligned}\sigma^2 &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} \\ &= \frac{(200250 - 258350)^2 + (355500 - 258350)^2 + \cdots + (268500 - 258350)^2}{10} \\ &= \frac{38579275000}{10} \\ &= 3857927500\end{aligned}$$

$$\begin{aligned}\sigma &= \sqrt{3857927500} \\ &\approx 62112.21\end{aligned}$$

Making ends meet, the calculation of the normalized value and the result of the calculation are as follows:

Total	Equation	Normalized
Rp. 200 250	$\frac{200250-258350}{62112.21}$	-0.93
Rp. 355 500	$\frac{355500-258350}{62112.21}$	1.56
Rp. 273 000	$\frac{273000-258350}{62112.21}$	0.23
Rp. 192 750	$\frac{192750-258350}{62112.21}$	-1.05
Rp. 242 500	$\frac{242500-258350}{62112.21}$	-0.25
Rp. 333 750	$\frac{333750-258350}{62112.21}$	1.21
Rp. 336 000	$\frac{336000-258350}{62112.21}$	1.25
Rp. 192 250	$\frac{192250-258350}{62112.21}$	-1.06
Rp. 189 000	$\frac{189000-258350}{62112.21}$	-1.11
Rp. 268 500	$\frac{268500-258350}{62112.21}$	0.16

Normalization using decimal scaling. When normalizing data using decimal scaling, you simply move the decimal point to the front or back (or more formally, multiply or divide the value by 10) until the largest value is equal to or less than 1. Looking at the data, it is suitable that we move the decimal point 6 places to the left (or more formally, dividing the value by 10^6). The following will be the normalized value after the decimal point is moved 6 places to the left:

Total	Normalized
Rp. 200 250	0.20
Rp. 355 500	0.35
Rp. 273 000	0.27
Rp. 192 750	0.19
Rp. 242 500	0.24
Rp. 333 750	0.33
Rp. 336 000	0.33
Rp. 192 250	0.19
Rp. 189 000	0.18
Rp. 268 500	0.26

Determine all frequent itemsets using the Apriori algorithm and then list the association rules of the transaction data with the min_sup = 30% and confidence = 80%.

Answer 3. With min_sup of 30%, this means that a item will be considered 'frequent' if it shows up in at least $30\% \times 10 = 3$ times in the transaction data. After analyzing the data further, the following is the list of frequent itemsets:

$$\{AP, OB, PC, WB, CR, GC, BM\}$$

This fortunately corresponds with every item in the transaction data. To get the association rule, I will be manually 'cross' the data with each other, checking all combinations of them if they cross above the confidence level (yes, there are $7 \times 7 = 49$ combination which I've manually checked). In this case, confidence is calculated as $\frac{\text{Amount of time the association is true}}{\text{Amount of time premise appears in txn}}$.

Crossing each and every product to each other, with a confidence of 0.8% (needs to satisfy 80% of purchases), we get the following association rules:

Association	Confidence (≥ 0.8)
$AP \implies CR$	$\frac{5}{6} = 0.8\overline{3}$
$OB \implies CR$	$\frac{5}{5} = 1.0$
$OB \implies BM$	$\frac{4}{5} = 0.8$
$PC \implies CR$	$\frac{5}{6} = 0.8\overline{3}$
$WB \implies CR$	$\frac{5}{5} = 1.0$
$GC \implies CR$	$\frac{5}{6} = 0.8\overline{3}$
$BM \implies OB$	$\frac{4}{5} = 0.8$
$BM \implies CR$	$\frac{5}{5} = 1$