Faculty of Computing and Information Technology/Department of Computer Science and Mathematics

BACS3013 Data Science

Tutorial 7 (Descriptive Analytics - part 2)

Q1. The task attributes of the four data mining tasks discussed in class are briefly described below:
Association rule and sequential pattern mining - Customer ID, Transaction ID and Item.
Classification/prediction - input and the class label attributes
Clustering mining - input attributes

The following are the data fields in the data mining server log:
User ID, Session ID, Dataset ID, MiningTask ID, Parameter Value, Accuracy

a) Which task will you perform to identify the data mining tasks that tend to be performed in the same session? Describe the attributes you choose and how they are mapped to the data mining task attributes listed above.

b) Which task will you perform to identify the sequence of data mining tasks that users tend to perform on the same data set over time? Describe the attributes you choose and how they are mapped to the data mining task attributes listed above.

c) Which task will you perform to determine if the Parameter Value level (low, medium or high) and the level of Parameter Value adjustment (small, moderate or large) tend to have a positive or negative impact on Accuracy. Describe the attributes you choose and how they are mapped to the data mining task attributes listed above.

Ans:

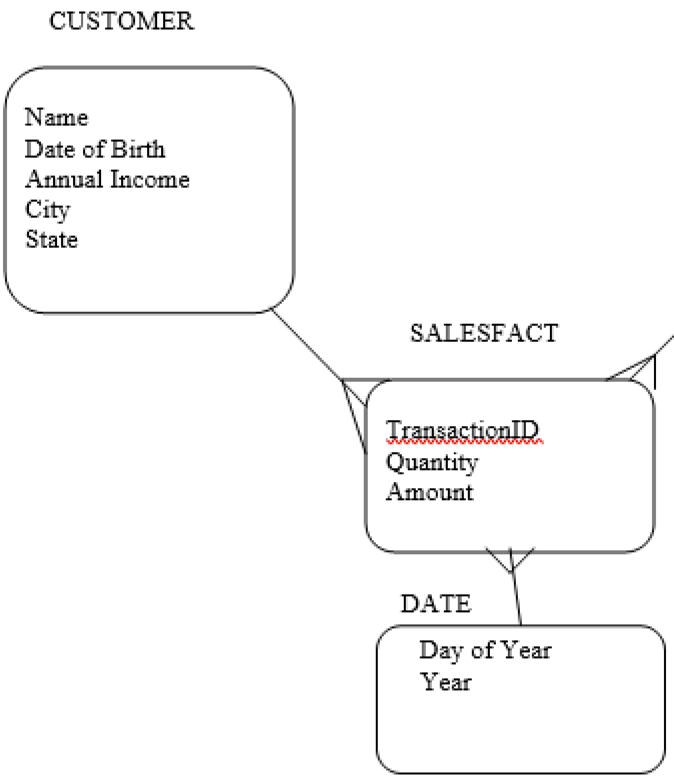
a)
I will suggest using association rule mining. In this data mining task, Session ID can be mapped to Transaction ID and MiningTask ID can be mapped to Item.

b)
I will suggest using sequential pattern mining. In this data mining task, Dataset ID can be mapped to Customer ID, Session ID can be mapped to Transaction ID and MiningTask ID can be mapped to Item.

c)
I will suggest using classification. In this data mining task, input attributes include parameter value level and level of parameter value adjustment, and the class label attribute is the impact on accuracy (i.e., positive, negative, or no impact).

BACS3013 Data Science

Q2. The following is the star schema for Sales Department of your company.



Note:

- TransactionID is used as the primary key in the fact table because there might be more than one transaction for each customer and product in a given day.
- The Introduction Date for a product is the date when it is first introduced into the market.

a) The clustering task was selected to identify *customer segmentation.* Suggest the attributes including derived attributes to be used in the clustering task and justify your answer.

b) Recommend a standardization or normalization method for the attributes in a distance function.

c) You are asked to recommend a classification/predication task to be performed on the above data set.

i. Specify the input and class label attributes you choose for this classification/prediction task. Give an example of business decision(s) that can benefit from the classification/prediction results using the input and class label attributes of your choice.

ii. Define and give an example of noise using the data set above.

iii. Assume that you will use a decision tree classifier. Specify and compare the different tree pruning approaches.

iv. Suppose you are using a neural network instead of a decision tree. List at least three possible parameters you want to tune to improve its performance during the training period.

BACS3013 Data Science

(Note: Q2 (c)(iii) & (iv) will be covered in Predictive Analytics)

Ans:

a) **Customer age, customer's annual income, # of days since a customer's last purchase, the total number of a customer's sales transactions, and the total amount of a customer's purchases. I am interested in groups of similar customers based on customer age, income and life-time value using the last three derived attributes.**

b) **For each attribute, calculate the mean value and the mean absolute deviation. Calculate the standardized value = (original value – mean value)/mean absolute deviation.**

c) **i) Input: Customer city, State, age, Income. Output: Product Type**
Business Analysis: These choices for input and output attributes enable us to understand the impact of customer demographics on product type preference. In marketing, this is called "customer segmentation."

ii) **Noise refers to records with the same input attribute values but different class labels. For example, in customer table, the same customer name with different city and state may be a noise: Is it because of erroneous input, or is it because the customer just moved>**

iii)

- **Prepruning: Halting creations of unreliable branches by statistically determine the goodness of further tree splits**
- **Postpruning: Remove unreliable branches from a full tree by minimizing error rates or required encoding bits**

iv) Hidden layer node number, learning rate, epochs, momentum, accuracy threshold, hidden layer number,

Q3. What are the similarities and differences between agglomerative and divisive approaches?

Ans:

Both are hierarchical clustering methods. Agglomerative clustering/approach is also known as bottom-up approach. It starts with one-point (singleton) clustrers and recursively merges two or more appropriate clusters. Divisive clustering is also known as top-down approach. It starts with one cluster of all data points and recursively splits the most appropriate cluster until a stopping criterion is achieved.

Q4. What are the advantage and disadvantage of hierarchical clustering?

Ans:

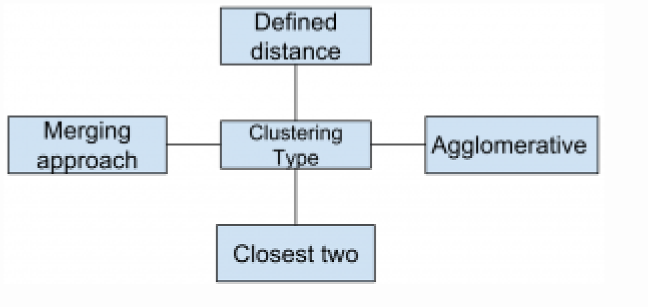
Advantage:

Flexibility (needs no fixed number of clusters). For instance, hierarchical clustering is suitable for semantic analysis and text mining. The number of clusters of corpora would best be divided into is unknown.

Disadvantage:

Expensive computational cost.

Q5. Which clustering type has characteristic shown in the figure below?



Ans:

Hierarchical clustering **groups data over a variety of scales by creating a cluster tree or dendrogram.**

Q6. Which of the following is finally produced by Hierarchical Clustering ?

a) final estimate of cluster centroids
b) tree showing how close things are to each other
c) assignment of each point to clusters
d) All of the Mentioned

Ans: b

Explanation:Hierarchical clustering is an agglomerative approach.